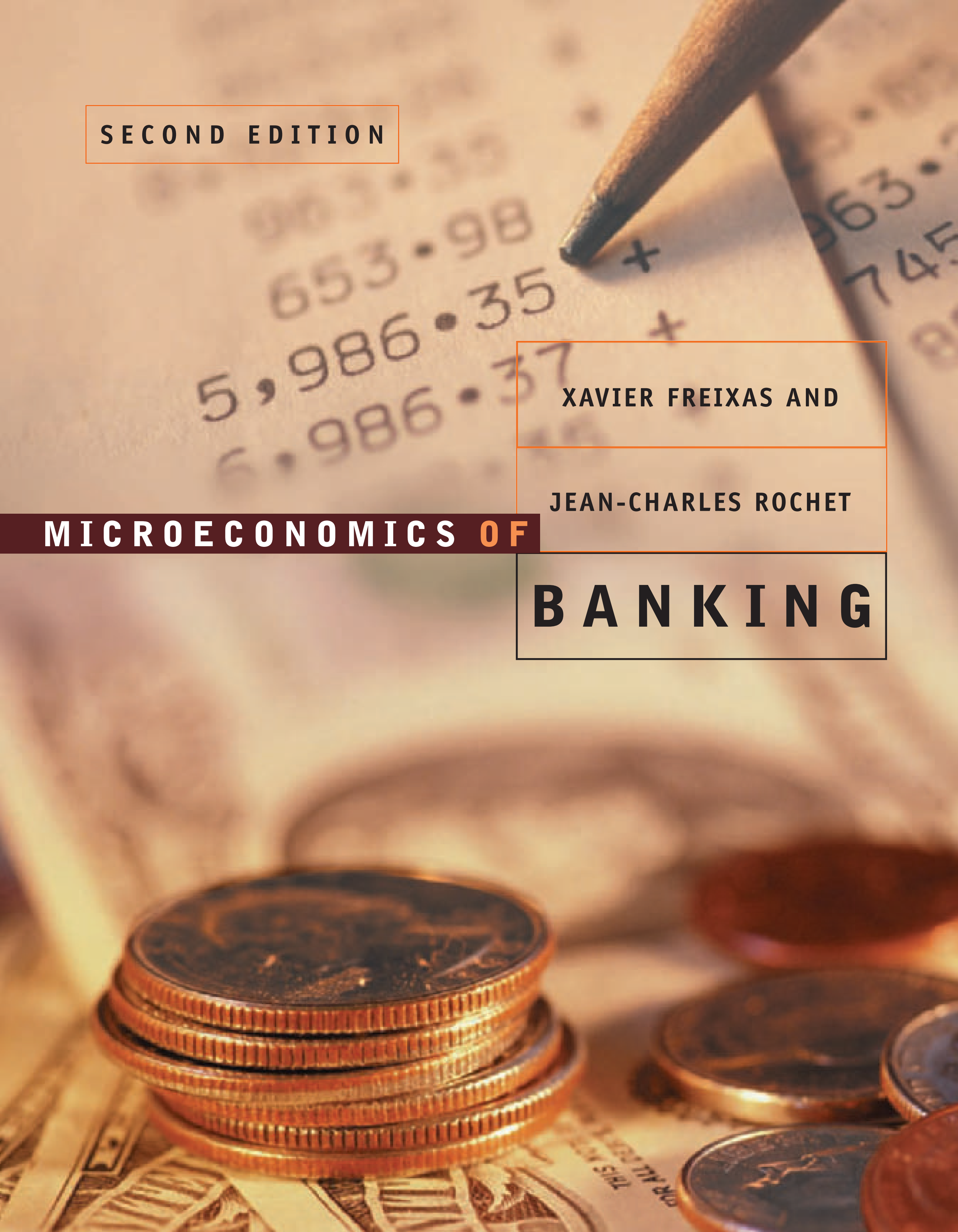
SECOND EDITION

XAVIER FREIXAS AND

JEAN-CHARLES ROCHET

# MICROECONOMICS OF BANKING



# Microeconomics of Banking

# Microeconomics of Banking

## Second Edition

**Xavier Freixas and Jean-Charles Rochet**

**The MIT Press**
**Cambridge, Massachusetts**
**London, England**

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email ⟨special_sales@mitpress.mit.edu⟩ or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Freixas, Xavier.
Microeconomics of banking / Xavier Freixas and Jean-Charles Rochet.—2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-06270-1 (hardcover : alk. paper)
1. Banks and banking. 2. Finance—Mathematical models. 3. Microeconomics. I. Rochet, Jean-Charles. II. Title.
HG1601.F74 2008
332.1—dc22 2007018937

10 9 8 7 6 5 4 3 2 1

A la mémoire de Jean-Jacques Laffont

# Contents

## Figures

# Preface

During the last three decades, the economic theory of banking has entered a process of change that has overturned economists' traditional view of the banking sector. Before that, the banking courses of most doctoral programs in economics, business, or finance focused either on management aspects (with a special emphasis on risk) or on monetary aspects and their macroeconomic consequences. Thirty years ago, there was no such thing as a microeconomic theory of banking, for the simple reason that the Arrow-Debreu general equilibrium model (the standard reference for microeconomics at that time) was unable to explain the role of banks in the economy.[1]

Since then, a new paradigm has emerged (the asymmetric information paradigm), incorporating the assumption that different economic agents possess different pieces of information on relevant economic variables and will use this information for their own profit. This paradigm has proved extremely powerful in many areas of economic analysis. In banking theory it has been useful in explaining the role of banks in the economy and pointing out the structural weaknesses of the banking sector (exposure to runs and panics, persistence of rationing on the credit market, recurrent solvency problems) that may justify public intervention.

This book provides a guide to this new microeconomic theory of banking. It focuses on the main issues and provides the necessary tools to understand how they have been modeled. We have selected contributions that we found to be both important and accessible to second-year doctoral students in economics, business, or finance.

## What Is New in the Second Edition?

Since the publication of the first edition of this book, the development of academic research on the microeconomics of banking has been spectacular. This second edition attempts to cover most of the publications that are representative of these new developments. Three topics are worth mentioning.

First, the analysis of competition between banks has been refined by paying more attention to nonprice competition, namely, competition through other strategic variables than interest rates or service fees. For example, banks compete on the level of the asset risk they take or the intensity of the monitoring of borrowers. These dimensions are crucial for shedding light on two important issues: the competition-stability trade-off and the effect of entry of new banks, both of concern for prudential regulation.

Second, the literature on the macroeconomic impact of the financial structure of firms has made significant progress on at least two questions: the transmission of monetary policy and the effect of capital requirements for banks on the functioning of the credit market.

Finally, the theoretical foundations of banking regulation have been clarified, even though the recent developments in risk modeling (due in particular to the new Basel accords on banks solvency regulation) have not yet led to a significant parallel development of economic modeling.

## Prerequisites

This book focuses on the theoretical aspects of banking. Preliminary knowledge of the institutional aspects of banking, taught in undergraduate courses on money and banking, is therefore useful. Good references are the textbooks of Mishkin (1992) or Garber and Weisbrod (1992). An excellent transition between these textbooks and the theoretical material developed here can be found in Greenbaum and Thakor (1995).

Good knowledge of microeconomic theory at the level of a first-year graduate course is also needed: decision theory, general equilibrium theory and its extensions to uncertainty (complete contingent markets) and dynamic contexts, game theory, incentives theory. An excellent reference that covers substantially more material than is needed here is Mas Colell, Whinston, and Green (1995). More specialized knowledge on contract theory (Salanié 1996; Laffont and Martimort 2002; Bolton and Dewatripont 2005) or game theory (Fudenberg and Tirole 1991; Gibbons 1992; Kreps 1990; Myerson 1991) is not needed but can be useful. Similarly, good knowledge of the basic concepts of modern finance (Capital Asset Pricing Model, option pricing) is recommended (see, e.g., Huang and Litzenberger 1988 or Ingersoll 1987). An excellent complement to this book is the corporate finance treatise of Tirole (2006). Finally, the mathematical tools needed are to be found in undergraduate courses in differential calculus and probability theory. Some knowledge of diffusion processes (in connection with Black-Scholes's option pricing formula) is also useful.

## Outline of the Book

Because of the discouraging fact that banks are useless in the Arrow-Debreu world (see section 1.7 for a formal proof), our first objective is to explain why financial intermediaries exist. In other words, what are the important features of reality that are overlooked in the Arrow-Debreu model of complete contingent markets? In chapter 2 we explore the different theories of financial intermediation: transaction costs, liquidity insurance, coalitions of borrowers, and delegated monitoring.

The second important aspect that is neglected in the complete contingent market approach is the notion that banks provide costly services to the public (essentially management of loans and deposits), which makes them compete in a context of product differentiation. This is the basis of the industrial organization approach to banking, studied in chapter 3.

Chapter 4 is dedicated to optimal contracting between a lender and a borrower. In chapter 5 we study the equilibrium of the credit market, with particular attention to the possibility of rationing at equilibrium, a phenomenon that has provoked important discussions among economists.

Chapter 6 is concerned with the macroeconomic consequences of financial imperfections. In chapter 7 we study individual bank runs and systemic risk, and in chapter 8 the management of risks in the banking firm. Finally, chapter 9 is concerned with bank regulation and its economic justifications.

## Teaching the Book

According to our experience, the most convenient way to teach the material contained in this book is to split it into two nine-week courses. The first covers the most accessible material of chapters 1–5. The second is more advanced and covers chapters 6–9. At the end of most chapters we have provided a set of problems, together with their solutions. These problems not only will allow students to test their understanding of the material contained in each chapter but also will introduce them to some advanced material published in academic journals.

## Acknowledgments

Our main debt is the intellectual influence of the principal contributors to the microeconomic theory of banking, especially Benjamin Bernanke, Patrick Bolton, Doug Diamond, Douglas Gale, Martin Hellwig, David Pyle, Joe Stiglitz, Jean Tirole, Robert Townsend, and several of their co-authors. We were also influenced by the ideas of Franklin Allen, Ernst Baltensperger, Sudipto Bhattacharya, Arnoud Boot, John

Boyd, Pierre André Chiappori, Mathias Dewatripont, Phil Dybvig, Gérard Gennotte, Charles Goodhart, Gary Gorton, Ed Green, Stuart Greenbaum, André Grimaud, Oliver Hart, Bengt Holmström, Jack Kareken, Nobu Kiyotaki, Hayne Leland, Carmen Matutes, Robert Merton, Loretta Mester, John Moore, Rafael Repullo, Tony Santomero, Elu Von Thadden, Anjan Thakor, Xavier Vives, Neil Wallace, David Webb, Oved Yosha, and Marie-Odile Yannelle. Some of them have been very helpful through their remarks and encouragement. We are also grateful to Franklin Allen, Arnoud Boot, Vittoria Cerasi, Gabriella Chiesa, Gerhard Clemenz, Hans Degryse, Antoine Faure-Grimaud, Denis Gromb, Loretta Mester, Bruno Parigi, François Salanié, Elu Von Thadden, and Jean Tirole, who carefully read preliminary versions of this book and helped us with criticism and advice.

The material of this book has been repeatedly taught in Paris (ENSAE), Toulouse (Master "Marchés et Intermédiaires Financiers"), Barcelona (Universitat Pompeu Fabra), Philadelphia (Wharton School), and Wuhan University. We benefited a lot from the remarks of our students. The encouragement and intellectual support of our colleagues in Toulouse (especially Bruno Biais, André Grimaud, Jean-Jacques Laffont, François Salanié, and Jean Tirole) and Barcelona (Thierry Foucault and José Marin) have also been very useful. Finally, we are extremely indebted to Claudine Moisan and Marie-Pierre Boé, who competently typed the (too many) different versions of this book without ever complaining about the sometimes contradictory instructions of the two co-authors.

The second edition benefited from the comments of many people, especially Judit Montoriol, Henri Pagès, and Henriette Prast.

## Note

1. This disappointing property of the Arrow-Debreu model is explained in chapter 1.

## References

Bolton, P., and M. Dewatripont. 2005. *Contract theory.* Cambridge, Mass.: MIT Press.

Fudenberg, D., and J. Tirole. 1991. *Game theory.* Cambridge, Mass.: MIT Press.

Garber, P., and S. Weisbrod. 1992. *The economics of banking, liquidity and money.* Lexington, Mass.: D. C. Heath.

Gibbons, R. 1992. *A primer on game theory.* New York: Wheatsheaf.

Greenbaum, S. I., and A. V. Thakor. 1995. *Contemporary financial intermediation.* Fort Worth, Texas: Dryden Press.

Huang, C. F., and D. Litzenberger. 1988. *Foundations for financial economics.* Amsterdam: North-Holland.

Ingersoll, J. E. 1987. *Theory of financial decision making.* Totowa, N.J.: Rowan and Littlefield.

Kreps, D. 1990. *Game theory and economic modelling.* Oxford: Clarendon Press.

Laffont, J. J., and D. Martimort. 2002. *The theory of incentives.* Princeton, N.J.: Princeton University Press.

Mas Colell, A., M. D. Whinston, and J. Green. 1995. *Microeconomic theory.* Oxford: Oxford University Press.

Mishkin, F. S. 1992. *The economics of money, banking and financial markets.* London: Scott, Foresman.

Myerson, R. 1991. *Game theory: Analysis of conflicts.* Cambridge, Mass.: Harvard University Press.

Salanié, B. 1996. *The theory of contracts.* Cambridge, Mass.: MIT Press.

Tirole, J. 2006. *Corporate finance.* Princeton, N.J.: Princeton University Press.

# Microeconomics of Banking

# 1 Introduction

## 1.1 What Is a Bank, and What Do Banks Do?

Banking operations may be varied and complex, but a simple operational definition of a bank is available: *a bank is an institution whose current operations consist in granting loans and receiving deposits from the public*. This is the definition regulators use when they decide whether a financial intermediary (this term is defined in chapter 2) has to submit to the prevailing prudential regulations for banks. This legal definition has the merit of insisting on the core activities of banks, namely, deposits and loans. Note that every word of it is important:

- The word *current* is important because most industrial or commercial firms occasionally lend money to their customers or borrow from their suppliers.[1]
- The fact that both loans are offered *and* deposits are received is important because it is the combination of lending and borrowing that is typical of commercial banks. Banks finance a significant fraction of their loans through the deposits of the public. This is the main explanation for the fragility of the banking sector and the justification for banking regulation. Some economists predict that commercial banks offering both loan and deposit transactions will someday disappear in favor of two types of specialized institutions,[2] on the one hand "narrow" banks or mutual funds, which invest the deposits of the public in traded securities, and on the other hand finance companies or credit institutions, which finance loans by issuing debt or equity.
- Finally, the term *public* emphasizes that banks provide unique services (liquidity and means of payment) to the general public. However, the public is not, in contrast with professional investors, armed to assess the safety and soundness of financial institutions (i.e., to assess whether individuals' interests are well preserved by banks). Moreover, in the current situation, a public good (access to a safe and efficient payment system) is provided by private institutions (commercial banks). These two reasons (protection of depositors, and the safety and efficiency of the payment system) have traditionally justified public intervention in banking activities.

Banks also play a crucial role in the allocation of capital in the economy. As Merton (1993, 20) states, "A well developed smoothly functioning financial system facilitates the efficient life-cycle allocation of household consumption and the efficient allocation of physical capital to its most productive use in the business sector." For centuries, the economic functions of the financial system were essentially performed by banks alone. In the last 30 years financial markets have developed dramatically, and financial innovations have emerged at a spectacular rate. As a result, financial markets are now providing some of the services that financial intermediaries used to offer exclusively. Thus, for example, a firm involved in international trade can now hedge its exchange rate risk through a futures market instead of using a bank contract. Prior to the development of futures markets, the banking sector was an exclusive provider of such services.

In order to provide a better understanding of how financial intermediation improves the allocation of capital in the economy, it is necessary to examine in more detail what functions banks perform. Contemporary banking theory classifies banking functions into four main categories:

- Offering liquidity and payment services
- Transforming assets
- Managing risks
- Processing information and monitoring borrowers

This, of course, does not mean that every bank has to perform each of these functions. Universal banks do, but specialized banks need not. In view of this classification, our initial definition of banks (as the institutions whose current operations consist in making loans and receiving deposits) may seem too simple. Therefore, to illustrate the proposed classification, the following sections examine how banks perform each of these functions.

## 1.2 Liquidity and Payment Services

In a world without transaction costs, like in the standard Arrow-Debreu model, there would be no need for money. However, as soon as one takes into account the existence of frictions in trading operations, it becomes more efficient to exchange goods and services for money, rather than for other goods and services, as in barter operations.[3] The form taken by money quickly evolved from commodity money (a system in which the medium of exchange is itself a useful commodity) to fiat money (a system in which the medium of exchange is intrinsically useless, but its value is guaranteed by some institution, and therefore it is accepted as a means of payment).[4] Historically, banks played two different parts in the management of fiat money: money change (exchange between different currencies issued by distinct institutions) and provision of payment services. These payment services cover both the management of clients' accounts and the finality of payments, that is, the guarantee by the bank that the debt of the payor (who has received the goods or services involved in the transaction) has been settled to the payee through a transfer of money.

### 1.2.1 Money Changing

Historically, the first activity of banks was money changing. This is illustrated by the etymology of the word: the Greek word for bank (*trapeza*) designates the balance that early money changers used to weigh coins in order to determine the exact quantity of precious metal the coins contained.[5] The Italian word for bank (*banco*) relates to the bench on which the money changers placed their precious coins.[6] These money-changing activities played a crucial role in the development of trade in Europe in the late Middle Ages.

The second historical activity of banks, namely, management of deposits, was a consequence of their money-changing activities. This is well documented, for example, in Kohn (1999). Early deposit banks were fairly primitive because of the necessity for both the payee (the deposit bank) and the payor to meet with a notary.[7] Most of the time, these deposits had a zero or even negative return because they were kept in vaults rather than invested in productive activities. If depositors considered it advantageous to exchange coins for a less liquid form of money, it was mainly because of the advantages of safekeeping, which reduced the risk of loss or robbery. Thus initially bank deposits were not supposed to be lent, and presumably the confidence of depositors depended on this information being public and credible. This means that deposit banks tried to build a reputation for being riskless.[8]

Apart from safekeeping services, the quality of coins was also an issue because coins differed in their composition of precious metals and the governments required the banks to make payments in good money. This issue had implications for the return paid on deposits. As Kindleberger (1993, 48) puts it, "The convenience of a deposit at a bank—safety of the money and the assurance that one will receive money of satisfactory quality—meant that bank money went to a premium over currency, which varied from zero or even small negative amounts when the safety of the bank was in question, to 9 to 10 percent." Still, once the coins themselves became of homogeneous quality, deposits lost this attractive feature of being convertible into "good money." However, because deposits were uninsured, the increased efficiency obtained by having a uniform value for coins (implying a decrease in transaction costs), with coins and bills exchanging at their nominal value, did not necessarily apply to deposits. This point was later considered of critical importance during the free banking episodes discussed in chapter 9.

### 1.2.2 Payment Services

Species proved to be inadequate for making large payments, especially at a distance, because of the costs and risks involved in their transportation. Large cash imbalances between merchants were frequent during commercial fairs, and banks played an important part in clearing merchants' positions. Clearing activities became especially important in the United States and Europe at the end of the nineteenth century, leading to modern payment systems, which are networks that facilitate the transfer of funds between the bank accounts of economic agents. The safety and efficiency of these payment systems have become a fundamental concern for governments and central banks, especially since the deregulation and internationalization of financial markets, which have entailed a large increase in interbank payments, both nationally and internationally.[9]

## 1.3 Transforming Assets

Asset transformation can be seen from three viewpoints: convenience of denomination, quality transformation, and maturity transformation. *Convenience of denomination* refers to the fact that the bank chooses the unit size (denomination) of its products (deposits and loans) in a way that is convenient for its clients. It is traditionally seen as one of the main justifications of financial intermediation. A typical example is that of small depositors facing large investors willing to borrow indivisible amounts. More generally, as Gurley and Shaw (1960) argued, in an early contribution, financial intermediaries provide the missing link between the financial products that firms want to issue and the ones desired by investors. Banks then simply play the role of intermediaries by collecting the small deposits and investing the proceeds into large loans.

*Quality transformation* occurs when bank deposits offer better risk-return characteristics than direct investments. This may occur when there are indivisibilities in the investment, in which case a small investor cannot diversify its portfolio. It may also occur in an asymmetric information situation, when banks have better information than depositors.

Finally, modern banks can be seen as transforming securities with short maturities, offered to depositors, into securities with long maturities, which borrowers desire. This *maturity transformation* function necessarily implies a risk, since the banks' assets will be illiquid, given the depositors' claims. Nevertheless, interbank lending and derivative financial instruments available to banks (swaps, futures) offer possibilities to limit this risk but are costly to manage for the banks' clients.

To clarify the distinction between the different functions performed by banks, it may be worth emphasizing that the three types of asset transformation that we are considering occur even in the absence of credit risk on the loans granted by the bank. A pawnbroker, a bank investing only in repos,[10] and a bank making only fully secured loans perform the three transformation functions we have mentioned: convenience of denomination, quality transformation, and maturity transformation.

## 1.4 Managing Risks

Usually, bank management textbooks define three sources of risk affecting banks: credit risk, interest rate risk, and liquidity risk.[11] These risks correspond to different lines in the banks' balance sheets. It is worth mentioning also the risks of off-balance-sheet operations, which have been soaring in the last two decades.[12] The following sections briefly sketch a historical account of the management of these different risks by banks. Chapter 8 offers a formal analysis of risk management in banks.

### 1.4.1 Credit Risk

When the first bank loans spread in Florence, Siena, and Lucca, and later in Venice and Genoa, lending was limited to financing the harvest that could be seen in the fields and appraised. Thus, credit risk was small. However, financing wars soon became an important part of banking activities.[13] Still, bankers tried to make their loans secure, either through collateral (jewels), through the assignment of rights (excise tax), or generally through the endorsement by a city (which could be sued in case of default, whereas kings could not be).

The riskiness of these loans seems to have increased through time. Initially banks used to arrange fully collateralized loans, an activity not intrinsically different from that of a pawnbroker. The change in the riskiness of bank loans can be traced back to the start of investment banking. Investment banking was performed by a different type of institution and was a different concept from traditional credit activity.[14] It introduced a different philosophy of banking because it involved advancing money to industry rather than being a simple lender and getting good guarantees. This implied making more risky investments and, in particular, buying stocks. This appraisal of risk on a loan is one of the main functions of modern bankers.

### 1.4.2 Interest Rate and Liquidity Risks

The asset transformation function of banks also has implications for their management of risks. Indeed, when transforming maturities or when issuing liquid deposits guaranteed by illiquid loans, a bank takes a risk. This is because the cost of funds—which depends on the level of short-term interest rates—may rise above the interest income, determined by the contractual interest rates of the loans granted by the bank. Even when no interest is paid on deposits, the bank may face unexpected withdrawals, which will force it to seek more expensive sources of funds. As a consequence, the bank will have to manage the combination of interest rate risk (due to the difference in maturity) and liquidity risk (due to the difference in the marketability of the claims issued and that of the claims held). The management of interest rate risk has become crucial for banks since the increase in the volatility of interest rates after the end of the Bretton-Woods fixed exchange system.

### 1.4.3 Off-Balance-Sheet Operations

In the 1980s competition from financial markets made it necessary for banks to shift to more value-added products, which were better adapted to the needs of customers. To do so, banks started offering sophisticated contracts, such as loan commitments, credit lines, and guarantees.[15] They also developed their offer of swaps, hedging contracts, and securities underwriting. From an accounting viewpoint, none of these operations corresponds to a genuine liability (or asset) for the bank but only to a conditional commitment. This is why they are classified as off-balance-sheet operations.

Different factors have fostered the growth of off-balance-sheet operations. Some are related to banks' desire to increase their fee income and to decrease their leverage; others are aimed at escaping regulation and taxes. Still, the very development of these services shows that nonfinancial firms now have a demand for more sophisticated, custom-made financial products.

Since banks have developed a know-how in managing risks, it is only natural that they buy and sell risky assets, whether or not they hold these assets on their balance sheets. Depending on the risk-return characteristics of these assets, banks may want to hedge their risk (that is, behave like someone who buys insurance) or, on the contrary, they may be willing to retain this risk (and take the position of someone who sells insurance). Given the fact that a bank's failure may have important externalities (see chapters 7 and 9), banking regulators must carefully monitor off-balance-sheet operations.

## 1.5 Monitoring and Information Processing

Banks have a specific part to play in managing some of the problems resulting from imperfect information on borrowers. Banks thus invest in the technologies that allow them to screen loan applicants and to monitor their projects.[16] According to Mayer (1988), this monitoring activity implies that firms and financial intermediaries develop long-term relationships, thus mitigating the effects of moral hazard.

This is clearly one of the main differences between bank lending and issuing securities in the financial markets. It implies that whereas bond prices reflect market information, the value of a bank loan results from this long-term relationship and is a priori unknown, both to the market and to the regulator.[17] In this sense we may say that bank loans are "opaque" (Merton 1993).

## 1.6 The Role of Banks in the Resource Allocation Process

Banks exert a fundamental influence on capital allocation, risk sharing, and economic growth (see Hellwig 1991). Gerschenkron (1962), in an early contribution, holds this influence to have been of capital importance for the development of some countries. Gerschenkron's position regarding the role of banks in economic growth and development has led to a continuing debate (Edwards and Ogilvie 1996). The historical importance of the impact of financial institutions on economic performance is still far from being well established. From a theoretical standpoint, the idea of "scarcity of funds" (which is difficult to capture in a general equilibrium model) could be useful in the study of economic development: underdeveloped economies with a low level of financial intermediation and small, illiquid financial markets may be unable to channel savings efficiently. Indeed, "large projects" that are essential to development, such as infrastructure financing, can be seen as unprofitable because of the high risk premia that are associated with them. This role of financial markets in economic development has now begun to be studied from a theoretical point of view, following in particular the contribution of Greenwood and Jovanovic (1990).[18]

Simultaneously, the fact that more bank-oriented countries such as Japan and Germany have experienced higher rates of growth in the 1980s has motivated additional research on the economic role of banks (Mayer 1988; Allen and Gale 1997). For instance, Allen and Gale (1995) closely examine the differences between the financial systems in Germany and in the United States.[19] They suggest that market-oriented economies are not very good in dealing with nondiversifiable risks: in the United States and Britain, for example, households hold around half of their assets in equities, whereas in bank-oriented economies such as Japan or Germany, households hold essentially safe assets. Banks' reserves work as a buffer against macroeconomic shocks and allow for better intertemporal risk sharing. The flip side of the coin is that bank-oriented economies are not very good at financing new technologies. Allen and Gale (2000) show that markets are much better for dealing with differences of opinion among investors about these new technologies.

## 1.7 Banking in the Arrow-Debreu Model

In order to explain the earlier statement that a microeconomic theory of banks could not exist before the foundations of the economics of information were laid (in the

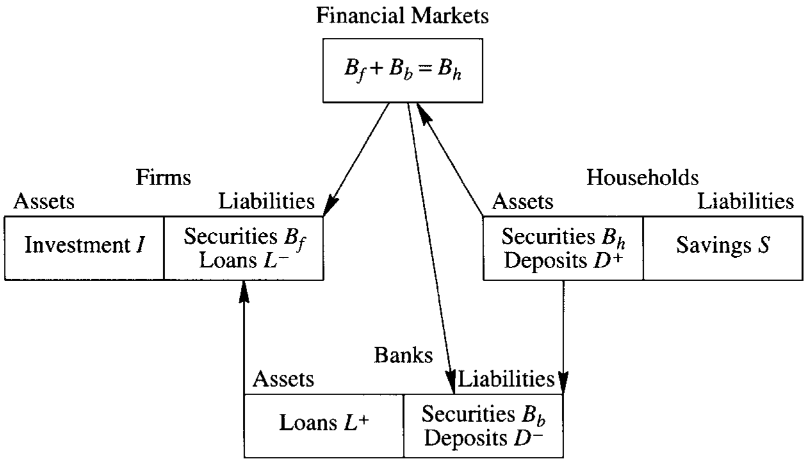



**Figure 1.1**
Financial decisions of economic agents.

early 1970s), this section presents a simple general equilibrium model à la Arrow-Debreu, extended to include a banking sector. To put things as simply as possible, the model uses a deterministic framework, although uncertainty could be introduced without any substantial change in the results, under the assumption of complete financial markets (Arrow 1953).

The financial decisions of economic agents in this simple model are represented in figure 1.1. Each type of agent is denoted by a particular subscript: $f$ for firms, $h$ for households, and $b$ for banks. For simplicity, the public sector (government and Central Bank) is omitted. A more complete diagram is presented in chapter 3 (fig. 3.1).

For simplicity, consider a two-dates model ($t = 1, 2$) with a unique physical good, initially owned by the consumers and taken as a numeraire. Some of it will be consumed at date 1, the rest being invested by the firms to produce consumption at date 2. All agents behave competitively. To simplify notations, the model assumes a representative firm, a representative consumer, and a representative bank.

### 1.7.1 The Consumer

The consumer chooses her consumption profile $(C_1, C_2)$, and the allocation of her savings $S$ between bank deposits $D_h$ and securities (bonds) $B_h$, in a way that maximizes her utility function $u$ under her budget constraints:

$$\mathscr{P}_h \begin{cases} \max u(C_1, C_2) \\ C_1 + B_h + D_h = \omega_1, & (1.1) \\ pC_2 = \Pi_f + \Pi_b + (1 + r)B_h + (1 + r_D)D_h, & (1.2) \end{cases}$$

where $\omega_1$ denotes her initial endowment of the consumption good, $p$ denotes the price of $C_2$, $\Pi_f$ and $\Pi_b$ represent respectively the profits of the firm and of the bank (distributed to the consumer-stockholder at $t = 2$), and $r$ and $r_D$ are the interest rates paid by bonds and deposits. Because, in this simplistic world, securities and bank deposits are perfect substitutes, it is clear that the consumer's program $(\mathscr{P}_h)$ has an interior solution only when these interest rates are equal:

$$r = r_D. \tag{1.3}$$

### 1.7.2 The Firm

The firm chooses its investment level $I$ and its financing (through bank loans $L_f$ and issuance of securities $B_f$) in a way that maximizes its profit:

$$\mathscr{P}_f \begin{cases} \max \Pi_f \\ \Pi_f = pf(I) - (1+r)B_f - (1+r_L)L_f, & (1.4) \\ I = B_f + L_f, & (1.5) \end{cases}$$

where $f$ denotes the production function of the representative firm and $r_L$ is the interest rate on bank loans. Again, because bank loans and bonds are here perfect substitutes, $\mathscr{P}_f$ has an interior solution only when

$$r = r_L. \tag{1.6}$$

### 1.7.3 The Bank

The bank chooses its supply of loans $L_b$, its demand for deposits $D_b$, and its issuance of bonds $B_b$ in a way that maximizes its profit:

$$\mathscr{P}_b \begin{cases} \max \Pi_b \\ \Pi_b = r_L L_b - rB_b - r_D D_b, & (1.7) \\ L_b = B_b + D_b. & (1.8) \end{cases}$$

### 1.7.4 General Equilibrium

General equilibrium is characterized by a vector of interest rates $(r, r_L, r_D)$ and three vectors of demand and supply levels—$(C_1, C_2, B_h, D_h)$ for the consumer, $(I, B_f, L_f)$ for the firm, and $(L_b, B_b, D_b)$ for the bank—such that

- each agent behaves optimally (his or her decisions solve $\mathscr{P}_h$, $\mathscr{P}_f$, or $\mathscr{P}_b$ respectively);
- each market clears

$I = S$ (good market)
$D_b = D_h$ (deposit market)
$L_f = L_b$ (credit market)
$B_h = B_f + B_b$ (bond market).

From relations (1.3) and (1.6) it is clear that the only possible equilibrium is such that all interest rates are equal:

$$r = r_L = r_D. \tag{1.9}$$

In that case, it is obvious from $\mathscr{P}_b$ that banks necessarily make a zero profit at equilibrium. Moreover, their decisions have no effect on other agents because households are completely indifferent between deposits and securities, and similarly firms are completely indifferent as to bank credit versus securities. This is the banking analogue of the Modigliani-Miller theorem (see, e.g., Hagen 1976) for the financial policy of firms.

**Result 1.1** If firms and households have unrestricted access to perfect financial markets, then in a competitive equilibrium:

- banks make a zero profit;
- the size and composition of banks' balance sheets have no effect on other economic agents.

This rather disappointing result extends easily to the case of uncertainty, provided financial markets are complete. Indeed, for each future state of the world $s$ ($s \in \Omega$), one can determine the price $p_s$ of the contingent claim that pays one unit of account in state $s$ and nothing otherwise. Now suppose a bank issues (or buys) a security $j$ (interpreted as a deposit or a loan) characterized by the array $x_s^j$ ($s \in \Omega$) of its payoffs in all future states of the world. By the absence of arbitrage opportunities, the price of security $j$ has to be

$$Z^j = \sum_{s \in \Omega} p_s x_s^j.$$

An immediate consequence is that all banks still make a zero profit, independent of the volume and characteristics of the securities they buy and sell. This explains why the general equilibrium model with complete financial markets *cannot* be used for studying the banking sector.

## 1.8 Outline of the Book

As we have just seen, the Arrow-Debreu paradigm leads to a world in which banks are redundant institutions. It does not account for the complexities of the banking industry. There are two complementary ways out of this disappointing result:

• The incomplete markets paradigm, which explains why financial markets *cannot* be complete and shows why banks (and more generally financial intermediaries) exist. This is the topic of chapter 2.

• The industrial organization approach to banking, which considers that banks essentially offer *services* to their customers (depositors and borrowers), and that financial transactions are only the visible counterpart to these services. As a consequence, the cost of providing these services has to be introduced, as well as some degree of product differentiation. This approach is studied in chapter 3.

In chapter 4 we explore in more detail the contractual relationship between a lender and a borrower. We examine the different considerations that influence the design of loan contracts: risk sharing, repayment enforcement, moral hazard, and adverse selection. In chapter 5 we study the credit market and explore the possible causes of equilibrium credit rationing. In chapter 6 we examine the macroeconomic consequences of financial imperfections. In chapter 7 we study the causes for the instability of the banking system. In chapter 8 we provide a formal analysis of the methods employed by bankers for managing the different risks associated with banking activities. Finally, we examine in chapter 9 the justifications and instruments of banking regulations.

## Notes

1. Even if it is recurrent, this lending activity, called trade credit, is only complementary to the core activity of these firms. For theoretical analyses of trade credit, see Biais and Gollier (1997) and Kiyotaki and Moore (1997).

2. Consider, for example, the title of the article by Gorton and Pennacchi (1993): "Money Market Funds and Finance Companies: Are They the Banks of the Future?"

3. The main reason is the famous argument of "double coincidence of wants" between traders.

4. For a theoretical analysis of commodity money, see Kiyotaki and Wright (1989; 1991).

5. Actually, a recent book by Cohen (1992) shows that in ancient Greece banks were already performing complex operations, such as transformation of deposits into loans. We thank Elu Von Thadden for indicating this reference to us.

6. When a bank failed, the bench was broken. This is the origin of the Italian word for *bankruptcy*, *bancarotta*, which means "the bench is broken."

7. It is customary to locate the origins of banking in England in the deposit activities of goldsmiths in the seventeenth century. Their capacity to deal with goldware and silverware made them into bankers. Still, as Kindleberger (1993) puts it, "The scriveners seem to have preceded the goldsmith as ones who accepted deposits. Needed to write out letters and contracts in a time of illiteracy, the scrivener became a skilled adviser, middleman, broker, and then lender who accepted deposits" (51).

8. Nevertheless, the need for the cities or the government to obtain cash could be such that the deposit bank could be forced to give credit to the city or to the king, as happened for the Taula de Canvi in Valencia and the Bank of Amsterdam. Also, Charles I of England confiscated the gold and silver that had been deposited in the Tower of London in 1640, and returned it only after obtaining a loan.

9. For an economic analysis of the risks involved in large payment interbank systems, see, for example, Rochet and Tirole (1996).

10. A repurchase agreement (repo) is a financial contract very similar to a fully collateralized short-term loan, the principal of which is fully guaranteed by a portfolio of securities (100 percent collateralization). For legal reasons, it is contractually implemented as if the borrower had sold balance sheet securities to the lender with a promise to buy them back later under specified conditions.

11. Two other sources of risk are not considered in this book: exchange rate risk, which affects banks involved in foreign exchange transactions, and operational risk, which concerns all financial institutions.

12. Note that these risks can also be decomposed into credit risk, interest rate risk, and liquidity risk.

13. This type of activity resulted in bankruptcy for several Italian bankers, such as the Bardi, the Peruzzi, and the Ricciardi (see, e.g., Kindleberger 1993).

14. In continental Europe the practice developed in the nineteenth century, with the Société Générale de Belgique or the Caisse Générale du Commerce et de l'Industrie (founded by Laffite in France).

15. We do not go into the details of these operations. The reader is referred to Greenbaum and Thakor (1995) for definitions and an analysis.

16. Screening and monitoring of projects can be traced back to the origins of banking, when bill traders identified the signatures of merchants and gave credit knowing the bills' quality, or even bought the bills directly (as in today's factoring activities).

17. Recent empirical studies (e.g., James 1987) have shown the importance of this specific role of banks.

18. More recently, Armendariz (1999) analyzes the role of government-supported financial institutions ("development banks") in less developed countries.

19. For another theoretical analysis of different banking systems, see Hauswald (1995).

## References


Allen, F., and D. Gale. 1994. *Financial innovation and risk sharing*. Cambridge, Mass.: MIT Press.

———. 1995. A welfare comparison of intermediaries in Germany and the U.S. *European Economic Review* 39 (2): 179–209.

———. 1997. Financial markets, intermediaries, and intertemporal smoothing. *Journal of Political Economy* 105 (3): 523–546.

———. 2000. *Comparing financial systems*. Cambridge, Mass.: MIT Press.

Armendariz de Aghion, B. 1999. Development banking. *Journal of Development Economics* 58: 83–100.

Arrow, K. 1953. *Le rôle des valeurs boursières pour la répartition la meilleure des risques*. Cahiers du Séminaire d'Econométrie. Paris.

Biais, B., and C. Gollier. 1997. Why do firms use trade credit: A signaling approach. *Review of Financial Studies* 10: 903–937.

Cohen, D. 1992. *Athenian economy and society: A banking perspective.* Princeton, N.J.: Princeton University Press.

Debreu, G. 1987. *Theory of value: An axiomatic analysis of economic equilibrium.* Cowles Foundation Monograph 17. New Haven, Conn.

Diamond, D. W., and R. G. Rajan. 2000. A theory of bank capital. *Journal of Finance* 55 (6): 2431–2465.

Edwards, J., and S. Ogilvie. 1996. Universal banks and German industrialization: A reappraisal. *Economic History Review* 49 (3): 427–446.

Flannery, M. J. 1994. Debt maturity and the deadweight cost of leverage: Optimally financing banking firms. *American Economic Review* 84: 320–331.

Gerschenkron, A. 1962. *Economic backwardness in historical perspective.* Cambridge, Mass.: Harvard University Press.

Gorton, G., and G. Pennacchi. 1993. Money market funds and finance companies: Are they the banks of the future? In *Structural change in banking*, ed. M. Klausner and L. White. New York: Irwin.

Greenbaum, S. I., and A. V. Thakor. 1995. *Contemporary financial intermediation*. Fort Worth, Texas: Dryden Press.

Greenwood, J., and B. Jovanovic. 1990. Financial development, growth and the distribution of income. *Journal of Political Economy* 98 (5): 1076–1107.

Gurley, J., and E. Shaw. 1960. *Money in the theory of finance.* Washington: Brookings Institution.

Hagen, K. P. 1976. Default risk, homemade leverage, and the Modigliani-Miller theorem: A note. *American Economic Review* 66 (1): 199–203.

Hauswald, R. 1995. Financial contracting, reorganization and mixed finance: A theory of banking systems. College Park: University of Maryland. Mimeograph.

Hellwig, M. 1991. Banking, financial intermediation and corporate finance. In *European financial integration*, ed. A. Giovannini and C. Mayer. Cambridge: Cambridge University Press.

James, C. 1987. Some evidence on the uniqueness of bank loans. *Journal of Financial Economics* 19 (2): 217–235.

Kindleberger, C. P. 1993. *A financial history of Western Europe*. Oxford: Oxford University Press.

Kiyotaki, N., and J. Moore. 1997. Credit-chains. London School of Economics. Mimeograph.

Kiyotaki, N., and R. Wright. 1989. On money as a medium of exchange. *Journal of Political Economy* 97: 927–954.

———. 1991. A contribution to the pure theory of money. *Journal of Economic Theory* 53 (2): 215–235.

Kohn, M. 1999. Early deposit banking. Working paper 99-03, Dartmouth College, Hanover, N.H.

Mayer, C. 1988. New issues in corporate finance. *European Economic Review* 32 (5): 1167–1183.

McAndrews, J., and W. Roberds. 1999. Payment intermediation and the origins of banking. Working paper, Federal Reserve Bank of Atlanta.

Merton, R. C. 1993. Operation and regulation in financial intermediation: A functional perspective. In *Operation and regulation of financial markets*, ed. P. Englund. Stockholm: Economic Council.

Qi, J. 1998. Deposit liquidity and bank monitoring. *Journal of Financial Intermediation* 7 (2): 198–218.

Rajan, R. G. 1998. The past and future of commercial banking viewed through an incomplete contract lens. *Journal of Money, Credit and Banking* 30: 524–550.

Rochet, J. C., and J. Tirole. 1996. Controlling risks in payments systems. *Journal of Money, Credit and Banking* 28 (4): 832–862.

# 2 The Role of Financial Intermediaries

Although this book is specifically focused on banks, this chapter adopts a broader perspective and studies financial intermediaries (FIs) in general. The first definition of an FI that may come to mind is that of *an economic agent who specializes in the activities of buying and selling (at the same time) financial claims*. This is analogous to the notion of *intermediary* (or retailer) in the theory of industrial organization as an agent who buys certain goods or services from producers and sells them to final consumers. The justification given by the theory of industrial organization for the existence of such intermediaries is the presence of *frictions* in transaction technologies (e.g., transportation costs). *Brokers* and *dealers*, operating on financial markets, are a clear example of such intermediaries in the financial sector. This paradigm can also provide a (simplistic) description of banking activities. Roughly speaking, banks can be seen as retailers of financial securities: they buy the securities issued by borrowers (i.e., they grant loans), and they sell them to lenders (i.e., they collect deposits).[1]

However, banking activities are in general more complex, for at least two reasons:

- Banks usually deal (at least partially) with financial *contracts* (loans and deposits), which cannot be easily resold, as opposed to financial *securities* (stocks and bonds), which are anonymous (in the sense that the identity of the holder is irrelevant) and thus easily marketable. Therefore, banks typically must *hold* these contracts in their balance sheets until the contracts expire.[2] (This is also true to some extent for insurance companies.)
- The characteristics of the contracts or securities issued by firms (borrowers) are usually different from those of the contracts or securities desired by investors (depositors). Therefore, as first argued by Gurley and Shaw (1960), and more recently by Benston and Smith (1976) and Fama (1980), banks (and also mutual funds and insurance companies) are there to *transform* financial contracts and securities.

Of course, in the ideal world of frictionless and complete financial markets, both investors and borrowers would be able to diversify perfectly and obtain optimal risk sharing. But as soon as one introduces indivisibilities (even small) and nonconvexities in transaction technologies, perfect diversification is no longer feasible and FIs are needed. This transaction costs approach (see section 2.1) does not in fact contradict the assumption of (approximately) complete markets. For instance, as argued by Hellwig (1991), the role of insurance companies is that of mutualizing idiosyncratic risks so that insured persons obtain approximately the same diversification as they would under complete markets.[3] A similar description could be given of mutual funds' activity. FIs can therefore be seen as *coalitions* (mutuals) of individual lenders or borrowers who exploit economies of scale or economies of scope in the transaction technology. As a result of the activities of FIs, individuals obtain almost perfect diversification.[4]

Of course, this approach is not completely satisfactory because these transaction costs are given exogenously. The nature of these costs must be explored. Even if physical and technological costs may have played a historical role in the emergence of FIs, the progress experienced recently in telecommunications and computers, as well as the related development of sophisticated financial instruments, implies that FIs would be bound to disappear if another, more fundamental form of transaction costs were not present. Therefore, the subject of *informational asymmetries*—whether ex ante (adverse selection), interim (moral hazard), or ex post (costly state verification)—is further explored in several sections of this book. These asymmetries generate market imperfections that can be seen as specific forms of transaction costs. These costs can be partially overcome by institutions that can be interpreted as FIs.

Section 2.2 discusses how the role of banks in providing liquidity insurance is related to these informational asymmetries. Following Diamond and Dybvig (1983), the discussion considers banks as "pools of liquidity" or "coalitions of depositors" that provide households with insurance against idiosyncratic liquidity shocks, supposedly privately observed. Alternatively, as emphasized by Allen and Gale (1997), depositors may obtain insurance against adverse market conditions.

Section 2.3 explores another interpretation of FIs as information-sharing coalitions. For example, when individual borrowers (firms) have private information on the characteristics of the projects they wish to finance, the competitive equilibrium can be inefficient (as discussed in Akerlof 1970). As shown by Leland and Pyle (1977), this problem can be partially overcome if firms can use their level of retained equity as a *signal* to investors (an adaptation of the theory developed by Spence 1973 for the job market). However, this signal has a cost because firms cannot obtain perfect risk sharing.[5] This cost—the informational cost of capital—can be seen as an *informational transaction cost*. Elaborating on Leland and Pyle (1977), Diamond (1984) and Ramakrishnan and Thakor (1984) were able to show that, under certain conditions, economies of scale were present. In other words, if firms are able to form coalitions (without internal communication problems), then the cost of capital per firm is a decreasing function of the number of firms in the coalition (size of the intermediary). Still in the context of adverse selection, coalitions of heterogeneous borrowers can also improve the market outcome by providing cross-subsidization inside the coalitions. An example is studied in Boyd and Prescott (1986).

Section 2.4 discusses the *delegated monitoring* theory of intermediation, first explored by Diamond (1984). The section uses the term more broadly than Diamond did, to refer to any activity aimed at preventing opportunistic behavior of the borrower, both interim (moral hazard) and ex post (costly state verification).

Monitoring typically involves increasing returns to scale, which implies that it is more efficiently performed by specialized firms. Therefore, individual lenders tend to delegate the monitoring activity instead of performing it themselves. This introduces a new problem: the information that the monitor provides may not be reliable (as modeled in Campbell and Kracaw 1980). Thus the monitor has to be given incentives to do the job properly. FIs can be seen as providing solutions to this incentive problem.[6] Several theories have been put forward:

- Diamond (1984) suggests that if investors can impose nonpecuniary penalties on a monitor who does not perform well, the optimal arrangement will look like a deposit contract. Moreover, by diversifying the loan portfolio, the monitor (interpreted as a banker) can make the cost of delegation as small as possible, getting close to offering riskless deposits.
- Calomiris and Kahn (1991) argue that demand deposits provide the adequate instrument for disciplining bank managers: if anything goes wrong, investors withdraw their deposits.[7]
- Holmström and Tirole (1997) invoke the financial involvement of the monitor in the project: outside investors require that the monitor participate in the financing. This gives rise to informational economies of scope between monitoring and lending activities, and explains the role of banking capital.

The early literature on the foundations of financial intermediation could not explain the coexistence of FIs and markets. In Diamond (1984) the increasing returns in the monitoring technology implied that a monopoly bank should emerge and replace financial markets.[8] In Diamond and Dybvig (1983) the existence of financial markets typically impairs the provision of liquidity insurance by deposit banks. A second generation of models has focused on the coexistence of financial markets and FIs. These models have in particular explained the financing choice of firms between issuing securities in the financial markets (direct finance) or borrowing from a bank (monitored finance). Different (complementary) explanations of this choice have been analyzed formally in the literature. According to these models, the firms who choose direct finance can be those with the best reputation (Diamond

1991), the highest level of collateral (Holmström and Tirole 1997), the best technology (Boot and Thakor 1997), or the best credit rating (Bolton and Freixas 2000). These models are presented in section 2.5.

## 2.1 Transaction Costs

The simplest way to justify the existence of FIs is to emphasize the difference between their inputs and their outputs and to view their main activity as transformation of financial securities. For example, banks transform deposits of convenient maturity, such as demand deposits without any restriction on minimum amount and with low risk, into loans with a longer maturity, in larger amounts, and with credit risk. FIs may thus be viewed as providing services of assets transformation. Attractive as it may be, *this scenario fails to explain why this assets transformation is not done by the borrowers themselves*. A consistent model must include the assumptions of economies of scale or economies of scope that make it profitable for separate units (the banks) to specialize in transforming the financial assets issued by the borrowers. These economies of scale or scope come from the existence of transaction costs,[9] which can be monetary but also include search costs[10] as well as monitoring and auditing costs (see sections 2.3 and 2.4).

The following sections present some of the classical transaction cost justifications of FIs by clarifying the implicit assumptions that each of them requires.

### 2.1.1 Economies of Scope

As mentioned in chapter 1, a primitive form of banking involved money changers who decided to offer deposit services because they had a comparative advantage in storing valuables. Having already a need for safekeeping places for their own inventories of coins and metals, they could easily offer analogous services to merchants and traders. Using modern vocabulary, we would say that economies of scope existed between money-changing and safekeeping activities. However, this explanation is incomplete because the economies of scope mentioned concern essentially payment and deposit services. To explain the development of commercial banking, economies of scope must exist also between deposit and credit activities. Although frequently alluded to, these economies of scope are not easy to pinpoint, either at the empirical or the theoretical level. It is true that in a location model, in which agents are geographically dispersed and face transportation costs, it is efficient for the same firm or the same branch to offer deposit and credit services in a single location. Similarly, the same clerk is more efficiently employed if he or she takes care simultaneously of customers' checking accounts and loan repayments. However, the same argument would also hold for any kind of services or activities; it is the "central place" story, which explains the existence of department stores or trade centers.

Something deeper must be involved in the economies-of-scope explanation of financial intermediation. A first explanation is given by portfolio theory. As discussed in chapter 8, if some investors are much less risk-averse than the others, these investors will in equilibrium short-sell (borrow) the riskless asset and invest more than their own wealth in the risky market portfolio. In a sense these investors have a comparative advantage in holding risky assets. A second source of scope economies is banks' expertise in managing liquidity risk, which allows them to offer credit lines as well as deposit services. This view has been put forward by Kashyap, Rajan, and Stein (1999). A third explanation, also given by portfolio theory, is diversification. If a positive correlation exists between the returns of two categories of securities, one having a positive expected excess return (over the riskless asset) and the other a negative expected excess return, the typical investor will hold a long position in the first one and a short position in the second one. If we call these investors *banks*, the first security *loans*, and the second one *deposits*, we have a diversification theory of financial intermediation. This theory, advanced by Pyle (1971), is explained in detail in chapter 8.

However, these portfolio theories of financial intermediation are not completely satisfactory; because of limited liability it is not possible to assimilate a deposit offered by an FI and a short position in a riskless asset (unless deposits are fully insured; see chapter 9). Similarly, the specificity of banks and insurance companies (as opposed to mutual funds) is that they deal essentially with nonmarketable securities: loans and insurance contracts. Therefore, another approach is needed for explaining economies of scope between, say, credit and deposit activities. This is where information asymmetries come in. If lenders have doubts on the credit worthiness of borrowers, they will trust more those borrowers that they know better (for instance, because they manage the borrowers' checking accounts and security portfolios).[11] Similarly, if depositors are uncertain about the true value of risky projects, they may agree to participate in the financing of these projects if they know that their banker has a personal stake in them. These issues are discussed in detail in the rest of this chapter.

### 2.1.2 Economies of Scale

Of course, an obvious justification for intermediation is the presence of fixed transaction costs, or more generally, increasing returns in the transaction technology. For instance, if a fixed fee is associated with any financial transaction, depositors or borrowers will tend to form coalitions and buy or sell together in order to divide the transaction costs. (This argument does not work with proportional transaction costs.) Similarly, because of indivisibilities, a coalition of investors will be able to hold a more diversified (and thus less risky) portfolio than the ones individual investors would hold on their own.

Another type of scale economy is related to liquidity insurance à la Diamond and Dybvig (see section 2.2 and chapter 7). By the law of large numbers, a large coalition of investors will be able to invest in illiquid but more profitable securities while preserving enough liquidity to satisfy the needs of individual investors. This argument is not specific to the banking industry but also valid for insurance activities and, more generally, for inventory management. To have a genuine specificity of banks (as opposed to other intermediaries) informational asymmetries must again be introduced. This is done in section 2.3 in the discussion of the signaling approach, originally advanced by Leland and Pyle (1977). These informational asymmetries are also crucial for explaining the superiority of banks over financial markets in the provision of liquidity insurance.

## 2.2 Coalitions of Depositors and Liquidity Insurance

A very natural idea for justifying the existence of depository institutions is to consider them as pools of liquidity that provide households with insurance against idiosyncratic shocks that affect their consumption needs. As long as these shocks are not perfectly correlated, the total cash reserve needed by a bank of size $N$ (interpreted as a coalition of $N$ depositors) increases less than proportionally with $N$. This is the basis for the fractional reserve system, in which some fraction of the deposits can be used to finance profitable but illiquid investments. However, this is also the source of a potential fragility of banks, in the event that a high number of depositors decide to withdraw their funds for reasons other than liquidity needs. An interesting modeling of these issues by Diamond and Dybvig (1983) is presented in detail in chapter 7. For the moment, a simplified version of this model is presented in order to capture the notion of liquidity insurance that was initially modeled by Bryant (1980).

### 2.2.1 The Model

Consider a one-good, three-dates economy in which a continuum of ex ante identical agents is each endowed with one unit of good at time $t = 0$. This good is to be consumed at dates $t = 1$ or $t = 2$. $C_t$ denotes consumption at date $t$. The simplest way to model liquidity shocks is to consider that consumers learn at $t = 1$ whether they will have to consume *early* (at $t = 1$, and the agent is said to be "of type 1" or "impatient"), in which case their utility is $u(C_1)$; or *late* (at $t = 2$, and the agent is said to be "of type 2" or "patient"), in which case their utility is $u(C_2)$. The utility function $u$ is assumed to be increasing and concave. For simplicity, there is no discounting. In ex ante terms the expected utility of a depositor is

$$U = \pi_1 u(C_1) + \pi_2 u(C_2), \tag{2.1}$$

where $\pi_1$ (resp. $\pi_2$) is the probability of being of "type 1" (resp. type 2).[12]

The good can be stored from one period to the next, or can be invested in an amount $I$, $0 \leq I \leq 1$, at $t = 0$, in a long-run technology. This technology provides $R > 1$ units of consumption at $t = 2$ but only $\ell < 1$ units of consumption if it has to be liquidated at $t = 1$. The following discussion compares different institutional arrangements and shows that a depository institution can improve the efficiency of the economy.

### 2.2.2 Characteristics of the Optimal Allocation

From an ex ante viewpoint, there is a unique symmetric[13] Pareto-optimal allocation $(C_1^*, C_2^*)$ that is easily obtained by computing

$$\max \pi_1 u(C_1) + \pi_2 u(C_2) \tag{2.2}$$

under the constraints

$$\pi_1 C_1 = 1 - I \quad \text{and} \quad \pi_2 C_2 = RI,$$

which, by eliminating $I$, can be aggregated into a single constraint:

$$\pi_1 C_1 + \pi_2 \frac{C_2}{R} = 1, \tag{2.3}$$

This optimal allocation satisfies the first-order condition:

$$u'(C_1^*) = Ru'(C_2^*), \tag{2.4}$$

expressing that the marginal rate of substitution between consumptions at dates 2 and 1 has to equal the return on the long-run technology.

We now turn to the study of different institutional arrangements and the allocations they generate.

### 2.2.3 Autarky

The simplest case, in which there is no trade between agents, is called autarky. Each agent chooses independently the quantity $I$ that will be invested in the illiquid technology, assumed to be perfectly divisible. If he has to consume early, then this investment will be liquidated at $t = 1$, yielding

$$C_1 = 1 - I + \ell I = 1 - I(1 - \ell). \tag{2.5}$$

On the contrary, if he has to consume late, he obtains

$$C_2 = 1 - I + RI = 1 + I(R - 1). \tag{2.6}$$

In autarky each consumer will select the consumption profile that maximizes his ex ante utility $U$, given by (2.1), under the constraints (2.5) and (2.6). Notice that $C_1 \leq 1$, with equality only when $I = 0$. Also $C_2 \leq R$, with equality only when $I = 1$. Efficiency is thus not reached because

$$\pi_1 C_1 + \pi_2 \frac{C_2}{R} < 1.$$

### 2.2.4 Market Economy

If agents are allowed to trade, welfare improves. In this simple context, it is enough to open at $t = 1$ a financial market in which agents can trade the good at $t = 1$ against a riskless bond[14] (that is, a promise to receive some quantity of the consumption good at $t = 2$). Let $p$ denote the price at $t = 1$ of the bond which, by convention, yields one unit of good at $t = 2$. Clearly, $p \leq 1$; otherwise people would prefer to store. By investing $I$ at $t = 0$, an agent can now obtain

$$C_1 = 1 - I + pRI, \tag{2.7}$$

if she needs to consume early (in which case she will sell $RI$ bonds). If, on the contrary, she needs to consume late, she will obtain

$$C_2 = \frac{1 - I}{p} + RI = \frac{1}{p}(1 - I + pRI), \tag{2.8}$$

since she can then buy $(1 - I)/p$ bonds at $t = 1$. $I$ is chosen ex ante by each agent in order to maximize her expected utility $\pi_1 u(C_1) + \pi_2 u(C_2)$. Notice that

$$C_2 = \frac{C_1}{p},$$

and both are linear functions of $I$. Since $I$ can be freely chosen by agents, the only possible interior equilibrium price is

$$p = \frac{1}{R}.$$

Otherwise either an excess supply or an excess demand of bonds will occur:

$$I = \begin{cases} 1 & \text{if } p > \dfrac{1}{R}, \\[2ex] 0 & \text{if } p < \dfrac{1}{R}. \end{cases}$$

The equilibrium allocation of the market economy is therefore $C_1^M = 1$, $C_2^M = R$, and the corresponding investment level is $I^M = \pi_2$. Notice that this market allocation Pareto-dominates the autarky allocation because there is no liquidation. Still, except in the very peculiar case in which

$$u'(1) = Ru'(R),$$

the market allocation $(C_1^M = 1, C_2^M = R)$ is not Pareto-optimal because condition (2.4) is not satisfied.

In particular, Diamond and Dybvig (1983) assume that $C \to Cu'(C)$ is decreasing.[15] In that case, since $R > 1$, we have

$$Ru'(R) < u'(1), \tag{2.9}$$

and the market allocation can be Pareto-improved by increasing $C_1^M$ and decreasing $C_2^M$:

$$C_1^M = 1 < C_1^*; \qquad C_2^M = R > C_2^*. \tag{2.10}$$

Thus the market economy does not provide perfect insurance against liquidity shocks and therefore does not lead to an efficient allocation of resources. This is because individual liquidity shocks are not publicly observable, and securities contingent on these shocks cannot be traded, leading to a problem of incomplete financial markets. The following discussion shows how a financial intermediary can solve this problem.

The reason that market allocation is not Pareto-optimal is that complete Arrow-Debreu contingent markets cannot exist; the state of the economy (the complete list of the consumers who need to consume early) is not observable by anyone. The only (noncontingent) financial market that can be opened (namely, the bond market) is not sufficient to obtain efficient risk sharing.

### 2.2.5 Financial Intermediation

The Pareto-optimal allocation $(C_1^*, C_2^*)$ characterized in section 2.2.2 can be implemented very easily[16] by a financial intermediary who offers a deposit contract stipulated as follows: in exchange for a deposit of one unit at $t = 0$, individuals can get either $C_1^*$ at $t = 1$ or $C_2^*$ at $t = 2$. In order to fulfill its obligations, the FI stores $\pi_1 C_1^*$ and invests $I = 1 - \pi_1 C_1^*$ in the illiquid technology. Thus we have established the following:

**Result 2.1** In an economy in which agents are individually subject to independent liquidity shocks, the market allocation can be improved by a deposit contract offered by a financial intermediary.

The balance sheet of the bank that offers the optimal deposit contract is very simple (fig. 2.1). Bank capital is not needed because liquidity shocks are perfectly diversifiable and loans are not risky.

A crucial assumption is that no individual withdraws at $t = 1$ if he or she does not have to. This assumption is not unreasonable, since it corresponds to a Nash

| Assets | Liabilities |
| --- | --- |
| Reserves $\pi_1 C_1^*$<br>Investment $I$ | Deposits $D = 1$ |

**Figure 2.1**
Bank balance sheet in Bryant-Diamond-Dybvig paradigm.

equilibrium behavior. Indeed, recall the first-order condition (2.4): $u'(C_1^*) = Ru'(C_2^*)$. Since $u'$ is decreasing and implies $R > 1$, it implies that $C_1^* < C_2^*$. In other words, a deviation by a single patient consumer (withdraw at $t = 1$ and store the good until $t = 2$) is never in that consumer's own interest. However, another Pareto-dominated Nash equilibrium exists (bank run equilibrium) in which all patient consumers withdraw (see chapter 7).

Note that an FI cannot coexist (in this simple setup) with a financial market. Indeed, if there is a bond market at $t = 1$, the equilibrium price is necessarily $p = 1/R$. Then the optimal allocation $(C_1^*, C_2^*)$ is not a Nash equilibrium anymore; (2.10) implies that

$$RC_1^* > R > C_2^*,$$

which means that late consumers are better off withdrawing early and buying bonds. This is of course a serious weakness of the model. Von Thadden (1996; 1997; 2004) has studied this question in a more general formulation that is discussed in chapter 7.

## 2.3 Coalitions of Borrowers and the Cost of Capital

The common assumption for all the models presented in this section is that entrepreneurs are better informed than investors about the quality of the projects they want to develop. This hidden information, or adverse selection, paradigm is explored in detail later. The current discussion demonstrates that this adverse selection paradigm can generate scale economies in the borrowing-lending activity, allowing interpretation of FIs as information-sharing coalitions. After a basic model of capital markets with adverse selection is introduced in section 2.3.1, the seminal contribution of Leland and Pyle (1977) is discussed in section 2.3.2. Leland and Pyle consider that entrepreneurs can signal the quality of their projects by investing more or less of their own wealth into them. In this way, they can partly overcome the adverse selection problem because "good" projects can be separated from "bad" projects by their level of self-financing. However, if entrepreneurs are risk-averse, this signaling is costly because "good" entrepreneurs are obliged to retain a fraction of the risk of their projects instead of obtaining full insurance on the financial markets. Leland and

Pyle then study coalitions of borrowers and show that the signaling cost increases less rapidly than the size of the coalition. In other words, if borrowers form partnerships, which Leland and Pyle interpret as FIs, they are able to obtain better financing conditions than by borrowing individually. This property is explained in section 2.3.3, and several related contributions are summarized in section 2.3.4.

### 2.3.1 A Simple Model of Capital Markets with Adverse Selection

The following model of competitive capital markets with adverse selection is used in several sections of this book. Consider a large number of entrepreneurs, each endowed with a risky project, requiring a fixed investment of a size normalized to 1 and yielding a random (gross) return denoted $\tilde{R}(\theta) = 1 + \tilde{r}(\theta)$. The net returns $\tilde{r}(\theta)$ of these investments follow a normal distribution of mean $\theta$ and variance $\sigma^2$. Whereas $\sigma^2$ is the same for all projects, $\theta$ differs across projects and is privately observed by each entrepreneur. However, the statistical distribution of $\theta$ in the population of entrepreneurs is common knowledge. The investors are risk-neutral and have access to a costless storage technology. The entrepreneurs have enough initial wealth $W_0$ to finance their projects ($W_0 > 1$), but they would prefer to sell these projects because they are risk-averse. They have an exponential Von Neumann–Morgenstern utility function $u(w) = -e^{-\rho w}$, where $w$ denotes their final wealth and $\rho > 0$ is their (constant) absolute index of risk aversion. If $\theta$ were observable, each entrepreneur would sell its project to the market at a price[17] $P(\theta) = E[\tilde{r}(\theta)] = \theta$ and would be perfectly insured.[18] The final wealth of an entrepreneur of type $\theta$ would be $W_0 + \theta$.

Suppose now that $\theta$ is private information and that entrepreneurs are indistinguishable by investors. As in Akerlof's market for lemons (1970), the price $P$ of equity will be the same for all firms, and in general only entrepreneurs with a lower expected return will sell their projects. Indeed, by self-financing its project, entrepreneur $\theta$ obtains[19]

$$Eu(W_0 + \tilde{r}(\theta)) = u\left(W_0 + \theta - \frac{1}{2}\rho\sigma^2\right),$$

whereas by selling it to the market, he obtains $u(W_0 + P)$. Therefore, entrepreneur $\theta$ will go to the financial market if and only if

$$\theta < \hat{\theta} = P + \frac{1}{2}\rho\sigma^2. \tag{2.11}$$

This means that only those entrepreneurs with a relatively low expected return ($\theta < \hat{\theta}$) will issue equity. This is exactly the adverse selection problem: by opening a financial market, investors select the low-quality entrepreneurs ($\theta < \hat{\theta}$) instead of the best ones ($\theta \geq \hat{\theta}$), who choose not to participate.

At equilibrium, the average return on equity will be equal to $P$ (because investors are risk-neutral):

$$P = E[\theta | \theta < \hat{\theta}]. \tag{2.12}$$

The equilibrium of the capital market with adverse selection is thus characterized by a price of equity $P$ and a cutoff level $\hat{\theta}$ such that relations (2.11) and (2.12) are satisfied. In general, the equilibrium outcome is inefficient. Assume, for instance, that the distribution of $\theta$ is binomial.[20] In other words, $\theta$ can take only two values: a low value $\theta_1$ with probability $\pi_1$, and a high value $\theta_2$ with probability $\pi_2$. Since investors are risk-neutral and entrepreneurs are risk-averse, first best efficiency requires that all entrepreneurs obtain 100 percent outside finance. By definition of the cutoff level, this means that $\hat{\theta} \geq \theta_2$. In that case, the price of equity equals

$$P = E[\theta] = \pi_1\theta_1 + \pi_2\theta_2.$$

Using (2.11) we obtain that this is only possible when

$$\pi_1\theta_1 + \pi_2\theta_2 + \frac{1}{2}\rho\sigma^2 \geq \theta_2,$$

or

$$\pi_1(\theta_2 - \theta_1) \leq \frac{1}{2}\rho\sigma^2. \tag{2.13}$$

In other words, the risk premium has to outweigh the adverse selection effect. If (2.13) is not satisfied, some entrepreneurs prefer to self-finance, and the equilibrium outcome is inefficient.[21]

### 2.3.2 Signaling through Self-Financing and the Cost of Capital

When (2.13) is not satisfied, the entrepreneurs who are endowed with good-quality projects ($\theta = \theta_2$) prefer to self-finance rather than to sell the entirety of their projects at a low price $P = E[\theta]$. In fact, they can limit themselves with *partial* self-finance if they can convince investors that the other entrepreneurs (who are endowed with low-quality projects, $\theta = \theta_1$) have no interest in doing the same (to "mimic" them, in the terminology of adverse selection models). In other words, deciding to self-finance a fraction $\alpha$ of the project will in that case signal to outside investors that this project is good. Intuitively, this is true when $\alpha$ is large enough. The "no mimicking" condition is

$$u(W_0 + \theta_1) \geq Eu(W_0 + (1 - \alpha)\theta_2 + \alpha\tilde{r}(\theta_1)). \tag{2.14}$$

The left side of (2.14) is the utility of a type $\theta_1$ entrepreneur who sells all his project at a low price $P_1 = \theta_1$. The right side represents his expected utility when he mimics type $\theta_2$, that is, sells only a fraction $(1-\alpha)$ of his project at a high unit price $P_2 = \theta_2$, but retains the risk on the remaining fraction $\alpha$. With normal returns and exponential utility this expected utility equals

$$u\left(W_0 + (1-\alpha)\theta_2 + \alpha\theta_1 - \frac{1}{2}\rho\sigma^2\alpha^2\right),$$

which gives a simplified version of (2.14):

$$\theta_1 \geq (1-\alpha)\theta_2 + \alpha\theta_1 - \frac{1}{2}\rho\sigma^2\alpha^2,$$

or

$$\frac{\alpha^2}{1-\alpha} \geq \frac{2(\theta_2-\theta_1)}{\rho\sigma^2}. \tag{2.15}$$

**Result 2.2 (Leland and Pyle 1977)** When the level of projects' self-financing is observable, there is a continuum of signaling equilibria, parameterized by a number $\alpha$ fulfilling (2.15),[22] and characterized by a low price of equity $P_1 = \theta_1$ for entrepreneurs who do not self-finance and a high price of equity $P_2 = \theta_2$ for entrepreneurs who self-finance a fraction $\alpha$ of their projects.

As is usual in signaling models (see Spence 1973), there is a continuum of equilibria, parameterized by the level $\alpha$ of self-financing by good-quality entrepreneurs. These equilibria can be Pareto-ranked because all lenders break even and $\theta_1$ entrepreneurs get the same outcome as in the full-information case. As for $\theta_2$ entrepreneurs, they obtain a utility level of

$$u\left(W_0 + \theta_2 - \frac{1}{2}\rho\sigma^2\alpha^2\right),$$

instead of $u(W_0 + \theta_2)$ in the full-information case. Expressed in terms of lost income, their informational cost of capital is therefore

$$C = \frac{1}{2}\rho\sigma^2\alpha^2, \tag{2.16}$$

which is increasing in the level $\alpha$ of self-financing. The Pareto-dominating signaling equilibrium corresponds to the minimum possible value of $\alpha$, which is defined implicitly by transforming (2.15) into an equality:

$$\frac{\alpha^2}{1-\alpha} = \frac{2(\theta_2-\theta_1)}{\rho\sigma^2}. \tag{2.17}$$

It is natural to focus on this Pareto-dominating equilibrium, which allows definition of the (minimum) cost of capital

$$C(\sigma) = \frac{1}{2}\rho\sigma^2\alpha^2(\sigma), \tag{2.18}$$

where $\alpha(\sigma)$ is defined implicitly by (2.17).

### 2.3.3 Coalitions of Borrowers

Here we illustrate the main idea of this section, namely, that in the presence of adverse selection, coalitions of borrowers can do better than individual borrowers. Suppose that $N$ identical entrepreneurs of type $\theta_2$ form a partnership and collectively issue securities in order to finance their $N$ projects. If the individual returns of each project are independently distributed, and if the $N$ entrepreneurs share equally both the proceeds of security issuing and the final returns, the situation is formally the same as before: the expected return per project is still $\theta_2$, but (and this is the only difference) the variance per project is now $\sigma^2/N$ (because of diversification). Since the function $\sigma \to C(\sigma)$ defined by (2.18) is increasing, the following result is obtained:

**Result 2.3 (Diamond 1984)** In the Leland-Pyle model (1977), the unit cost of capital decreases with the size of the coalition of borrowers.

***Proof*** It is necessary only to prove that $\sigma \to C(\sigma)$ is increasing. But relation (2.17) implies both that $\sigma \to \alpha(\sigma)$ is decreasing (since

$$\alpha \to \frac{\alpha^2}{1-\alpha}$$

is increasing on $[0, 1]$) and that

$$\frac{1}{2}\rho\sigma^2\alpha^2(\sigma) = (\theta_2 - \theta_1)(1 - \alpha(\sigma)).$$

Therefore, (2.18) shows that $\sigma \to C(\sigma)$ is increasing, which was to be proved. ∎

Holding companies can be interpreted as such coalitions of borrowers. They can be considered as financial intermediaries who issue equity on the financial markets and invest in several subsidiaries (the $N$ entrepreneurs of our model). The cost of financing is lower for the holding company than for the subsidiaries.[23]

### 2.3.4 Suggestions for Further Reading

The Leland and Pyle (1977) model justifies FIs by considering the benefits obtained by borrowers when they form coalitions, provided they are able to communicate truthfully the quality of their projects within the coalition. But the framework of adverse selection (where the quality of projects is observable only by some investors) is sufficiently rich to study other possible justifications of FIs by coalition formation.

An agent endowed with private information faces two types of problems in order to benefit from this information. First, if she tries to sell her information directly, she will be confronted with a classic credibility problem: the potential buyers may not be convinced that the information is true. Second, the profits she might obtain through trading on her information might be too small with respect to the cost of obtaining this information. These profits might even be zero if prices are fully revealing, leading to the well-known paradox of Grossman and Stiglitz (1980). Campbell and Kracaw (1980) and Allen (1990) have studied the incentive issues associated with this problem and how they could be solved by the creation of FIs.

Ramakrishnan and Thakor (1984) have discovered another form of scale economies caused by asymmetric information, similar to the one already discussed. They study the case of security analysts who produce some information that is valuable to a risk-neutral principal (investor). This principal observes a signal that is positively correlated to the effort spent by the analyst in producing the information. Ramakrishnan and Thakor (see problem 2.9.3) compute the optimal contract (incentive scheme) between the principal and the analyst (who is risk-averse). Their crucial result is that if the analysts are able to collude (form coalitions), sign separate contracts with different principals, and mutualize their remunerations, they increase their total expected surplus. Millon and Thakor (1985) consider a variant of this model, with two additional elements: *information reusability* (the fact that gathering information about one project provides some information about similar projects) and *internal communication problems* (which tend to limit the optimal size of FIs).

Boyd and Prescott (1986) consider an economy with two types of agents (entrepreneurs) endowed with either a good or a bad project. Each entrepreneur knows the quality of his project. In a perfect information setting it would be optimal to implement all the good projects, implement some of the bad projects, and have the rest of the agents who are endowed with bad projects invest in the good projects. The stock market cannot implement the optimum because agents who have bad projects and want them to be profitable have no incentive to reveal their type. Nevertheless, a coalition of agents (an FI) could do better because the group could allow for cross-subsidization, decreasing the returns for good types and increasing the returns for bad types in such a way that each agent has an incentive to reveal truthfully the characteristic of his project. In this way, coalitions of heterogeneous agents can improve the market equilibrium outcome. This is a standard phenomenon in economies with adverse selection because the equilibrium outcome can be inefficient, even in a second-best sense, when incentive compatibility constraints are introduced (see, e.g., Rothschild and Stiglitz 1976).

Finally, Gorton and Pennacchi (1990) emphasize the qualitative asset transformation activity of banks, which typically finance risky investments through riskless deposits. In an adverse selection world, where some agents have private information on these risky investments, riskless deposits (which are not sensitive to this private information) may be particularly suited for uninformed agents. Notice, however, that FIs are not a necessary ingredient of that story, as the authors point out; riskless bonds, directly issued by firms, could do the same job as riskless deposits.

## 2.4 Financial Intermediation as Delegated Monitoring

In a context of asymmetric information, monitoring could clearly be a way to improve efficiency. Following Hellwig (1991), this discussion uses the term *monitoring* in a broad sense to mean

- *screening* projects a priori in a context of adverse selection (Broecker 1990);
- *preventing* opportunistic behavior of a borrower during the realization of a project (moral hazard) (Holmström and Tirole 1997);
- *punishing* (Diamond 1984) or *auditing* (Townsend 1979; Gale and Hellwig 1985; Krasa and Villamil 1992) a borrower who fails to meet contractual obligations; this is the context of costly state verification.

Whereas these monitoring activities clearly improve the efficiency of lender-borrower contracts with asymmetric information, they can very well be performed by the individual lenders themselves or more accurately by specialized firms: rating agencies, security analysts, or auditors. The delegated monitoring theory of financial intermediation (originally advanced by Diamond 1984) suggests that banks have a comparative advantage in these monitoring activities. For this theory to work, several ingredients are needed:

- Scale economies in monitoring, which implies that a bank typically finances many projects.
- Small capacity of investors (as compared to the size of investment projects), which implies that each project needs the funds of several investors.
- Low costs of delegation: the cost of monitoring or controlling the FI itself has to be less than the surplus gained from exploiting scale economies in monitoring or controlling investment projects.

The basic justification of banks put forward by Diamond (1984) is that banks economize on monitoring costs. To see why, consider a framework in which $n$ identical risk-neutral firms seek to finance projects. Each firm requires an investment of one unit, and the returns of each firm are identically and independently distributed.

The cash flow $\tilde{y}$ that a firm obtains from its investment is a priori unobservable to lenders.

**Monitoring Borrowers**

By paying $K$ (a monitoring cost) the lender is able to observe the realized cash flow and enforce the contractual repayment $\tilde{y}$. We assume that lending is profitable:

$$E[\tilde{y}] > 1 + r + K, \tag{2.19}$$

where $r$ is the riskless interest rate.

Assume also that each investor owns only $1/m$, so that $m$ of them are needed for financing one project, and that the total number of investors is at least $mn$, so that all projects can be financed. Direct lending would then imply that each of the $m$ investors monitors the firm he has financed: the total monitoring cost would be $nmK$ (see figure 2.2).

In this context, if a bank emerges, each investor would have to pay the cost of monitoring the bank, and the bank would still have to pay the cost of monitoring the $n$ firms. The total cost would be $nK + nmK$, so the bank would only bring in an additional cost corresponding to an additional layer in the monitoring process, which would be grossly inefficient. The idea of Diamond is that banks' incentives can be provided by an alternative technology: auditing.

**Audit Technology**

The bank's incentives to repay depositors are provided by the threat of a bank closure (bankruptcy). The bank promises a fixed deposit rate $r_D$ to depositors and is audited only if returns on the bank's assets are insufficient to repay depositors. The expected auditing cost has the advantage of being a fixed cost, independent of

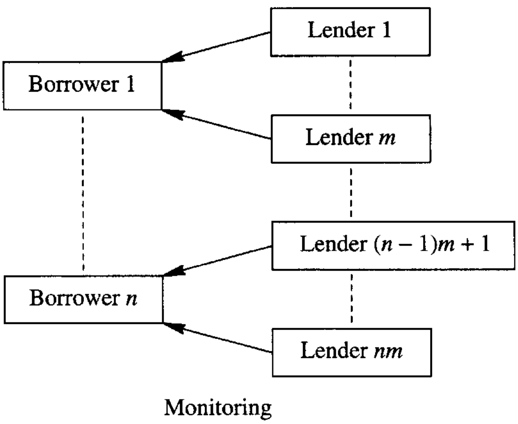



**Figure 2.2**
Direct finance: Each lender monitors its borrower (total cost $nmK$).

the number of investors. It is given by

$$C_n = n\gamma \Pr(\tilde{y}_1 + \tilde{y}_2 + \cdots + \tilde{y}_n < (1 + r_D)n), \tag{2.20}$$

where $\gamma$ is the unit cost of audit. We assume that the cost of audit is proportional to the volume of assets: this is why $\gamma$ is multiplied by $n$ in formula (2.20).[24] We also assume that $K < C_1$, which means that if the firm had a unique financier, it would be efficient to choose the monitoring technique.

If a bank (FI) emerges, it has to choose a monitoring technology for its loans and for its deposits. Regarding its loans, it can choose to monitor each borrowing firm (total cost $nK$) or to sign a debt contract with each of them (total cost $nC_1$). Since $K < C_1$, the first solution is preferable. The bank is therefore a delegated monitor, which monitors borrowers on behalf of lenders (see figure 2.3).

The equilibrium level of $r_D^n$ (the nominal interest rate on deposits) and the expected cost of audit clearly depend on $n$. They are determined implicitly by

$$E[\min(\tilde{Z}_n, 1 + r_D^n)] = 1 + r, \tag{2.21}$$

and

$$C_n = n\gamma \Pr(\tilde{Z}_n < 1 + r_D^n), \tag{2.22}$$

where $\tilde{Z}_n = (1/n)\sum_{i=1}^n \tilde{y}_i - K$ is the net return on the bank's assets. Now delegated monitoring will be more efficient than direct lending if and only if

$$nK + C_n < nmK. \tag{2.23}$$

**Result 2.4 (Diamond 1984)** If monitoring is efficient ($K < C_1$), investors are small ($m > 1$), and investment is profitable ($E(\tilde{y}) > 1 + r + K$), financial intermedia-

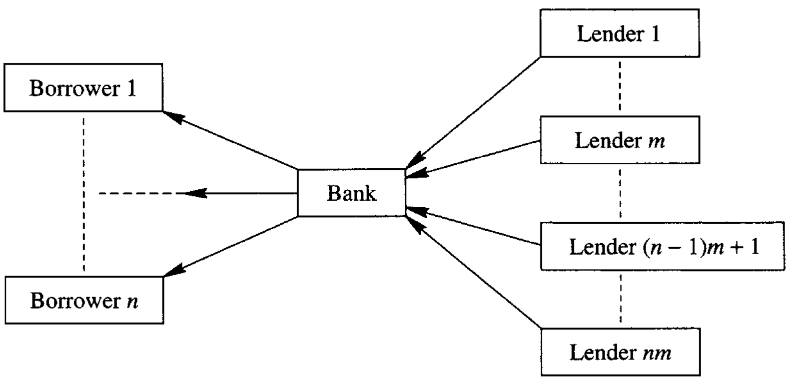



**Figure 2.3**
Intermediated finance: Delegated monitoring (total cost $nK + C_n$).

tion (delegated monitoring) dominates direct lending as soon as $n$ is large enough (diversification).

***Proof*** Condition (2.23) must be proved. Dividing it by $n$ yields an equivalent form:

$$K + \frac{C_n}{n} < mK.$$

Since $m > 1$, it is enough to prove that $C_n/n$ tends toward zero as $n$ tends toward infinity.

The strong law of large numbers dictates that $(1/n)\sum_{i=1}^{n} \tilde{y}_i$ converges almost surely to $E(\tilde{y})$. Since $E(\tilde{y}) > 1 + r + K$, relation (2.21) implies $\lim_n r_D^n = r$ (deposits are asymptotically riskless). Therefore, by (2.22),

$$\lim_n \frac{C_n}{n} = \lim_n \Pr(\tilde{Z}_n < 1 + r_D^n) = \Pr(E(\tilde{y}) - K < 1 + r) = 0. \qquad \blacksquare$$

Krasa and Villamil (1992) construct a model in which only ex post monitoring (auditing) is available, so that the problem of monitoring the monitors has to be solved. But if there are enough independent projects, the probability of the bank's insolvency goes to zero, and so does the cost of monitoring the bank. Duplication of monitoring costs is avoided because only the bank will have to monitor the firms.

Another interesting contribution that builds on Diamond's (1984) model of delegated monitoring is an article by Cerasi and Daltung (2000), who introduce considerations on the internal organization of banks as a possible explanation for the fact that scale economies in the banking sector seem to be rapidly exhausted, whereas Diamond's model predicts that banking should be a natural monopoly. The idea is that, in reality, monitoring is not performed by the banker, but by loan officers, who in turn have to be monitored by the banker. This additional delegation becomes more and more costly as the size of the bank increases because more and more officers have to be hired. Therefore, a trade-off exists between the benefits of diversification (which, as in Diamond 1984, improve the incentives of the banker) and the costs of internal delegation (which increase with the size of the bank).

A series of authors, notably Calomiris and Kahn (1991), Flannery (1994), Qi (1998), Rajan (1992), and Diamond and Rajan (2001), view the demandable character of bank liabilities as an instrument to prevent opportunistic behavior by bank managers. In particular, bank runs can be seen as a credible punishment for bank managers who do not monitor their borrowers. This suggests that the lending capacity of a bank can be increased if it is financed by demandable deposits. However, McAndrews and Roberds (1995), in their study of banking in medieval Bruges, suggest that the historical causation might well be inverse: "Facilitating payments (which determined the liability structure of the bank) was the initial and key role of banks, and ... banks' role in financial intermediation grew out of their original role of payment intermediation'' (29).

## 2.5 The Choice between Market Debt and Bank Debt

So far, the discussion has focused on why FIs exist and on the ''uniqueness of bank loans'' (to recall the title of an important article by James (1987), who shows that financial markets tend to react positively when they learn that a quoted firm has obtained a bank loan). However, direct access by firms to financial markets has experienced strong development in recent years (as part of the so-called disintermediation process), especially among large firms. Therefore, to be complete, this section analyzes the choice between direct and intermediate finance. Since in practice direct debt is less expensive than bank loans, it is usually considered that loan applicants are only those agents that cannot issue direct debt on financial markets. This discussion explains the coexistence of the two types of finance, based on moral hazard, which prevents firms without enough assets from obtaining direct finance.

### 2.5.1 A Simple Model of the Credit Market with Moral Hazard

Consider a model in which firms seek to finance investment projects of a size normalized to 1. The riskless rate of interest is normalized to zero. Firms have a choice between a good technology, which produces $G$ with probability $\pi_G$ (and zero otherwise), and a bad technology, which produces $B$ with probability $\pi_B$. Assume that only good projects have a positive net (expected) present value (NPV), $\pi_G G > 1 > \pi_B B$, but that $B > G$, which implies $\pi_G > \pi_B$. Assume also that the success of the investment is verifiable by outsiders, but not the firm's choice of technology nor the return.[25] Thus the firm can promise to repay some fixed amount $R$ (its nominal debt) only in case of success. The firm has no other source of cash, so the repayment is zero if the investment fails. The crucial element of this model is that the value $R$ of the nominal indebtedness of the firm determines its choice of technology. Indeed, in the absence of monitoring, the firm will choose the good technology if and only if this gives it a higher expected profit:

$$\pi_G(G - R) > \pi_B(B - R). \tag{2.24}$$

Since $\pi_G > \pi_B$, (2.24) is equivalent to

$$R < R_C = \frac{\pi_G G - \pi_B B}{\pi_G - \pi_B}, \tag{2.25}$$

where $R_C$ denotes the critical value of nominal debt above which the firm chooses the bad technology. From the lender's viewpoint, the probability $\pi$ of repayment therefore depends on $R$:

$$\pi(R) = \begin{cases} \pi_G & \text{if } R \le R_C, \\ \pi_B & \text{if } R > R_C. \end{cases}$$

In the absence of monitoring, a competitive equilibrium of the credit market is obtained for $R$ such that

$$\pi(R)R = 1. \tag{2.26}$$

Because $\pi_B R < 1$ for $R \le B$, equilibrium is only possible when the $G$ technology is implemented. This implies $R < R_C$, and thus $\pi_G R_C \ge 1$, which is only satisfied when moral hazard is not too important. If $\pi_G R_C < 1$, the equilibrium involves no trade, and the credit market collapses because good projects cannot be financed, and bad projects have a negative net present value (NPV).

Now a monitoring technology is introduced. At a cost $C$, specialized firms interpreted as banks can prevent borrowers from using the bad technology. Assuming perfect competition between banks, the nominal value of bank loans at equilibrium (denoted $R_m$, where $m$ stands for monitor) is determined by the break-even condition

$$\pi_G R_m = 1 + C. \tag{2.27}$$

For bank lending to appear at equilibrium, two conditions are needed:

- The nominal repayment $R_m$ on bank loans at equilibrium has to be less than the return $G$ of successful firms. Given condition (2.27), which determines $R_m$, this is equivalent to

$$\pi_G G - 1 > C. \tag{2.28}$$

In other words, the monitoring cost has to be less than the NPV of the good project. This is a very natural condition; without it, monitoring would be inefficient.

- Direct lending, which is less costly, has to be impossible:

$$\pi_G R_C < 1. \tag{2.29}$$

Therefore, bank lending appears at equilibrium for intermediate values of the probability

$$\pi_G \left( \pi_G \in \left[ \frac{1+C}{G}, \frac{1}{R_C} \right] \right),$$

provided this interval is not empty. Thus we have established the following:

**Result 2.5** Assume that the monitoring cost $C$ is small enough so that

$$\frac{1}{R_C} > \frac{1+C}{G}.$$

There are three possible regimes of the credit market at equilibrium:

1. If

$$\pi_G > \frac{1}{R_C}$$

(high probability of success), firms issue direct debt at a rate

$$R_1 = \frac{1}{\pi_G}.$$

2. If

$$\pi_G \in \left[\frac{1+C}{G}, \frac{1}{R_C}\right]$$

(intermediate probability of success), firms borrow from banks at a rate

$$R_2 = \frac{1+C}{\pi_G}.$$

3. If

$$\pi_G < \frac{1+C}{G}$$

(small probability of success), the credit market collapses (no trade equilibrium).

### 2.5.2 Monitoring and Reputation

This section is adapted from Diamond (1991). Its objective is to show, in a dynamic extension of the previous model (with two dates, $t = 0, 1$), that successful firms can build a reputation that allows them to issue direct debt instead of using bank loans, which are more expensive. In order to capture this notion of reputation, assume that firms are heterogeneous; only some fraction $f$ of them has the choice between the two technologies. The rest have access only to the bad one, and bank monitoring has no effect on them.

Under some conditions of the parameters, the equilibrium of the credit market will be such that

- at $t = 0$, all firms borrow from banks;
- at $t = 1$, the firms that have been successful at $t = 0$ issue direct debt while the rest still borrow from banks;
- banks monitor all the firms who borrow from them.

This example starts with the case of successful firms. Because of result 2.5, they will be able to issue direct debt if and only if

$$\pi_S > \frac{1}{R_C}, \tag{2.30}$$

where $\pi_S$ is the probability of repayment at date 2, conditionally on success at date 0 (and given that all firms have been monitored at $t = 0$). Bayes' formula gives the following:

$$\pi_S = \frac{P(\text{success at } t = 0 \text{ and } t = 1)}{P(\text{success at } t = 0)} = \frac{f\pi_G^2 + (1 - f)\pi_B^2}{f\pi_G + (1 - f)\pi_B}. \tag{2.31}$$

If (2.30) is satisfied, successful firms will be able to issue direct debt at a rate $R_S = 1/\pi_S$. On the other hand, the probability of success at $t = 1$ of the firms that have been unsuccessful at $t = 0$ is

$$\pi_U = \frac{f\pi_G(1 - \pi_G) + (1 - f)\pi_B(1 - \pi_B)}{f(1 - \pi_G) + (1 - f)(1 - \pi_B)}. \tag{2.32}$$

Result 2.5 implies that if

$$\frac{1 + C}{G} < \pi_U < \frac{1}{R_C},$$

these unsuccessful firms will borrow from banks, at a rate

$$R_U = \frac{1 + C}{\pi_U}.$$

In order to complete the picture, it is necessary only to establish that, at $t = 0$, for the adequate values of the different parameters, all firms (which are then indistinguishable) choose bank lending.

The symbol $\pi_0$ denotes the unconditional probability of success at $t = 0$ (recall that strategic firms choose the good technology because they are monitored):

$$\pi_0 = f\pi_G + (1 - f)\pi_B.$$

The notion of reputation building comes from the fact that $\pi_U < \pi_0 < \pi_S$, that is, the probability of repayment of its debt by the firm is initially $\pi_0$ but increases if the firm is successful ($\pi_S$) and decreases in the other case ($\pi_U$). Because of that, the critical level of debt, $R_C^0$ (above which moral hazard appears) at $t = 0$ is higher than in the static case. Indeed, firms know that if they are successful at $t = 0$, they will obtain cheaper finance ($R_S$ instead of $R_U$) at date 1. If $\delta < 1$ denotes the discount factor, the critical level of debt above which strategic firms choose the bad project at $t=0$ (denoted by $R_C^0$) is now defined by

$$\pi_B[B - R_C^0 + \delta\pi_G(G - R_S)] + (1-\pi_B)\delta\pi_G(G - R_U)$$
$$= \pi_G[G - R_C^0 + \delta\pi_G(G - R_S)] + (1-\pi_G)\delta\pi_G(G - R_U).$$

The left-hand side of this equality is the discounted sum of expected profits of a firm that chooses the bad technology at $t=0$. Note that at $t=1$ it will choose the good technology, either because it can issue direct debt at rate $R_S$ after a success or because it borrows from a bank at rate $R_U$ after being unsuccessful. The right-hand side represents the discounted sum of expected profits for a firm that chooses the good technology at $t=0$ (and thus also at $t=1$). Solving for $R$, we obtain

$$R_C^0 = \frac{\pi_G G - \pi_B B}{\pi_G - \pi_B} + \delta\pi_G(G - R_S) - \delta\pi_G(G - R_U),$$

which gives, after simplification:

$$R_C^0 = R_C + \delta\pi_G(R_U - R_S).$$

The following is the complete result:

**Result 2.6** Under the following assumptions,

$$\pi_0 \le \frac{1}{R_C^0}, \qquad \pi_S > \frac{1}{R_C}, \quad \text{and} \quad \frac{1}{R_C} > \pi_U > \frac{1+C}{G},$$

the equilibrium of the two-period version of Diamond's model is characterized as follows:

1. At $t=0$, all firms borrow from banks at a rate

$$R_0 = \frac{1+C}{\pi_0}.$$

2. At $t=1$, successful firms issue direct debt at a rate

$$R_S = \frac{1}{\pi_S},$$

whereas the rest borrow from banks at a rate

$$R_u = \frac{1+C}{\pi_U},$$

which is higher than $R_0$.

***Proof*** It is necessary only to apply result 2.5 repeatedly, after adjusting the parameters for the different cases. Part 1 of result 2.6 comes from part 2 of result 2.5, since by assumption

$$\frac{1+C}{G} < \pi_U < \pi_0 \leq \frac{1}{R_C^0}.$$

Part 2 of result 2.6 comes from parts 1 and 2 of result 2.5, since by assumption

$$\pi_G^S > \frac{1}{R_C} \quad \text{and} \quad \frac{1+C}{G} < \pi_U < \pi_0 \leq \frac{1}{R_C^0} < \frac{1}{R_C}. \qquad \blacksquare$$

Although this model is very simple, it captures several important features of credit markets:

- Firms with a good reputation can issue direct debt.[26]
- Unsuccessful firms pay a higher rate than new firms ($R_U > R_0$).
- Moral hazard is partially alleviated by reputation effects ($R_C^0 > R_C$).

### 2.5.3 Monitoring and Capital

This section is adapted from Holmström and Tirole (1997), who consider a simple model that elegantly captures the notion of substitutability between capital and monitoring at both the level of the firms and the level of banks. They obtain delegated monitoring without the complete diversification assumption of Diamond (1984). The moral hazard issue at the level of the bank is solved by bank capital. In a sense, their model is poles apart from Diamond (1984). They assume perfect correlation between the projects financed by banks, whereas Diamond assumes project independence.

More specifically, Holmström and Tirole's model considers three types of agents: (1) firms (borrowers), represented by the index $f$; (2) monitors (banks), represented by the index $m$; and (3) uninformed investors (depositors), represented by the index $u$. Each industrial project (owned by firms) costs the same amount $I$ and returns $y$ (which is verifiable) in case of success (and nothing in case of failure). There are two types of projects: a good project with a high probability of success $p_H$, and a bad project with a low probability of success $p_L$ ($p_H - p_L$ is denoted by $\Delta p$). Bad projects give a private benefit to the borrower; this is the source of moral hazard.[27] Monitoring the firm (which entails a nonpecuniary cost $C$) implies a reduction of this benefit from $B$ (without monitoring) to $b$ (with monitoring). Investors are risk-neutral, are uninformed (they are not able to monitor firms), and have access to an alternative investment of gross expected return $(1 + r)$. It is assumed that only the good project has a positive expected net present value even if the private benefit of the firm is included:

$$p_H y > 1 + r > p_L y + B.$$

Firms differ only in their capital $A$ (assumed to be publicly observable). The distribution of capital among the (continuum) population of firms is represented by the cumulative function $G(\cdot)$. Finally, the capital of banks is exogenous. Since it is assumed that bank asset returns are perfectly correlated, the only relevant parameter is total banking capital $K_m$, which determines the total lending capacity of the banking industry. The following are different possibilities through which a firm can find outside finance.

**Direct Lending**

A firm can borrow directly from uninformed investors by promising a return $R_u$ (in case of success) in exchange for an initial investment $I_u$. Since firms must be given incentives to choose the good project, there is an upper bound on $R_u$:

$$p_H(y - R_u) \geq p_L(y - R_u) + B \Leftrightarrow R_u \leq y - \frac{B}{\Delta p}. \tag{2.33}$$

The individual rationality constraint of uninformed investors implies an upper bound on $I_u$:

$$p_H R_u \geq (1 + r)I_u \Rightarrow I_u \leq \frac{p_H R_u}{1 + r} \leq \frac{p_H}{1 + r}\left(y - \frac{B}{\Delta p}\right). \tag{2.34}$$

Therefore, the project can be financed only if the firm has enough capital:

$$A + I_u \geq I \Rightarrow A \geq \bar{A}(r) \stackrel{\text{def}}{=} I - \frac{p_H}{1 + r}\left(y - \frac{B}{\Delta p}\right). \tag{2.35}$$

**Intermediated Lending**

If the firm does not have enough capital for issuing direct debt, it can try to borrow $I_m$ from a bank (in exchange for a return $R_m$ in case of success), together with a direct borrowing of $I_u$ from uninformed investors (in exchange for a return $R_u$ in case of success). The incentive compatibility constraint of the firm becomes

$$p_H(y - R_u - R_m) \geq p_L(y - R_u - R_m) + b \Leftrightarrow R_u + R_m \leq y - \frac{b}{\Delta p}. \tag{2.36}$$

The bank also must be given incentives to monitor the firm:

$$p_H R_m - C \geq p_L R_m \Leftrightarrow R_m \geq \frac{C}{\Delta p}. \tag{2.37}$$

Because banking finance is always more expensive than direct finance, the firm will borrow the least possible amount from the bank,

$$I_m = I_m(\beta) \stackrel{\text{def}}{=} \frac{p_H R_m}{\beta} = \frac{p_H C}{\beta \Delta p},$$

where $\beta$ denotes the expected rate of return that is demanded by the bank. The firm will obtain the rest,

$$I_u = \frac{p_H R_u}{1+r},$$

from uninformed investors. Therefore, constraint (2.37) is binding. Now, as a consequence of (2.36) and (2.37),

$$R_u \le y - \frac{b+C}{\Delta p},$$

which implies

$$I_u \le \frac{p_H}{1+r}\left(y - \frac{b+C}{\Delta p}\right).$$

Therefore, the project will be financed if and only if

$$A + I_u + I_m \ge I \Rightarrow A \ge \underline{A}(\beta, r) \stackrel{\text{def}}{=} 1 - I_m(\beta) - \frac{p_H}{1+r}\left(y - \frac{b+C}{\Delta p}\right). \tag{2.38}$$

Finally, the rate of return $\beta$ is determined by the equality between supply and demand of bank capital:

$$K_m = [G(\bar{A}(r)) - G(\underline{A}(\beta, r))]I_m(\beta), \tag{2.39}$$

where $K_m$ denotes the total capital of the banking industry (taken to be exogenous), $G(\bar{A}(r)) - G(\underline{A}(\beta, r))$ represents the number ("proportion") of firms that obtain loans, and $I_m(\beta)$ represents the size of each loan (the quantity lent by the bank). The right side of (2.39) being a decreasing function of $\beta$, there is a unique equilibrium.

**Result 2.7** At equilibrium, only well-capitalized firms ($A \ge \bar{A}$) can issue direct debt. Firms with intermediate capitalization ($\underline{A}(\beta, r) \le A < \bar{A}$) borrow from banks, and undercapitalized firms ($A \le \underline{A}(\beta, r)$) cannot invest.

Figure 2.4 categorizes firms by the type of finance available to them: those that cannot find external finance, those that obtain bank loans, and those that are funded directly in the financial markets.

The equilibrium values of $r$ (the riskless rate) and $\beta$ (the gross return on bank loans) are thus determined by two conditions: the equilibrium equation of the market for banking capital (2.39) and the equilibrium condition on the financial market. The savings supply $S(r)$ equals the demand for funds $D(\beta, r, C)$, defined by

$$D(\beta, r, C) = \int_{A(\beta, r)}^{\bar{A}(r)} (I - I_m - A)\, dG(A) + \int_{\bar{A}(r)}^{\bar{A}} (I - A)\, dG(A). \tag{2.40}$$

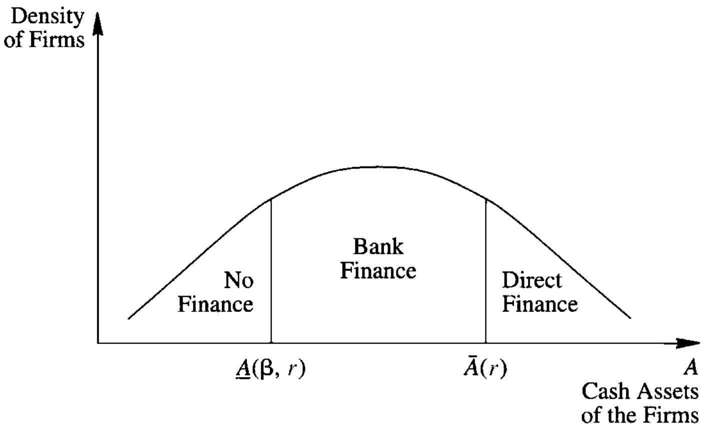



**Figure 2.4**
Firms categorized by type of finance.

Holmström and Tirole consider also a more general model, with a variable investment level. They study the effects of three types of financial shocks: a *credit crunch*, which corresponds to a decrease in $K_m$, the capital of the banking industry; a *collateral squeeze*, which corresponds to a negative shock on firms' assets; and a *savings squeeze*, which corresponds to a downward shift in the function $S$. They show the following properties:

**Result 2.8** Let $r$ and $\beta$ denote the equilibrium returns on financial markets and on bank loans, respectively. Then,

- a credit crunch decreases $r$ and increases $\beta$;
- a collateral squeeze decreases $r$ and $\beta$;
- a savings squeeze increases $r$ and decreases $\beta$.

### 2.5.4 Financial Architecture

The structure of Boot and Thakor's (1997) model is particularly rich and complex. They provide a formal analysis of a financial system where financial markets and banks coexist. The financial market equilibrium is reminiscent of Kyle (1985), as informed traders face an exogenous liquidity demand. The way the banking industry is modeled is close to Holmström and Tirole (1997), except that banks emerge as coalitions of investors with monitoring ability, as in Ramakrishnan and Thakor (1984).

The model considers two types of agents: investors and firms. Investors have three different options: they can become a financial analyst (or informed trader), become a banker, or deposit their money in a bank. If they are all used at equilibrium, these three options must give the same net expected return. Since the informed trader and the banker options are costly, they must generate an expected excess return that covers exactly these costs. This condition determines the proportions of informed traders, bankers, and depositors in the economy. Firms are heterogeneous. They are characterized by an observable parameter $\theta$ (interpreted as their reputation). With probability $\theta$ they invest in projects with a positive NPV, but with the complement probability $(1 - \theta)$ they are subject to moral hazard. They can then divert the funds to a negative NPV project that gives them private benefits.

Moral hazard can be avoided if a firm is monitored by a bank. The main result of Boot and Thakor's model is that good-reputation firms (such that $\theta > \hat{\theta}$, where $\hat{\theta}$ is endogenous) issue direct debt, and other firms borrow from banks.

In addition to justifying the coexistence of banks and financial markets, Boot and Thakor consider the effect of information transmission via asset prices. They do so by assuming that firms have the option to make an additional investment, which is only profitable in a good environment, and that informed traders can assess this environment. Their information being partly reflected into asset prices, financial markets create an additional surplus, which benefits both firms and informed traders.

This model has an interesting implication regarding the relation between the level of financial market development and the choice of financing source (financial architecture). At an early stage of financial development, moral hazard dominates. This increases the value created by banks and decreases the value created by financial markets, which is only generated if good projects are selected. At the same time, the initial lack of sophistication of financial market traders may imply that the cost they have to incur in order to assess the firm's environment is large. Therefore, in the early stages of financial developments bank finance will dominate. Conversely, as the financial system evolves, credit ratings will decrease the importance of moral hazard (as in Diamond 1991), and the role of financial markets in the economy will expand.

### 2.5.5 Credit Risk and Dilution Costs

Bolton and Freixas (2000) explore the coexistence of financial markets and financial intermediaries in a world where borrowers differ in their credit risks. Beyond providing a natural framework for analyzing financial intermediation, their goal is also to understand why equity issuing and bond financing are found predominantly in mature and relatively safe firms, whereas bank finance (or other forms of intermediated finance) is the only source of funding for start-up firms and risky ventures (see Petersen and Rajan (1994) and (1995)).

Financial market imperfection is associated with asymmetric information between firms and investors, which leads to informational dilution costs, as in Myers and Majluf (1984).

Banks' role is to produce interim monitoring. They are able to renegotiate the loans, choosing whether to liquidate the firm or to keep it in business in an efficient way. In contrast, renegotiation of a publicly issued bond is impossible, so default leads to liquidation. This assumption is in line with empirical evidence (Gilson and Lang 1990) and also consistent with the prevalence of covenants in loan contracts that trigger the renegotiation process.

Still, banks themselves must bear an intermediation cost, $\gamma$, stemming either from a monitoring cost or a dilution cost (as banks themselves have to issue equity), which is here assumed to be exogenous.

**The Model**

For simplicity, all investors are risk-neutral, and riskless interest rates are normalized to zero. A continuum of firms have to choose whether to be financed by a bank loan, by issuing equity, or by issuing a bond.

The firms' investment projects are characterized by an initial outlay of 1 at date $t = 0$, yielding a return $y > 1$ in case of success or 0 in case of failure, both at times $t = 1$ and $t = 2$. These projects can be liquidated at $t = 1$ for a resale value $A > 0$.

Firms are heterogeneous and differ by the (observable) probability $p$ of success at time $t = 1$. This probability $p$ is interpreted as the credit rating of the firm and is supposed to be distributed on the interval $[\underline{p}, 1]$, with $\underline{p} < \frac{1}{2}$.

There is adverse selection regarding time $t = 2$ cash flows. There are two types of firms: good ones (in proportion $\nu$), which are successful (obtain a return $y$) at time $t = 2$ with probability 1, and bad ones (in proportion $1 - \nu$), which have zero probability of success. Each firm knows its type, but creditors' beliefs (at $t = 0$) are uniform across firms.

Consequently, because of adverse selection, the cost of borrowing \$1 at $t = 1$ is a promised repayment of $\$(1/\nu)$. Yet, good borrowers know they will repay at time $t = 2$ with certainty, so that under full information their cost of borrowing \$1 should be \$1. The dilution cost is thus $1/\nu - 1$ per dollar borrowed.

**Firms' Financial Choice**

Firms choose among three different financial instruments (we assume that firms cannot combine them). Since the bad firms will mimic the good ones, we only have to care about the choices of the good firms:

- *Bond financing* implies a repayment $R$ at date $t = 1$ (in case of success) and nothing at $t = 2$. In case of default at time $t = 1$, the firm is declared bankrupt and is liquidated.

- *Equity issue*: a share $a \in [0, 1]$ of the cash flows generated by the firm is sold to the investors.
- *Bank debt* implies a repayment $\hat{R}$ at $t = 1$, and nothing at $t = 2$. If the firm defaults at $t = 1$, there is renegotiation and the bank is able to extract the entire surplus at $t = 2$ because it can observe the probability of success at date 2.

Each of the three financial instruments has its pros and cons. Equity financing eliminates inefficient liquidations but generates high dilution costs for good firms. On the contrary, bond financing has lower dilution costs but entails inefficient bankruptcy costs for good firms. Finally, bank loans have the benefits of both because there is efficient renegotiation in case of default and limited dilution costs, but there is an intermediation cost.

For each instrument, it is easy to compute the profits of good firms, considering that investors require a non-negative return. Bad firms systematically mimic the choices of good firms to avoid being identified.

**Bond Financing**

The zero-profit condition for investors is

$$1 = pR + (1 - p)A.$$

This nominal return $R$ is feasible $(R < y)$ if $py + (1 - p)A > 1$, and the expected profit of good firms is then

$$\Pi_B = p(y - R) + py.$$

Replacing $R$ by its value (given by the investors' zero profit condition), we get

$$\Pi_B = 2py - 1 + (1 - p)A.$$

**Equity Issue**

A fraction $a$ of the firm's capital is sold to outside investors. Because of adverse selection about the probability of success at $t = 2$, there is a dilution cost. Outside shareholders only anticipate an expected cash flow $vy$ at $t = 2$.

The zero-profit condition for outside shareholders is

$$1 = a[py + vy],$$

and the expected profit of good firms is

$$\Pi_E = (1 - a)[py + y].$$

Replacing $a$ by its value (given by the banks' zero-profit condition), we get

$$\Pi_E = \left(y - \frac{1}{p + v}\right)[p + 1].$$

**Bank Debt**

The zero-profit condition for banks is[28]

$$1 + \gamma = p\hat{R} + (1 - p)[A + v(y - A)],$$

and the expected profit of goods firms is

$$\Pi_{BL} = p(y - \hat{R}) + py.$$

Replacing $p\hat{R}$ by its value given by the investors' zero-profit condition, we get

$$\Pi_{BL} = 2py - 1 - \gamma + (1 - p)(A + v(y - A)).$$

By comparing these expected profits for different values of $p$ and $v$, we derive the optimal financing choice of the firms (fig. 2.5). Equity financing dominates when dilution costs are low, whereas bond financing dominates when credit risk is low or dilution costs are high.

## 2.6 Liquidity Provision to Firms

The Bryant-Diamond-Dybvig paradigm, presented in section 2.2, insists on the role of banks in providing liquidity insurance to depositors. This section summarizes briefly a more recent contribution by Holmström and Tirole (1998), which models

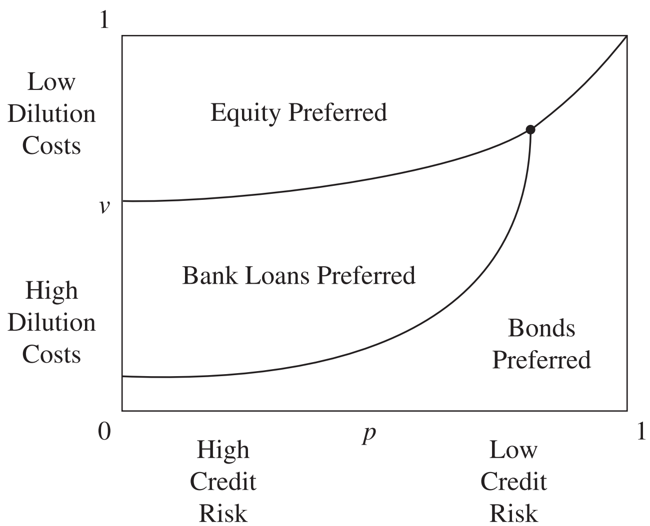



**Figure 2.5**
Optimal financing choices of firms.

the demand for liquidity by firms. This model shares a fundamental element with Diamond and Dybvig (1983): banks provide insurance against liquidity shocks. However, the demand for this insurance does not come from the risk aversion of depositors, as in Diamond and Dybvig (1983). In Holmström and Tirole (1998) depositors are risk-neutral. The demand for liquidity insurance comes instead from firms, which are subject to moral hazard problems.

As in Holmström and Tirole (1997) these moral hazard problems imply that firms have to self-finance some fraction of their investments, even if these investments have a positive expected NPV. In Holmström and Tirole (1998), firms are potentially subject to liquidity shocks. If such a shock occurs, the firm has to inject some additional funds into its project after the initial investment has been sunk. Moral hazard problems imply that firms have to self-finance some fraction of these liquidity needs. This can be obtained either by maintaining cash reserves or by securing a credit line with a bank.

In fact, the first solution is dominated: it would be inefficient for the firm to replace the credit line by the holding of liquid assets. If it did so, it would be by diverting funds from profitable investments. This would be costly because in equilibrium the self-financing constraint binds, and the firm's return on investment is larger than the bank's. The opportunity cost of holding liquid assets is therefore higher than the cost of the credit line.

The superiority of bank finance over direct finance in this context lies in the commitment possibilities that banks have and that financial markets typically don't have, an exception being derivative markets like futures and options. But these markets require the institution of a clearinghouse, which can be considered a financial intermediary.

So, to summarize, banks allow firms to insure against liquidity shocks by committing to finance them in the future (through credit lines), even if such funding is not profitable ex post. By so doing, they allow more firms to continue their projects, thus improving the overall efficiency of the economy.

## 2.7 Suggestions for Further Reading

The coexistence of public debt and bank loans is also justified by Diamond (1997), who assumes *limited participation* of investors in the financial markets, for example, because some of them do not trade every day. This possibility implies that assets offered for sale in the market will not attract bids from all possible buyers and that a low resale price will be anticipated. As a consequence, the investment in long-term assets will tend to be depressed. In such a situation, banks can emerge endogenously to solve the liquidity problem generated by limited participation. Indeed, when banks have a large volume of deposits, the liquidity needs of depositors are predictable, and limited participation is not an issue. An interesting outcome of this model is that it yields some predictions on financial development: as market participation increases, the market becomes more efficient, and the banking sector shrinks.

Another interesting paper explaining the coexistence of public debt and bank loans is Besanko and Kanatas (1993), which develops an attractive model sharing some of the main features of Holmström and Tirole's. Indeed, the former model also includes moral hazard on behalf of the firms, partly solved by monitoring services performed by the banks. But a bank must be provided with the correct incentives to monitor its borrowers, and this occurs only when the bank has a sufficient stake in the firm. (As in Holmström and Tirole, the monitoring activity is nonobservable, which creates a second moral hazard problem.) Also, once the bank lends to a firm and has incentives to monitor it, the firm can borrow from the security markets, which can piggyback on the bank's monitoring services.

The probability of success of the investments depends on the effort of entrepreneurs, which is not directly observable. This is the source of the moral hazard problem. However, banks can influence entrepreneurs' efforts through monitoring activities, the cost of which increases with the effort level that is required from the entrepreneurs. The equilibrium that is obtained is characterized by the fact that each firm combines direct lending and intermediated lending. Also, there is always a positive amount of monitoring because it is not possible to reach the best effort level. Finally, substituting bank financing for direct financing increases firms' stock price, a fact that is in accordance with the main empirical findings. This model differs from that of Holmström and Tirole (1997) in two ways. Collateral plays no role in Besanko and Kanatas (1993), and banks are not restricted in the amount of capital they are able to raise. As a consequence, a credit crunch cannot occur.

Also in the same vein as Holmström and Tirole (1997), Repullo and Suarez (1995) develop a model of financial intermediation with a more general specification of the moral hazard problem and use it to explore the choice of the structure of short-term credit (commercial paper versus bank loans) over the business cycle.

Bhattacharya and Chiesa (1995) study the problem of proprietary information disclosure by financiers. The word *proprietary* refers to the fact that the borrowing firms can be hurt if their competitors on the product market obtain this information. On the other hand, the lender is likely to gather this information in the monitoring process. This question may be particularly important in the context of R&D financing. Bhattacharya and Chiesa argue that in such a context bilateral bank-borrower relationships may be superior to multilateral lending. Similar arguments are modeled along the same lines by Yosha (1995a; 1995b).

Finally, Kashyap, Rajan, and Stein (1999) explore a new explanation for possible scope economies between deposit taking and lending. They argue that loan commitments, the contracts by which banks allow firms to borrow liquidity when they need it (up to some maximal amount), are very similar to demand deposit contracts (the only difference being that deposits are owned by the firm). In particular, the liquid reserves held by banks (to protect themselves against the risk that depositors withdraw their money) can also be used as a buffer against the risk that firms draw on their credit lines. Since the opportunity cost of holding liquid reserves is a convex function of the amount of reserves, a bank that offers both deposit contracts and loan commitments is more efficient than two separate banks that specialize (provided that liquidity risks on deposits and loan commitments are independent, so that diversification works). Kashyap, Rajan, and Stein (1999) show that empirical evidence supports their thesis. Banks that collect a large amount of deposits are more likely to offer more loan commitments.

## 2.8 Problems

### 2.8.1 Strategic Entrepreneurs and Market Financing

Consider an economy with one good, a continuum of risk-neutral investor-depositors with an aggregate measure of savings of $S$, and a continuum of risk-neutral entrepreneurs with no wealth and an aggregate measure $M$ ($M$ is the maximum demand for investment) of two nonobservable types:

• Bad entrepreneurs in proportion $1 - \mu$ invest 1 in a bad project, which is successful with probability $p_B$, in which case it yields a return of $B$. If the project fails, the return is zero.

• Strategic entrepreneurs in proportion $\mu$ have a choice between implementing a bad project and choosing a good one, which is successful with probability $p_G$, in which case it yields a return of $G$. Both projects require an investment of 1. The strategic entrepreneur choices are as follows:

$$
\begin{array}{cccc}
 & t=0 & t=1 & \text{Probability} \\
G & -1 & \begin{bmatrix} G \\ 0 \end{bmatrix} & \begin{bmatrix} p_G \\ 1-p_G \end{bmatrix} \\
B & -1 & \begin{bmatrix} B \\ 0 \end{bmatrix} & \begin{bmatrix} p_B \\ 1-p_B \end{bmatrix}
\end{array}
$$

We assume that $p_G G > 1 > p_B B$, and $B > G$, that interest rates are normalized to zero, and that firms issue bonds with a return $R$ in case of success and zero otherwise.

1. Compute the range of values for $\mu$ and for $R$ for which the strategic entrepreneur will have an incentive to implement a good project.

2. Under what conditions will a market equilibrium exist when $S > M$ and when $S < M$?

3. Should we characterize the equilibrium by the equality of supply and demand?

### 2.8.2 Market versus Bank Finance

Consider an economy with a continuum of firms with zero wealth that implement random projects with an initial investment of 1 and a return of $X$ with probability $\theta$ and zero otherwise. There are two types of firms, good firms in a proportion $v_H$, which have a probability of success $\theta_H$, and bad firms in a proportion $v_L$, which have a probability of success $\theta_L$ ($\theta_H > \theta_L$ and $\theta_L X < 1 < \bar{\theta} X < \theta_H X$, where $\bar{\theta} \equiv v_H \theta_H + v_L \theta_L$). Firms do not have initial wealth.

Investors are assumed to be risk-neutral and cannot distinguish the two types of firms. Interest rates are normalized to zero.

1. Assuming investors obtain a zero expected return, compute the nominal interest rate on a bond issued by a firm.

2. Assume a competitive banking industry exists. By paying a sunk cost $C$ per firm at the initial period, banks are able to identify a firm's type and then decide whether to lend or not. Under what conditions does the bank have an incentive to monitor the firm? Under what conditions will all good firms be attracted by competitive bank loan conditions? How many competitive equlilibria will exist?

3. If several equilibria exist, how do they compare in terms of efficiency, that is, in terms of output maximization, if $C$ is a real cost?

*Hint:* Banks will compete with the bond market for the good firms, so in equilibrium some proportion $\gamma$ ($0 \leq \gamma \leq 1$) of good firms will issue bonds.

### 2.8.3 Economies of Scale in Information Production

This problem is inspired by Ramakrishnan and Thakor (1984). Consider an agent (e.g., a security analyst) who is able to produce some information that is valuable to a risk-neutral principal (an investor). The principal observes a signal $\beta$ positively correlated with the effort $e$ spent by the agent in producing the information. For simplicity, assume that both $\beta$ and $e$ are binomial:

$$\text{Proba}(\beta = 1|e = 1) = p > \text{Proba}(\beta = 1|e = 0) = q,$$

$$\text{Proba}(\beta = 0|e = 1) = 1 - p < \text{Proba}(\beta = 0|e = 0) = 1 - q.$$

The contract between the principal and the agent specifies the agent's wage $Z$ as a function of $\beta$. The utility of the agent is

$$V(Z, e) = u(Z) - Ce,$$

where $u$ is concave and increasing, and $C$ denotes the cost of effort.

1. Compute the expected cost for the principal of inducing information production by the agent. It is defined as the minimum expected value wage schedule $\beta \to Z(\beta)$ such that

- the agent makes an effort (incentive compatibility constraint);
- the agent accepts the contract (individual rationality constraint).

The reservation utility of the agent (the utility level the agent can obtain outside) is denoted by $R$.

2. Consider now the case of two agents (with no communication problems between them) who are able to sign separate contracts with the principal and equally share their total receipts. Show that they are better off in the coalition.

### 2.8.4 Monitoring as a Public Good and Gresham's Law

The following model formalizes the idea that an economy using several risky means of payment (monies) issued by competing banks is confronted with free-rider and lemon problems, associated with Gresham's law: "Bad money drives away good money." Consider a model with $N$ identical banks ($n = 1, \ldots, N$), each having $M$ identical depositors with a unit deposit. The depositors of bank $n$ are indexed by the couple $(m, n)$ where $m = 1, \ldots, M$. Each bank issues bank notes that can be used as a store of value or circulated as a means of payment. The quality $q_n$ of the notes issued by bank $n$ (related to its probability of failure) increases according to the monitoring efforts spent by each of the bank's depositors. Assume the following simple specification:

$$q_n = \sum_{m=1}^{M} e(m, n) + \theta,$$

where $e(m, n)$ represents the effort spent by depositor $(m, n)$ in monitoring his bank's management, and $\theta$ represents the intrinsic quality of the bank. It is assumed that $q_n$ is known only to the depositors of bank $n$. The utility of depositor $(m, n)$ is thus

$$U(m,n) = \begin{cases} q_n - \dfrac{1}{2}\gamma e^2(m,n) & \text{when he stores his bank notes,} \\ P - \dfrac{1}{2}\gamma e^2(m,n) & \text{when he circulates them,} \end{cases}$$

where $C(e) = \frac{1}{2}\gamma e^2$ represents the cost of effort, and $P$ is the market price for money in circulation. As in Akerlof's (1970) market for lemons, this price is identical for all circulating monies: $P = k\bar{q}$, where $\bar{q}$ is the average quality of circulating monies, and $k > 1$ represents the utility gained from using money as a means of payment. From the expression $U(m, n)$, money $n$ circulates if and only if $P \geq q_n$.

1. Show that in any symmetric situation $(e(m, n) \equiv e, q_n \equiv q)$ all monies circulate, and the utility of each depositor is

$$U = kq - \frac{1}{2}\gamma e^2,$$

where $q = Me + \theta$.

2. Show that in the best situation all monies circulate and

$$e(m, n) \equiv e^* = \frac{kM}{\gamma}; \qquad q_n \equiv q^* = \frac{kM^2}{\gamma} + \theta.$$

3. Show that in a symmetric Nash equilibrium all monies circulate but

$$e(m, n) \equiv e^{**} = \frac{k}{\gamma}; \qquad q_n \equiv q^{**} = \frac{kM}{\gamma} + \theta.$$

Therefore, the quality of money is dramatically insufficient (free-rider effect).

4. Suppose that banks have different intrinsic qualities $\theta_1 < \theta_2 < \cdots < \theta_N$.

4a. Show that if circulating monies were distinguishable, the best effort levels would be the same as in question 2.

4b. Determine the characteristics of a Nash equilibrium, and show that the free-rider effect is aggravated by a lemon problem.

4c. Assume that $k$ is less than the number $N^*$ of monies in circulation. Is Gresham's law satisfied?

4d. In the particular case in which $N = 2$, find conditions under which only money 1 circulates at equilibrium.

### 2.8.5 Intermediation and Search Costs

This problem is adapted from Gehrig (1993). Consider an economy with a continuum of potential buyers and sellers, characterized by their valuations $b$ and $s$ for a given good. The valuations $b$ and $s$ are uniformly distributed on the interval $[0, 1]$ and are publicly observed.

1. If there is a central marketplace, show that the (Walrasian) equilibrium involves the upper half of the buyers $\left(b \geq \frac{1}{2}\right)$ trading with the lower half of the sellers $\left(s \leq \frac{1}{2}\right)$ at a price $p^* = \frac{1}{2}$. Compute the total surplus.

2. Assume that traders meet only individually. When buyer $b$ meets seller $s$, they trade at price

$$\frac{b+s}{2}$$

(provided $b \geq s$). For simplicity, rule out other bargaining solutions and more complex search strategies. Compute the expected total surplus.

3. Introduce an intermediary who sells at an ask price $\hat{b}$ and buys at a bid price $\hat{s}$. Show that the upper part of the distribution of buyers $(b \geq b^*)$ and the lower part of the distribution of sellers $(s \leq s^*)$ trade with the intermediary, whereas the rest still search for a direct trade. Compute $b^*$ and $s^*$ as functions of $\hat{b}$ and $\hat{s}$.

4. Compute the bid and ask prices that maximize the profit of the intermediary (monopoly situation). Show that some traders are better off than in a competitive situation (question 1).

### 2.8.6 Intertemporal Insurance

Allen and Gale (1997) suggest that banks may be better than markets for providing intertemporal insurance. Their point is illustrated by the effects of the oil shock in the early 1970s. While the real value of shares listed on the New York Stock Exchange fell by almost half, in Germany the financial system was able to absorb the shock rather than pass it on to investors. This was at the cost of doing less well in the 1980s.

Consider the following overlapping generations model. Agents are born with a unit of endowment and live for two periods. They have to choose how to allocate their savings between a safe asset (storage technology) and an infinite-lived risky asset that produces i.i.d. cash flows $(y_t)_t$, equal to 0 with probability 0.5, and equal to 1 with the same probability.

During the first period of her life an agent born at date $t$ is faced with the budget constraint

$$C_{1t} + S_t + p_t x_t = 1, \tag{2.41}$$

where $C_{1t}$ denotes the consumption at time $t$ of the agent born at date $t$; $S_t$ is the investment in the storage technology; $p_t$ is the price of the risky asset; and $x_t$ is the quantity of the risky asset she buys.

At time 2 the budget constraint becomes

$$C_{2t+1} = S_t + p_{t+1}x_t + y_{t+1}x_t, \tag{2.42}$$

where $C_{2t+1}$ is the consumption at time $t+1$ of the agent born at date $t$. For simplicity, we assume that there is a continuum of mass 1 of agents in each generation.

The initial supply of the riskless asset is $S_0 = 0$, and the supply of the risky asset is fixed and normalized to 1. The utility function of each agent is

$$u(C_{1t}, C_{2t+1}) = \ln C_{1t} + \ln C_{2t+1}.$$

We look for a stationary Markov market equilibrium, that is, an asset price $p(y_t)$ and consumption and savings decisions that maximize the utility of consumers, clear the markets, and only depend on the current state of the economy, that is, given by the cash flow $y_t$ produced by the risky technology at the current date $t$.

1. Show that the equilibrium price function is necessarily constant: $p(y_t) \equiv p$.
2. Show that all savings are invested in the risky asset.
3. Compute the demand for the risky asset $x(p)$ by the young generation.
4. Compute the equilibrium price and allocations. Suppose that a bank is created at date 0. It starts with reserves $R_0 = \frac{1}{2}$, financed by a tax on generation 0. Agents of all other generations deposit their endowments in the bank and are allowed to withdraw $\frac{3}{4}$ at each date. The bank owns the risky technology.
5. Show that the dynamics of the reserves $R_t$ of the bank are given by

$$R_{t+1} = \max\left(0, R_t + y_t - \frac{1}{2}\right).$$

Whenever $R_{t+1} = 0$, the bank is closed and the economy returns to the market solution.

6. Show that the expected utility of all generations (except generation 1) is higher with the bank than without it.

## 2.9 Solutions

### 2.9.1 Strategic Entrepreneurs and Market Financing

1. A market equilibrium will exist provided that investors and firms satisfy their participation constraint, and firms have an incentive to choose the good technology:

$$(\mu p_G + (1 - \mu) p_B) R \equiv \bar{p} R \geq 1 \tag{2.43}$$

and

$$p_G(G - R) \geq p_B(B - R)$$

for $R \leq G$, which is equivalent to

$$R \leq R^* \equiv \frac{p_G G - p_B B}{p_G - p_B}. \tag{2.44}$$

A necessary condition for (2.43) and (2.44) to be simultaneously satisfied is that

$$\frac{1}{\bar{p}} \leq \frac{p_G G - p_B B}{p_G - p_B}. \tag{2.45}$$

Conversely, if condition (2.45) is satisfied, there exists a range of values for $R$ such that both (2.43) and (2.44) are satisfied. Notice that condition (2.45) can be rewritten as

$$\mu \geq \mu^*,$$

with

$$\mu^* = \frac{\frac{1}{R^*} - p_B}{p_G - p_B}.$$

Thus the existence of market finance depends upon the existence of a sufficiently high number of strategic entrepreneurs.

2. Where $S > M$, there is an excess of savings, and competition will drive the rates down until the point where (2.43) is satisfied with equality, implying $R = 1/\bar{p}$. Where $S < M$ case, we have

$$R = \frac{p_G G - p_B B}{p_G - p_B}.$$

3. There is no equality of demand and supply in $S < M$ case because at the prevailing rate entrepreneurs are not indifferent between having a loan or not. They strictly prefer to have a loan. Thus $M - S$ of them are rationed.

### 2.9.2 Market versus Bank Finance

1. Let $\bar{\theta} \equiv v_H \theta_H + v_L \theta_L$. Then a bond has to repay an amount $R_B$ such that $\bar{\theta} R_B = 1$. This payment is feasible because $1/\bar{\theta} \leq X$.

2. Let $R_L$ be the repayment on a bank loan. Denote by $\gamma$ the proportion of good firms that issue bonds. Then define

$$\theta(\gamma) = \frac{\gamma v_H \theta_H + v_L \theta_L}{\gamma v_H + v_L}$$

as the average probability of repayment. As intuition suggests, $\theta(\gamma)$ is increasing in $\gamma$. From this value we obtain the bond market interest rate $R_B(\gamma)$ by solving the equation $\theta(\gamma) R_B = 1$. The interest rate

$$R_B(\gamma) = \frac{\gamma v_H + v_L}{\gamma v_H \theta_H + v_L \theta_L}$$

is decreasing in $\gamma$. Since $\theta_L X < 1$, we obtain that $\theta(\gamma)X = 1$ will be satisfied for some value $\hat{\gamma} > 0$.

Banks will lend only with a probability $\nu_H$, so their profits will be given by $\nu_H \theta_H R_L - C$. Individual rationality for banks that invest an amount $\nu_H$ implies $\nu_H \theta_H R_L \geq \nu_H + C$, so the zero-profit condition becomes

$$R_L = \frac{1 + \frac{C}{\nu_H}}{\theta_H}.$$

To distinguish the different types of equilibria, we consider those with $\gamma = 0$, those with $\gamma = 1$, and those with $\gamma \in (0, 1)$. Note that $\gamma$ is endogenous and determined in equilibrium.

- *Banks-only equilibrium* ($\gamma = 0$). The condition for such an equilibrium to exist is that $R_L \leq X$ (the loan is feasible) and $R_L \leq R_B(0)$ (the loan is attractive for borrowers). Since for $\gamma = 0$ the market for bonds does not exist (or equivalently, $R_B(0) > X$), only the first constraint is relevant.

Using the zero-profit condition for banks,

$$\frac{C}{\nu_H} \leq \theta_H X - 1. \tag{2.46}$$

The economic interpretation: $C/\nu_H$, the cost per unit of loan that is granted, should be small with respect to the project's net present value.

- *Bonds-only equilibrium* ($\gamma = 1$). The condition is here the opposite one, $R_B(1) \leq R_L \leq X$. Using the competitive values for $R_B(1)$ and $R_L$, the condition becomes

$$\frac{1}{\bar{\theta}} \leq \frac{1 + \frac{C}{\nu_H}}{\theta_H}, \tag{2.47}$$

or equivalently,

$$\frac{1}{\bar{\theta}} - \frac{1}{\theta_H} \leq \frac{C}{\nu_H}.$$

The economic interpretation: the monitoring cost per unit of loan $C/\nu_H$ is excessively large in comparison with the gains from screening that allow going from a default probability of $\bar{\theta}$ to a probability of $\theta_H$.

- *Interior solution*. For an interior solution to exist, we need $R_L = R_B(\gamma)$, implying

$$\frac{\gamma \nu_H + \nu_L}{\gamma \nu_H \theta_H + \nu_L \theta_L} = \frac{1 + \frac{C}{\nu_H}}{\theta_H},$$

which provides a unique solution in $\gamma$.

So, to summarize, since conditions (2.46) and (2.47) depend upon different parameters, it is possible that neither of them is satisfied (no funding), which will occur when monitoring is too expensive and the population of borrowers is of bad quality; that only one is satisfied; or that both are satisfied.

When this is the case, that is, for $C/\nu_H$ within the interval $(1/\bar{\theta} - 1/\theta_H, \theta_H X - 1)$, three equilibria coexist, depending on how borrowers coordinate and choose bank loans or bond finance. Clearly, the pure bank equilibrium ($\gamma = 0$) occurs, as does the equilibrium with bonds only ($\gamma = 1$). By continuity of $R_B(\gamma)$, the third equilibrium will also exist because we have

$$R_B(1) < \frac{1 + \frac{C}{\nu_H}}{\theta_H} < R_B(\hat{\gamma}) = X.$$

3. Compute the aggregate net output (zero interest rates allow us to subtract time 0 input from time 1 output).

With bonds only, a fraction of negative net present value is implemented, and the total output will be

$$\nu_H \theta_H (X - 1) - \nu_L \theta_L (1 - X).$$

With banks only, the total output will be

$$\nu_H \theta_H (X - 1) - C.$$

In the interior solution, the output will be

$$(\gamma \nu_H \theta_H (X - 1) - \nu_L \theta_L)(1 - X) + (1 - \gamma) \nu_H \theta_H (X - 1) - C.$$

This expression implies duplication of screening costs because firms identified as good will choose the bond market. This case is always dominated by the previous cases, for $\gamma = 0$ or $\gamma = 1$.

Consequently only one of the equilibria is efficient. The banks-only equilibrium is efficient if

$$C < \nu_L \theta_L (1 - X),$$

that is, if the cost of monitoring is lower than the cost of inefficient investment by the bad firms. The bonds-only equilibrium is efficient in the opposite case.

### 2.9.3 Economies of Scale in Information Production

1. Let $Z(\beta)$ denote the agent's wage schedule in the optimal contract, and set

$$W_0 = u(Z(0)), \qquad W_1 = u(Z(1)).$$

The agent will make an effort if and only if the incentive compatibility constraint is satisfied, namely,

$$pW_1 + (1-p)W_0 - C \geq qW_1 + (1-q)W_0,$$

or equivalently,

$$W_1 - W_0 \geq \frac{C}{p-q}.$$

The agent will accept the contract if and only if the individual rationality constraint is satisfied, namely,

$$pW_1 + (1-p)W_0 - C \geq R.$$

The optimal contract will be such that both constraints are binding, which gives

$$\begin{cases} W_0 = R - \dfrac{q}{p-q}C, \\ W_1 = R + \dfrac{1-q}{p-q}C. \end{cases}$$

2. Suppose that each of the two agents separately signs the contract with the principal and that they decide to equalize their wages (mutual insurance). Denote the individual signals received by the principal on the performance of each agent as $\beta_1$ and $\beta_2$. Then if $\beta_1 = \beta_2$, the agents gain nothing by pooling their wages. However, if $\beta_1 \neq \beta_2$, each of them gets

$$\frac{Z(0) + Z(1)}{2}.$$

Since they are risk-averse, they are better off in the coalition. In fact, their expected utility gain is exactly

$$\Delta U = \underbrace{2p(1-p)}_{\substack{\text{probability} \\ \text{that } \beta_1 \neq \beta_2}} \left[ \underbrace{u\left(\frac{Z(0)+Z(1)}{2}\right)}_{\substack{\text{average wage} \\ \text{in coalition}}} - \underbrace{\frac{1}{2}u(Z(0)) - \frac{1}{2}u(Z(1))}_{\substack{\text{risky wage} \\ \text{outside coalition}}} \right].$$

### 2.9.4 Monitoring as a Public Good and Gresham's Law

1. In a symmetric situation,

$$q_n \equiv q < P = kq \qquad (k > 1).$$

Thus all monies circulate, and the utility of any depositor is

$$U = kq - \frac{1}{2}\gamma e^2,$$

where $q = Me + \theta$.

2. The best situation is symmetric because utilities are concave and costs are convex. Therefore, all monies circulate and $e$ is chosen to maximize $U$:

$$U = k(Me + \theta) - \frac{1}{2}\gamma e^2.$$

Thus,

$$e = e^* = \frac{kM}{\gamma}; \qquad q = q^* = Me^* + \theta = \frac{kM^2}{\gamma} + \theta.$$

3. In a symmetric Nash equilibrium all monies circulate, but each depositor takes the efforts of others as given. Therefore $e$ is chosen to maximize

$$U(e) = k((M-1)e^{**} + \theta + e) - \frac{1}{2}\gamma e^2.$$

Thus,

$$e = e^{**} = \frac{k}{\gamma}; \qquad q = q^{**} = \frac{kM}{\gamma} + \theta.$$

Clearly, $e^{**} \ll e^*$, and $q^{**} \ll q^*$.

4a. Since $\theta_i$ does not affect the marginal impact of effort on quality, the best level of effort is the same as in question 2.

4b. In a Nash equilibrium the marginal utility of effort for a depositor varies according to whether his money circulates. In the first case it is equal to $k/N^*$, where $N^*$ denotes the number of monies in circulation, and in the second case it is equal to 1. Therefore, the Nash equilibrium level of effort equals

$$e_1^* = \frac{k}{N^*\gamma} \quad \text{in the first case,}$$

$$e_2^* = \frac{1}{\gamma} \quad \text{in the second case.}$$

4c. If $N^* > k$ (which is assumed), then $e_1^* < e_2^*$. Money $n$ circulates if and only if $P \geq q_n = \theta_n + Me_1^*$, and therefore Gresham's law is satisfied: good-quality monies are driven out of the market.

4d. When $N = 2$, the conditions for money 1 to be the only circulating money at equilibrium are $P = kq_1 < q_2$, with

$$q_1 = \theta_1 + \frac{Mk}{\gamma} \quad \text{and} \quad q_2 = \theta_2 + \frac{M}{\gamma}.$$

This is summarized by

$$k\theta_1 + \frac{Mk^2}{\gamma} < \theta_2 + \frac{M}{\gamma}.$$

### 2.9.5 Intermediation and Search Costs

1. If there is a central marketplace, all trades take place at the same price $p$. The demand and supply functions are

$$D(p) = \int_p^1 db = 1 - p \quad \text{and} \quad S(p) = \int_0^p ds = p.$$

Therefore the equilibrium price is $p^* = \frac{1}{2}$, and trade takes place between the upper half of buyers $(b \geq p^*)$ and the lower half of sellers $(s \leq p^*)$. The total surplus is

$$\int_{1/2}^1 b\, db - \int_0^{1/2} s\, ds = \frac{1}{4}.$$

2. If traders meet only individually, the ex post surplus equals $\max(b - s, 0)$. For any number $y$, we denote by $y_+$ the value $\max(y, 0)$. The expected surplus becomes

$$\int_0^1 \int_0^1 (b - s)_+ \, db\, ds = \frac{1}{6}.$$

3. A buyer of type $b$ has to compare her surplus $(b - \hat{b})$ if she buys from the intermediary to her expected surplus

$$E\left[\frac{(b - s)_+}{2}\right]$$

(computed on the relevant population of sellers) if she trades at random. The difference between these two expressions is clearly increasing in $b$ because the marginal surplus is 1 in the case of a transaction with the intermediary and less than $\frac{1}{2}$ in the case of direct trade. Therefore, for $b$ larger than some cutoff level $b^*$, the buyer will buy from the intermediary. By symmetry, a seller of type $s$ will sell to the intermediary if and only if $s \leq s^*$. The cutoff levels $b^*$ and $s^*$ are jointly determined by the following two equations:

$$\begin{cases} b^* - \hat{b} = E\left[\dfrac{(b^* - s)_+}{2}\middle| s \geq s^*\right], \\ \hat{s} - s^* = E\left[\dfrac{(b - s^*)_+}{2}\middle| b \leq b^*\right], \end{cases}$$

where the two expectations are conditioned by the fact that the other party to the transaction (the seller in the first equation, the buyer in the second) does not trade with the intermediary. Easy computations lead to a transformation of these conditions into

$$\begin{cases} b^* - \hat{b} = \dfrac{1}{4}\dfrac{(b^* - s^*)^2}{1 - s^*}, \\ \hat{s} - s^* = \dfrac{1}{4}\dfrac{(b^* - s^*)^2}{b^*}. \end{cases}$$

Feasibility for the intermediary (supply equals demand) implies

$$b^* = 1 - s^*,$$

which, because of the preceding equations, gives

$$\hat{b} = 1 - \hat{s}.$$

This could be expected from the symmetry of the problem. Now everything can be expressed in terms of $\hat{s}$:

$$\hat{s} - 1 + b^* = \frac{1}{4}\frac{(2b^* - 1)^2}{b^*},$$

which gives

$$b^* = \frac{1}{4\hat{s}}, \qquad s^* = 1 - \frac{1}{4\hat{s}}.$$

4. The profit of the intermediary is

$$\pi = (\hat{b} - \hat{s})s^*,$$

or using the preceding expressions,

$$\begin{aligned} \pi &= (1 - 2\hat{s})\left(1 - \frac{1}{4\hat{s}}\right), \\ &= \frac{3}{2} - 2\hat{s} - \frac{1}{4\hat{s}}. \end{aligned}$$

This is maximum for

$$\hat{s} = \frac{1}{2\sqrt{2}},$$

which gives

$$b^* = 1 - s^* = \frac{\sqrt{2}}{2}.$$

Comparing this result with the competitive situation examined in question 1 shows that the gains from trade are not completely exploited by the intermediary. This is not surprising, given the monopolistic situation. But this has an interesting consequence. Consider a buyer whose valuation $b$ lies just below the competitive price $p^* = \frac{1}{2}$. In a competitive situation he obtains a zero surplus because he does not buy in the marketplace, and all sellers with whom he could have traded (with a cost $s \leq b$) have gone to the marketplace. This is not true with a monopolistic intermediary who has set a bid price

$$\hat{s} = \frac{1}{2\sqrt{2}} < \frac{1}{2}.$$

Therefore if $b > \hat{s}$, the buyer obtains a positive expected surplus by searching for a seller with a cost parameter $s$ between $\hat{s}$ and $b$.

### 2.9.6 Intertemporal Insurance

1. The old generation inelastically supplies one unit of the risky asset independently of $y_t$. Since preferences are smooth, the price of the risky asset must be constant.

2. The net return on the risky asset is always positive because it can be resold at the same price it was bought, and it distributes non-negative cash flows. Thus it dominates the storage technology, and no one invests in the safe asset ($S_t \equiv 0$).

3. Since $S_t \equiv 0$, the budget constraints of the agent born at date $t$ give

$$C_{1t} = 1 - px_t \quad \text{and} \quad C_{2t} = (p + y_{t+1})x_t.$$

The agent's expected utility can be expressed as a function of $x_t$ only:

$$Eu = \ln(1 - px_t) + E[\ln((p + y_{t+1})x_t)],$$

which simplifies into

$$Eu = \ln(1 - px_t) + \ln x_t + E[\ln(p + y_{t+1})].$$

It is maximum when

$$\frac{-p}{1 - px_t} + \frac{1}{x_t} = 0,$$

or

$$x_t = x(p) = \frac{1}{2p}.$$

4. The supply of the risky asset is 1, therefore the equilibrium price is $p = \frac{1}{2}$. The allocations are $C_{1t} \equiv \frac{1}{2}$, $C_{2t} = \frac{1}{2} + y_{t+1}$. Therefore the old generation assumes all the risk: there is no intertemporal insurance.

5. If the bank promises $\frac{3}{4}$ to each of its depositors (young or old), it has to finance $\frac{3}{2}$ withdrawals at each date with one unit of deposits and the cash flow $y_t$ produced by the risky asset. Since reserves cannot be negative, we have

$$R_{t+1} = \max\left(0, R_t + y_t - \frac{1}{2}\right).$$

The bank is closed at date $t+1$ if and only if $R_t = \frac{1}{2}$ and $y_t = 0$. It can be shown that the bank survives forever with a probability $\frac{1}{3}$.

6. The expected utility of a generation $t > 0$ is $U_B = 2 \ln \frac{3}{4}$ if the bank is not closed and $U_M = \ln \frac{1}{2} + \frac{1}{2} \ln \frac{3}{2} + \frac{1}{2} \ln \frac{1}{2}$ (the market solution) if it is closed. It is easy to check that $U_B > U_M$.

## Notes

1. This interpretation of banking activities is explored in section 8.2.3.

2. This is not true if the bank can *securitize* its loans. However, asymmetric information limits the possibilities of securitization, as discussed later in this chapter. Without securitization, bankruptcy issues become important. If the grocery store around the corner fails and is immediately replaced by another store, its customers essentially lose nothing. This is not the case if a bank fails and is replaced by a new bank. Indeed, the new bank does not have information about the old bank's borrowers and may not be ready to renew their loans. This issue is discussed in chapter 9.

3. For instance, in the case of $N$ individuals confronted with simple independent risks, a single mutual insurance company offering $N$ insurance contracts (one per individual) generates the same diversification as $2^N$ contingent markets would. Indeed, in the Arrow-Debreu framework of state-contingent securities, complete markets are obtained when there is a contingent security for each state of the world. A state of the world is a complete description of the economy (which individuals have an accident, and which have no accident); there are $2^N$ such states, and therefore $2^N$ securities are needed. When $N$ is large, this number becomes astronomical.

4. The ownership structure of real FIs is another problem. The distinction between "genuine" mutuals owned and managed by their customers and stockholder-owned FIs can be analyzed within the general context of corporate governance (see Bhattacharya and Thakor 1993 and the references therein for a discussion of this issue in the specific context of FIs).

5. This is similar to Bester's (1985) solution to the credit-rationing problem of Stiglitz and Weiss (1981), in which firms have to provide collateral to signal the quality of their projects (see chapter 5).

6. Notice, however, that *reputation* is another mechanism for solving the reliability problem. It applies in particular to rating agencies or security analysts, which are not included in our definition of FIs.

7. An argument we fully explore in chapter 9 because it concerns market discipline.

8. See, however, Cerasi and Daltung (2000) for a model of internal agency problems within banks that generates U-shaped average costs for banks.

9. Benston and Smith (1976) state, "The raison d'être for this industry [of financial intermediaries] is the existence of transaction costs" (215).

10. Gehrig (1993) studies an interesting model of trade with search costs, in which the introduction of a monopolistic intermediary surprisingly improves the situation of some traders (see problem 2.9.5).

11. This idea has been modeled by Vale (1993), for instance.

12. Therefore, $\pi_1 + \pi_2 = 1$.

13. Since agents are ex ante identical, we only consider symmetric allocations $(C_1, C_2)$, where an agent's consumption profile does not depend on the agent's identity.

14. We explain later why a riskless bond is the only financial security that can be traded in this market.

15. This is equivalent to the condition

$$\frac{-Cu''(C)}{u'(C)} > 1,$$

which Diamond and Dybvig interpret as saying that the relative index of risk aversion is larger than 1. The idea is that starting from the allocation $(C_1 = 1, C_2 = R)$, depositors are willing to buy insurance against the risk of being of type 1. But risk aversion alone is not enough to imply this, because this liquidity insurance is costly. Increasing expected consumption at $t = 1$ by $\varepsilon$ (i.e., $\pi_1 C_1 = \pi_1 + \varepsilon$) is obtained by decreasing long-term investment of the same amount and therefore decreasing expected consumption at $t = 2$ by $R\varepsilon$ (i.e., $\pi_2 C_2 = R(\pi_2 - \varepsilon)$). Therefore, a stronger condition than just risk aversion is necessary; it is the condition that $C \to Cu'(C)$ is decreasing. Since there is no aggregate uncertainty, a more natural interpretation of this condition is that the intertemporal elasticity of substitution is larger than 1.

16. Chapter 7 discusses the potential coordination failures (bank runs) that may destabilize this arrangement.

17. By convention the project is sold before the investment is undertaken. Thus the total cost to the investor is $1 + P(\theta)$. For any integrable random variable $\tilde{x}$, the notation $E(\tilde{x})$ represents its expectation.

18. In this case, self-finance would be useless and costly. It would entail incomplete insurance for entrepreneurs.

19. This model uses the well-known fact that if $\tilde{x}$ is a normal random variable, and $u(w) = -e^{\rho w}$, then

$$E[u(\tilde{x})] = u\left[E(\tilde{x}) - \frac{1}{2}\rho \operatorname{var}(\tilde{x})\right].$$

20. Leland and Pyle (1977) consider the case in which $\theta$ has a continuous distribution on some interval $[\underline{\theta}, \bar{\theta}]$, which implies the use of more sophisticated techniques.

21. To obtain a complete characterization of equilibria, the reader can check that two inefficient equilibria can arise:

- A pure strategy equilibrium in which only bad projects issue equity, and therefore $P = \theta_1$
- A mixed strategy equilibrium in which $P = \theta_2 - \frac{1}{2}\rho\sigma^2$, and some good-quality projects (but not all of them) issue equity

22. There is a symmetric constraint that good types have no interest to mimic bad types, but it is typically not binding. It is equivalent to

$$\alpha^2 \leq \frac{2(\theta_2 - \theta_1)}{\rho\sigma^2}.$$

23. An alternative interpretation is that a single entrepreneur seeks financing for $N$ independent projects. The implication of result 2.3 is then that project financing involves increasing returns to scale. However, this interpretation has no implication for financial intermediation.

24. Haubrich (1989) considers an alternative way of preventing moral hazard, by assuming that the lender and the borrower are in a long-term relationship in which truthful reporting is induced through punishment schemes that depend on the whole sequence of past messages. Systematic understatements of cash flows can thus be identified and punished. However, a crucial assumption made by Haubrich is the absence of discounting for the future.

25. If the return ($G$ or $B$) were observable, the choice of technology could be inferred ex post. This is why we have to make this disputable assumption. The model of Holmström and Tirole (see section 2.5.3) does not suffer from this drawback.

26. Gorton (1996) has applied this idea to the banks themselves and tested it on a sample of U.S. banks that issued bank notes during the free banking era (1836–1860). His results confirm the existence of a reputation effect.

27. This specification of moral hazard is probably more satisfactory than the one used in sections 2.5.1 and 2.5.2. Indeed, that previous specification relied on the awkward assumption that the success of an investment was verifiable, but not its return.

28. Recall that there is an exogenous cost intermediation $\gamma$—we assume $py + (1 - p)[A + v(y - A)] > 1 + \gamma$.

## References


Akerlof, G. A. 1970. The market for lemons: Qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Allen, F. 1990. The market for information and the origin of financial intermediation. *Journal of Financial Intermediation* 1: 3–30.

Allen, F., and D. Gale. 1997. Financial markets, intermediaries, and intertemporal smoothing. *Journal of Political Economy* 105 (3): 523–546.

Benston, G., and C. W. Smith. 1976. A transaction cost approach to the theory of financial intermediation. *Journal of Finance* 31: 215–231.

Besanko, D., and G. Kanatas. 1993. Credit market equilibrium with bank monitoring and moral hazard. *Review of Financial Studies* 6: 213–232.

Bester, H. 1985. Screening vs rationing in credit markets with imperfect information. *American Economic Review* 75 (4): 850–855.

Bhattacharya, S., and G. Chiesa. 1995. Proprietary information, financial intermediation and research incentives. *Journal of Financial Intermediation* 4: 328–357.

Bhattacharya, S., and A. V. Thakor. 1993. Contemporary banking theory. *Journal of Financial Intermediation* 3: 2–50.

Bolton, P., and X. Freixas. 2000. Equity, bonds, and bank debt: Capital structure and financial market equilibrium under asymmetric information. *Journal of Political Economy* 108 (2): 324–351.

Boot, A., and A. V. Thakor. 1997. Financial system architecture. *Review of Financial Studies* 10 (3): 693–733.

Boyd, J., and E. Prescott. 1986. Financial intermediary coalitions. *Journal of Economic Theory* 38: 211–232.

Broecker, T. 1990. Credit worthiness tests and interbank competition. *Econometrica* 58: 429–452.

Bryant, J. 1980. A model of reserves, bank runs and deposit insurance. *Journal of Banking and Finance* 4: 335–344.

Calomiris, C. W., and C. M. Kahn. 1991. The role of demandable debt in structuring optimal banking arrangements. *American Economic Review* 81 (3): 497–513.

Campbell, T. S., and W. A. Kracaw. 1980. Information production, market signalling, and the theory of financial intermediation. *Journal of Finance* 35: 863–882.

Cerasi, V., and S. Daltung. 2000. The optimal size of a bank: Costs and benefits of diversification. *European Economic Review* 44 (9): 1701–1726.

Diamond, D. W. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51 (3): 393–414.

———. 1991. Monitoring and reputation: The choice between bank loans and directly placed debt. *Journal of Political Economy* 99: 689–721.

———. 1997. Liquidity, banks, and markets. *Journal of Political Economy* 105 (5): 928–956.

Diamond, D. W., and P. H. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91 (3): 401–419.

Diamond, D. W., and R. G. Rajan. 2001. Banks and Liquidity. *American Economic Review* 91 (2): 422–425.

Fama, E. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6 (1): 39–57.

Flannery, M. J. 1994. Debt maturity and the deadweight cost of leverage: Optimally financing banking firms. *American Economic Review* 84 (1): 320–331.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

Gehrig, T. 1993. Intermediation in search markets. *Journal of Economics and Management Strategy* 2: 97–120.

Gilson, S.K.J., and L. Lang. 1990. Troubled debt restructurings: An empirical study of private reorganization of firms in default. *Journal of Financial Economics* 27: 315–353.

Gorton, G. 1996. Reputation formation in early bank note markets. *Journal of Political Economy* 104 (2): 346–397.

Gorton, G., and G. Pennacchi. 1990. Financial intermediaries and liquidity creation. *Journal of Finance* 45 (1): 49–71.

Greenbaum, S. I., G. Kanatas, and I. Venezia. 1989. Equilibrium loan pricing under the bank-client relationship. *Journal of Banking and Finance* 13: 221–235.

Greenbaum, S. I., and A. V. Thakor. 1987. Bank funding modes: Securitization vs deposits. *Journal of Banking and Finance* 11: 379–401.

Grossman, S. J., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Gurley, J., and E. Shaw. 1960. *Money in the theory of finance*. Washington: Brookings Institution.

Haubrich, J. 1989. Financial intermediations: Delegated monitoring and long-term relationships. *Journal of Banking and Finance* 13 (1): 9–20.

Hellwig, M. 1991. Banking, financial intermediation and corporate finance. In *European financial integration*, ed. A. Giovannini and C. Mayer. Cambridge: Cambridge University Press.

Holmström, B., and J. Tirole. 1997. Financial intermediation, loanable funds, and the real sector. *Quarterly Journal of Economics* 112 (3): 663–691.

———. 1998. Private and public supply of liquidity. *Journal of Political Economy* 106: 1–40.

James, C. 1987. Some evidence on the uniqueness of bank loans. *Journal of Financial Economics* 19 (2): 217–235.

Kashyap, A. K., R. G. Rajan, and J. C. Stein. 1999. Banks as liquidity providers: An explanation for the co-existence of lending and deposit taking. DOI 10.2139/ssm.156748.

Krasa, C. M., and A. P. Villamil. 1992. Monitoring the monitor: An incentive structure for a financial intermediary. *Journal of Economic Theory* 57: 197–221.

Kroszner, R. S., and R. G. Rajan. 1994. Is the Glass-Steagall Act justified? A study of the U.S. experience with universal banking before 1933. *American Economic Review* 84 (4): 810–832.

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1335.

Leland, H. E., and D. H. Pyle. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32: 371–387.

McAndrews, J., and W. Roberds. 1995. Banks, payments, and coordination. *Journal of Financial Intermediation* 4 (4): 305–327.

———. 1999. A general equilibrium analysis of check float. *Journal of Financial Intermediation* 8 (4): 353–377.

Mester, L. J. 1992. Traditional and nontraditional banking: An information-theoretic approach. *Journal of Banking and Finance* 16 (3): 545–566.

Millon, M. H., and A. V. Thakor. 1985. Moral hazard and information sharing: A model of financial information gathering agencies. *Journal of Finance* 40 (5): 1403–1422.

Myers, S. C., and N. S. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.

Padilla, J. A., and M. Pagano. 2000. Sharing default information as a borrower discipline device. *European Economic Review* 44 (10): 1951–1980.

Pagano, M., and T. Jappelli. 1993. Information sharing in credit markets. *Journal of Finance* 48 (5): 1693–1718.

Petersen, M., and R. Rajan. 1995. The effect of credit market competition on lending relationships. *Quarterly Journal of Economics* 110: 407–443.

Petersen, M. A., and R. G. Rajan. 1994. The benefits of lending relationships: Evidence from small business data. *The Journal of Finance* 49 (1): 3–37.

Pyle, D. 1971. On the theory of financial intermediation. *Journal of Finance* 26 (3): 737–747.

Qi, J. 1998. Deposit liquidity and bank monitoring. *Journal of Financial Intermediation* 7 (2): 198–218.

Rajan, R. G. 1992. Insiders and outsiders: The choice between informed and arm's-length debt. *Journal of Finance* 47 (4): 1367–1400.

Ramakrishnan, R.T.S., and A. V. Thakor. 1984. Information reliability and a theory of financial intermediation. *Review of Economic Studies* 51: 415–432.

Repullo, R., and J. Suarez. 1995. Credit markets and real economic activity: A model of financial intermediation. Centro de Estudios Monetarios y Financieros, Madrid. Mimeograph.

Rothschild, M., and J. Stiglitz. 1976. Equilibrium in competitive insurance markets. *Quarterly Journal of Economics* 90 (4): 629–649.

Sharpe, S. 1990. Asymmetric information, bank lending and implicit contracts: A stylized model of customer relationships. *Journal of Finance* 45 (4): 1069–1087.

Spence, M. A. 1973. Job market signaling. *Quarterly Journal of Economics* 87 (3): 355–374.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21 (2): 265–293.

Vale, B. 1993. The dual role of demand deposits under asymmetric information. *Scandinavian Journal of Economics* 95 (1): 77–95.

Von Thadden, E. L. 1996. Optimal liquidity provision and dynamic incentive compatibility. Cahiers de Recherches Economiques 9604. Université de Lausanne, Switzerland.

———. 1997. The term structure of investment and the banks' insurance function. *European Economic Review* 41 (7): 1355–1374.

———. 2004. Asymmetric information, bank lending and implicit contracts: The winner's curse. *Finance Research Letters* 1: 11–23.

Yafeh, Y., and O. Yosha. 1995. Large shareholders and banks: Who monitors and how. Discussion paper 1178, Centre for Economic Policy Research, London School of Economics.

Yosha, O. 1995a. Information disclosure costs and the choice of financing source. *Journal of Financial Intermediation* 4: 3–20.

———. 1995b. Arm's length financing and competition in product markets: A welfare analysis. Discussion paper, Tel Aviv University, Israel.

# 3 The Industrial Organization Approach to Banking

The previous chapter presented the asymmetric information justifications of financial intermediation. This chapter focuses on the second pillar of the microeconomic theory of banking, namely the industrial organization (IO) approach. The objective here is twofold. On the one hand, the discussion explores the implications of standard IO theory on the behavior of the banking firm in order to clarify the notions of competitive pricing and market power as well as the implications of monopolistic competition. Still, this view of banks is quite limited because it ignores the specificity of banks and basically reduces banks to ordinary firms.

On the other hand, the implications of banks' specific characteristics on equilibrium prices and quantities is examined. This means taking into account a wider set of variables and strategies than those available to the standard firm. Indeed, banks are able to choose their level of risk, their level of monitoring, and the level of their investment in specific relationships with their customers. This has important consequences on the functioning of the credit market. A correct modeling of these issues is crucial for a better understanding of some empirical results or puzzles that are still lacking a complete theoretical foundation:

- Banks' quoted interest rates (both on deposits and on lows) are sticky, in the sense that they vary less than the market interest rates (Hannah and Berger 1991).
- Despite the development of Internet banking, the median distance between a lending bank and its small-firm customer is 4 miles in the United States (Petersen and Rajan 2002) or 2.25 km (1.4 miles) in Belgium (Degryse and Ongena 2005b). Even if these distances have increased, they still show that distance does matter.
- The possibility of credit rationing (Petersen and Rajan 1995).
- The existence of a "winner's course" that makes banks' lending to new customers less profitable for entrants than for incumbents (Shaffer 1998).

Finally, correct modeling of the banking industry is crucial for understanding the competition versus stability trade-offs that are essential for guiding competition policy in the banking industry.

We start by ignoring some important specificities of banking activities (the risk and informational aspects of these activities). Although simplistic, this approach provides a rich set of models for tackling different issues: monetary policy, market failures (network externalities, switching costs), and some aspects of banking regulation. This chapter focuses on the implications of modeling commercial banks as independent entities that optimally react to their environment, instead of simply considering the banking sector as a passive aggregate, as in the standard approach to monetary policy often found in macroeconomic textbooks.

As in the previous chapter, banks are defined as financial intermediaries that *buy* (possibly nonmarketable) securities of a certain type *(loans)* and *sell* securities of another type *(deposits)*.[1] This discussion takes as given the banking technology (the cost of managing these loans and deposits) and looks at the equilibrium of the banking sector under alternative specifications for the type of competition that prevails in this sector. The chapter starts with the polar cases of *perfect competition* (section 3.1) and *monopoly* (section 3.2). Then it moves to alternative (and possibly more realistic) paradigms: *monopolistic competition* (section 3.3) and *nonprice competition* (section 3.5). Section 3.4 briefly discusses the determination of the range of a bank's activities. Section 3.6 is dedicated to *relationship banking*. Finally, section 3.7 presents the new theory of two-sided markets and its application to payment cards.

## 3.1 A Model of a Perfectly Competitive Banking Sector

### 3.1.1 The Model

This chapter models banking activity as the production of deposit and loan services. Banking technology is represented by a cost function $C(D, L)$, interpreted as the cost of managing a volume $D$ of deposits and a volume $L$ of loans.[2] There are $N$ different banks (indexed by $n = 1, \ldots, N$) with the same cost function $C(D, L)$ that satisfies the usual assumptions of convexity (which implies decreasing returns to scale) and regularity ($C$ is twice differentiable).

The typical balance sheet of a bank is therefore as follows:

| Assets | Liabilities |
|---|---|
| Reserves $R_n$<br>Loans $L_n$ | Deposits $D_n$ |

More precisely, the difference $R_n$ between the volume of deposits $D_n$ that bank $n$ has collected and the volume of loans $L_n$ that the bank has granted is divided into two terms: cash reserves $C_n$ (transferred by bank $n$ on its account at the Central Bank) and the bank's (net) position $M_n$ (positive or negative) on the interbank market. The difference between these two terms is that $C_n$ typically bears no interest and is therefore optimally chosen at its minimum level defined by the regulator. $C_n$ equals a proportion $\alpha$ of deposits. Thus, for all $n$,

$$C_n = \alpha D_n. \tag{3.1}$$

The coefficient $\alpha$ of compulsory reserves may be used as a policy instrument through which the Central Bank tries to influence the quantity of money in circulation in the economy. To complete the picture, a description of the real sector is needed, which consists of three types of agents: the government (including the Central Bank), the firms, and the households. The role of commercial banks is to collect the savings $S$ of households so as to finance the investment needs $I$ of firms. Finally, the government finances its deficit $G$ by issuing securities $B$ (Treasury bills) and high-powered money $M_0$ (the monetary base)[3] used by commercial banks to finance their compulsory reserves at the Central Bank. This model ignores currency (the cash holdings of households and relations with foreign countries); money consists only of the sum of deposits collected by commercial banks ($D = \sum_{n=1}^{N} D_n$). Similarly, the monetary base $M_0$ equals the sum of the reserves of commercial banks on their accounts at the Central Bank (this is the equilibrium condition on the interbank market):

$$M_0 = \sum_{n=1}^{N} C_n = \alpha D. \tag{3.2}$$

In this simplistic framework, the increments in the aggregated balances of each category of agents are as shown in figure 3.1.[4]

### 3.1.2 The Credit Multiplier Approach

The usual description of monetary policy that can be found in elementary macroeconomics textbooks relies on this aggregate description. In this view, a change of the monetary base $M_0$ or an open market operation (a change in $B$) has a direct effect on money and credit because by the preceding conditions,

$$D = \frac{M_0}{\alpha} = \frac{G - B}{\alpha},$$

$$L = M_0\left(\frac{1}{\alpha} - 1\right) = (G - B)\left(\frac{1}{\alpha} - 1\right). \tag{3.3}$$

The money multiplier is defined by the effect of a marginal change in the monetary base (or an open market operation)[5] on the quantity of money in circulation:

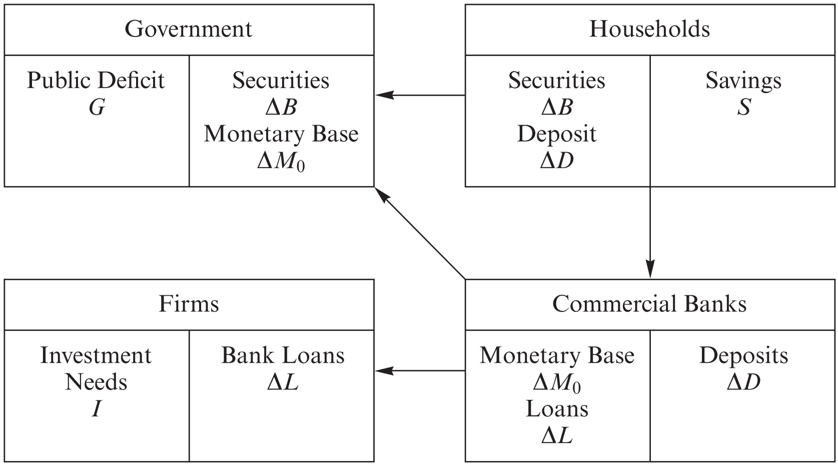



**Figure 3.1**
Increments in aggregated balances of various agents.

$$\frac{\partial D}{\partial M_0} = -\frac{\partial D}{\partial B} = \frac{1}{\alpha} > 0.$$

Similarly, the credit multiplier is defined as the effect on credit of such marginal changes:

$$\frac{\partial L}{\partial M_0} = -\frac{\partial L}{\partial B} = \frac{1}{\alpha} - 1 > 0.$$

The trouble with this simplistic description is that banks are taken as passive entities. Also, modern monetary policy is more accurately described as interventions on the rate $r$ at which the Central Bank refinances commercial banks (assumed equal to the interbank rate). These interventions affect the behavior of commercial banks and therefore the equilibrium interest rates on deposits $(r_D)$ and loans $(r_L)$. To analyze these effects we need to model the individual behavior of commercial banks.

### 3.1.3 The Behavior of Individual Banks in a Competitive Banking Sector

In a competitive model, banks are supposed to be price takers. They take as given the rate $r_L$ of loans, the rate $r_D$ of deposits, and the rate $r$ on the interbank market. Taking into account the management costs, the profit of a bank is given by

$$\pi = r_L L + rM - r_D D - C(D, L),$$

where $M$, the net position of the bank on the interbank market, is given by

$$M = (1 - \alpha)D - L. \tag{3.4}$$

Therefore, $\pi$ can be rewritten as

$$\pi(D, L) = (r_L - r)L + (r(1 - \alpha) - r_D)D - C(D, L). \tag{3.5}$$

Thus the bank's profit is the sum of the intermediation margins on loans and deposits, net of management costs. Because of the assumptions on the cost function $C$, profit-maximizing behavior is characterized by the first-order conditions:

$$\begin{cases} \dfrac{\partial \pi}{\partial L} = (r_L - r) - \dfrac{\partial C}{\partial L}(D, L) = 0, \\ \dfrac{\partial \pi}{\partial D} = (r(1 - \alpha) - r_D) - \dfrac{\partial C}{\partial D}(D, L) = 0. \end{cases} \tag{3.6}$$

**Result 3.1**

1. A competitive bank will adjust its volume of loans and deposits in such a way that the corresponding intermediation margins, $r_L - r$ and $r(1 - \alpha) - r_D$, equal its marginal management costs.
2. As a consequence, an increase in $r_D$ will entail a decrease in the bank's demand for deposits $D$. Similarly, an increase in $r_L$ will entail an increase in the bank's supply of loans $L$. The cross-effects depend on the sign of

$$\frac{\partial^2 C}{\partial D \partial L}.$$

When

$$\frac{\partial^2 C}{\partial D \partial L} > 0$$

(resp. $< 0$), an increase in $r_L$ entails a decrease (resp. an increase) in $D$, and an increase in $r_D$ entails an increase (resp. a decrease) in $L$. When costs are separable,

$$\frac{\partial^2 C}{\partial D \partial L} = 0,$$

cross-effects are nil.

The economic interpretation of the conditions on

$$\frac{\partial^2 C}{\partial L \partial D}$$

is related to the notion of *economies of scope*. When

$$\frac{\partial^2 C}{\partial L \partial D} < 0,$$

an increase in $L$ has the consequence of decreasing the marginal cost of deposits. This is a particular form of economies of scope because it implies that a universal bank that jointly offers loans and deposits is more efficient than two separate entities, specialized, respectively, on loans and deposits. On the contrary, when

$$\frac{\partial^2 C}{\partial L \partial D} > 0,$$

there are *diseconomies of scope*.

***Proof*** Part 1 follows directly from (3.6). Part 2 is obtained by totally differentiating the same system of equations, that is, applying the implicit function theorem to (3.6). For instance, differentiating (3.6) with respect to $r_L$ yields the following result:

$$\begin{cases} 1 = \dfrac{\partial^2 C}{\partial L \partial D} \dfrac{dD}{dr_L} + \dfrac{\partial^2 C}{\partial L^2} \dfrac{dL}{dr_L}, \\ 0 = \dfrac{\partial^2 C}{\partial D^2} \dfrac{dD}{dr_L} + \dfrac{\partial^2 C}{\partial L \partial D} \dfrac{dL}{dr_L}. \end{cases}$$

The Cramer determinant of this system,

$$\delta = \left( \frac{\partial^2 C}{\partial L \partial D} \right)^2 - \frac{\partial^2 C}{\partial L^2} \cdot \frac{\partial^2 C}{\partial D^2},$$

is negative, since $C$ is convex. Solving this system by Cramer's formulas gives

$$\frac{dL}{dr_L} = -\frac{1}{\delta} \frac{\partial^2 C}{\partial D^2} \quad \text{and} \quad \frac{dD}{dr_L} = \frac{1}{\delta} \frac{\partial^2 C}{\partial D \partial L}.$$

Therefore

$$\frac{dL}{dr_L}$$

has the same sign as

$$\frac{\partial^2 C}{\partial D^2},$$

which is positive, and

$$\frac{dD}{dr_L}$$

has the same sign as

$$-\frac{\partial^2 C}{\partial D \partial L}.$$

The consequences of a change in $r_D$ are analyzed in exactly the same terms. ■

### 3.1.4 The Competitive Equilibrium of the Banking Sector

When there are $N$ different banks (indexed by $n = 1, \ldots, N$), each of them is characterized by a loan supply function $L^n(r_L, r_D, r)$ and a deposit demand function $D^n(r_L, r_D, r)$, defined as previously. Let $I(r_L)$ be the investment demand by firms (which, in this simple framework, is equal to their demand for loans, since they do not issue securities), and $S(r_D)$ the savings function of households. (Assume for simplicity that banking deposits and Treasury bills $B$ are perfect substitutes for households: at equilibrium their interest rate is therefore the same.) The competitive equilibrium will be characterized by three equations:

$$I(r_L) = \sum_{n=1}^{N} L^n(r_L, r_D, r) \qquad \text{(loans market)}, \tag{3.7}$$

$$S(r_D) = B + \sum_{n=1}^{N} D^n(r_L, r_D, r) \qquad \text{(savings market)}, \tag{3.8}$$

$$\sum_{n=1}^{N} L^n(r_L, r_D, r) = (1-\alpha) \sum_{n=1}^{N} D^n(r_L, r_D, r) \qquad \text{(interbank market)}, \tag{3.9}$$

where $B$ is the net supply of Treasury bills. Equation (3.9) expresses the fact that the aggregate position of all banks on the interbank market is zero. More generally, a term corresponding to the injection (or drain) of cash by the Central Bank can be added to (or subtracted from) this equation, in which case $r$ becomes a policy variable chosen by the Central Bank. Alternatively, $r$ could be determined in the international capital markets, with an additional term corresponding to the net inflow (or outflow) of funds in (or from) the country. In both cases, $r$ becomes exogenous and (3.9) disappears.

In the case of constant marginal costs of intermediation ($C'_L \equiv \gamma_L, C'_D \equiv \gamma_D$), a simpler characterization of equilibrium is obtained. Equations (3.7) and (3.8) are replaced by a direct determination of $r_L$ and $r_D$,[6] deduced from (3.6):

$$r_L = r + \gamma_L, \tag{3.10}$$

$$r_D = r(1-\alpha) - \gamma_D. \tag{3.11}$$

Then the interest rate $r$ on the interbank market is determined by (3.9), which can also be written as

$$I(r_L) = (1-\alpha)[S(r_D) - B],$$

or

$$S(r(1-\alpha) - \gamma_D) - \frac{I(r+\gamma_L)}{1-\alpha} = B. \tag{3.12}$$

These equations allow us to determine the macroeconomic effects of a marginal shift in the reserve coefficient $\alpha$, or of an open market operation (change in the level of $B$) on the equilibrium level of interest rates $r_L$ and $r_D$. As result 3.2 establishes, the consequences of open market operations and changes in reserve requirements are more complex when the reactions of individual banks are taken into account.

**Result 3.2**

1. An issue of Treasury bills by the government (an increase in $B$) entails a decrease in loans and deposits.[7] However, the absolute values are smaller than in the standard model of section 3.1:

$$\left|\frac{\partial D}{\partial B}\right| < 1, \qquad \left|\frac{\partial L}{\partial B}\right| < 1-\alpha.$$

2. If the reserve coefficient $\alpha$ increases, the volume of loans decreases, but the effect on deposits is ambiguous.

***Proof***

1. Differentiating (3.12) with respect to $B$ (taking into account that $r$ is a function of $B$) yields the following result:

$$\left\{(1-\alpha)S'(r_D) - \frac{I'(r_L)}{(1-\alpha)}\right\}\frac{dr}{dB} = 1.$$

Thus

$$\frac{dr}{dB} = \frac{1}{(1-\alpha)S'(r_D) - \frac{I'(r_L)}{1-\alpha}} > 0.$$

The effect on $D$ of a change in $B$ is now easily obtained, since

$$D(r_D) = S(r_D) - B \quad \text{and} \quad r_D = r(1-\alpha) - \gamma_0.$$

Thus

$$\frac{\partial D}{\partial B} = S'(r_D)(1-\alpha)\frac{dr}{dB} - 1 = \frac{S'(r_D)(1-\alpha)}{(1-\alpha)S'(r_D) - \frac{I'(r_L)}{1-\alpha}} - 1 = \frac{1}{\frac{(1-\alpha)^2 S'(r_D)}{I'(r_L)} - 1}.$$

Since $S'(r_D) > 0$ and $I'(r_L) < 0$,

$$\frac{\partial D}{\partial B} < 0 \quad \text{and} \quad \left|\frac{\partial D}{\partial B}\right| < 1.$$

As for the effect on $L$ (which equals $I$), $L = (1-\alpha)D$; therefore

$$\frac{\partial L}{\partial B} = (1-\alpha)\frac{\partial D}{\partial B}.$$

Thus

$$\frac{\partial L}{\partial B} < 0 \quad \text{and} \quad \left|\frac{\partial L}{\partial B}\right| < 1-\alpha.$$

2. Differentiating (3.12) with respect to $\alpha$ yields

$$\left\{(1-\alpha)S'(r_D) - \frac{I'(r_L)}{1-\alpha}\right\}\frac{dr}{d\alpha} = rS'(r_D) + \frac{I(r_L)}{(1-\alpha)^2} > 0.$$

Since $S'(r_D) > 0$ and $I'(r_L) < 0$,

$$\frac{dr}{d\alpha} > 0.$$

Now, $r_L = r + \gamma_L$ and $r_D = r(1-\alpha) - \gamma_D$. Therefore, if $\alpha$ increases, $r_L$ also increases, and thus the volume of loans decreases. However, the effect on $r_D$ (and deposits) is ambiguous:

$$\frac{dr_D}{d\alpha} = -r + (1-\alpha)\frac{dr}{d\alpha}. \qquad \blacksquare$$

The second part of result 3.2 may be somewhat surprising given that the first-order conditions state that the deposit rate is a decreasing function of the reserve coefficient $\alpha$. But the interbank market rate here is endogenous, and given by (3.12). If the extreme opposite assumption were made of an exogenous interbank market rate $r$ (either controlled by the Central Bank through open market operations or determined by the interest rate on international capital markets under a regime of fixed exchange rates), then the rate on loans would be unaffected by reserve requirements, and only the deposit rate would adjust, as seen from (3.10) and (3.11).

## 3.2 The Monti-Klein Model of a Monopolistic Bank

The assumption of perfect competition may not seem really appropriate for the banking sector, where there are important barriers to entry. An imperfect competition model (oligopoly) is probably more appropriate. For expository reasons, this discussion first studies the Monti-Klein model, which in its simplest version is poles apart from the perfectly competitive model because it considers a monopolistic bank.

### 3.2.1 The Original Model

The Monti-Klein model considers a monopolistic bank confronted with a downward-sloping demand for loans $L(r_L)$ and an upward-sloping supply of deposits $D(r_D)$. In fact, it is more convenient to work with their inverse functions, $r_L(L)$ and $r_D(D)$. The bank's decision variables are $L$ (the amount of loans) and $D$ (the amount of deposits), since its level of equity is assumed to be given. Using the same assumptions and notations as before, the profit of the bank is easily adapted from (3.5), the only difference being that the bank now takes into account the influence of $L$ on $r_L$ (and of $D$ on $r_D$). Assume that the bank still takes $r$ as given, either because it is fixed by the Central Bank or because it is determined by the equilibrium rate on international capital markets:

$$\pi = \pi(L, D) = (r_L(L) - r)L + (r(1 - \alpha) - r_D(D))D - C(D, L). \tag{3.13}$$

The bank's profit is, as before, the sum of the intermediation margins on loans and on deposits minus management costs. In order for the maximum of $\pi$ to be characterized by the first-order conditions, assume that $\pi$ is concave. The first-order conditions, which equate marginal revenue and marginal cost, are

$$\frac{\partial \pi}{\partial L} = r_L'(L)L + r_L - r - C_L'(D, L) = 0, \tag{3.14}$$

$$\frac{\partial \pi}{\partial D} = -r_D'(D)D + r(1 - \alpha) - r_D - C_D'(D, L) = 0. \tag{3.15}$$

Now the elasticities of the demand for loans and the supply of deposits are introduced:[8]

$$\varepsilon_L = -\frac{r_L L'(r_L)}{L(r_L)} > 0 \quad \text{and} \quad \varepsilon_D = \frac{r_D D'(r_D)}{D(r_D)} > 0.$$

The solution $(r_L^*, r_D^*)$ of (3.14) and (3.15) can then be characterized by[9]

$$\frac{r_L^* - (r + C_L')}{r_L^*} = \frac{1}{\varepsilon_L(r_L^*)}, \tag{3.16}$$

$$\frac{r(1-\alpha) - C_D' - r_D^*}{r_D^*} = \frac{1}{\varepsilon_D(r_D^*)}. \tag{3.17}$$

These equations are simply the adaptation to the banking sector of the familiar equalities between Lerner indices (price minus cost divided by price) and inverse elasticities. The greater the market power of the bank on deposits (resp. loans), the smaller the elasticity and the higher the Lerner index. The competitive model corresponds to the limit case of infinite elasticities, which goes back to equation (3.6). Therefore, the intuitive result is that intermediation margins are higher when banks have a higher market power.

**Result 3.3** A monopolistic bank sets its volume of loans and deposits in such a way that the Lerner indices equal inverse elasticities.

An immediate consequence of this result is that intermediation margins will be adversely affected if substitutes to banking products appear on financial markets (e.g., when households have access to money market funds as substitutes for banking deposits, and when firms issue securities on financial markets as a substitute for bank loans).

An interesting consequence follows:

**Result 3.4** If management costs are additive, the bank's decision problem is separable. The optimal deposit rate is independent of the characteristics of the loan market, and the optimal loan rate is independent of the characteristics of the deposit market.

### 3.2.2 The Oligopolistic Version

Of course, one may question the practical relevance of these results, since the banking industry is clearly not controlled by a unique firm. In fact, the main interest of the Monti-Klein model is that it can easily be reinterpreted as a model of imperfect (Cournot) competition between a finite number $N$ of banks, which is a more accurate description of reality.[10] Indeed, consider the case of $N$ banks (indexed by $n = 1, \dots, N$) supposed for simplicity to have the same cost function, taken to be linear:

$$C(D, L) = \gamma_D D + \gamma_L L.$$

A Cournot equilibrium of the banking industry is an $N$-tuple of couples $(D_n^*, L_n^*)_{n=1,\dots,N}$ such that for every $n$, $(D_n^*, L_n^*)$ maximizes the profit of bank $n$ (taking the volume of deposits and loans of other banks as given). In other words, for every $n$, $(D_n^*, L_n^*)$ solves

$$\max_{(D_n, L_n)} \left\{ \left( r_L \left( L_n + \sum_{m \neq n} L_m^* \right) - r \right) L_n \right.$$
$$\left. + \left( r(1-\alpha) - r_D \left( D_n + \sum_{m \neq n} D_m^* \right) \right) D_n - C(D_n, L_n) \right\}.$$

It is easy to see that there is a unique equilibrium, in which each bank sets $D_n^* = D^*/N$ and $L_n^* = L^*/N$. The first-order conditions are

$$\frac{\partial \pi_n}{\partial L_n} = r_L'(L^*)\frac{L^*}{N} + r_L(L^*) - r - \gamma_L = 0,$$

$$\frac{\partial \pi_n}{\partial D_n} = -r_D'(D^*)\frac{D^*}{N} + r(1-\alpha) - r_D(D^*) - \gamma_D = 0.$$

These first-order conditions can also be rewritten as

$$\frac{r_L^* - (r + \gamma_L)}{r_L^*} = \frac{1}{N\varepsilon_L(r_L^*)}, \tag{3.18}$$

$$\frac{r(1-\alpha) - \gamma_D - r_D^*}{r_D^*} = \frac{1}{N\varepsilon_D(r_D^*)}. \tag{3.19}$$

Comparing these with (3.16) and (3.17), one can see that the only difference between the monopoly case and the Cournot equilibrium is that the elasticities are multiplied by $N$. With this simple adaptation, the Monti-Klein model can be reinterpreted as a model of imperfect competition with two limiting cases: $N = 1$ (monopoly), and $N = +\infty$ (perfect competition).

Notice that equations (3.18) and (3.19) provide a possible test of imperfect competition on the banking sector. Indeed, from these equations, the sensitivity of $r_L^*$ and $r_D^*$ to changes in the money market rate $r$ depends on $N$, which is a proxy for the intensity of competition ($N = 1$ may be interpreted as pure cartelization, whereas $N = +\infty$ corresponds to perfect competition). Assuming for simplicity that elasticities are constant,

$$\frac{\partial r_L^*}{\partial r} = \frac{1}{1 - \frac{1}{N\varepsilon_L}} \quad \text{and} \quad \frac{\partial r_D^*}{\partial r} = \frac{1-\alpha}{1 + \frac{1}{N\varepsilon_D}}.$$

Therefore, as the intensity of competition increases ($N$ grows), $r_L^*$ (resp. $r_D^*$) becomes less (resp. more) sensitive to changes in $r$.

### 3.2.3 Empirical Evidence

Although the Monti-Klein model presents a very simplified approach to the banking activity, it is clear that the model provides a series of conclusions that seem particularly natural and appealing, and that can be confirmed with empirical evidence. Of course, there are also features of deposit contracts that the model cannot explain, but it seems reasonable that market power will lead banks to quote lower deposit rates and higher rates on loans. Indeed, the empirical findings, since an early contribution by Edwards (1964), show that the different interest rates charged on loans can be viewed as reflecting different elasticities of demand.

An interesting test of the industrial organization models of banking is provided by examining the reaction of deposit rates and credit rates to the fluctuations of money market interest rates (say, the T-bill rate). Berger and Udell (1992) show, for example, that credit rates are relatively sticky, in the sense that an increase of 1 percent in the T-bill rate only leads to an increase of 50 basis points (0.5 percent) of credit rates. Hannah and Berger (1991) show that this stickiness increases with market concentration, in accordance with the predictions derived from IO models. However, some other empirical findings are more difficult to explain by simple IO models. For example, Newmark and Sharpe (1992) show that the deposit rate response to T-bill rate fluctuations is asymmetric. Indeed, Newmark and Sharpe establish that deposit rates adjust faster when they are low, and slower when they are high.

The empirical studies of competition in the banking industry have documented the effect of market concentration on both deposit and loan spreads. The excellent surveys of Berger et al. (2004) and Degryse and Ongena (2005a) provide a rigorous perspective on this field.

## 3.3 Monopolistic Competition

The concept of monopolistic competition, first introduced by Chamberlin (1933), has been extensively used in the theory of industrial organization. It can be summarized as follows. As soon as there is some degree of differentiation between the products sold by competing firms, price competition will lead to less extreme outcomes than in pure Bertrand models. One of the most popular models of monopolistic competition is the location model of Salop (1979), in which product differentiation is generated by transportation costs. This section presents three applications of the Salop model to the banking sector, in increasing order of complexity, designed to address three different questions: (1) Does free competition lead to an optimal number of banks (section 3.3.1)? (2) What is the effect of deposit rate regulation on credit rates (section 3.3.2)? (3) Does free competition lead to an appropriate level of interbank cooperation in automated teller machine (ATM) networks (section 3.3.3)?

### 3.3.1 Does Free Competition Lead to the Optimal Number of Banks?

The simplest formulation of the banking version of the Salop model considers a continuum of depositors, each endowed with one unit of cash and uniformly distributed along a circle. There are $n$ banks (indexed by $i = 1, \ldots, n$), located on the same circle, that collect the deposits from the public and invest them into a riskless technology (or security) with a constant rate of return $r$. Depositors do not have access to this technology; they can only deposit their money in a bank. Moreover, when each depositor does so, he or she incurs a transportation cost $tx$, proportional[11] to the distance $x$ between the depositor's location and that of the bank.[12] The total length of the circle is normalized to 1, and the total mass of depositors is denoted $D$.

Depositors being uniformly distributed, the optimal organization of the banking industry corresponds to a symmetric location of the $n$ banks. The maximal distance traveled by a consumer is $1/2n$, and the sum of all depositors' transportation costs can be computed by dividing the circle into $2n$ equal arcs,

$$2n \int_0^{1/2n} txD\,dx = \frac{tD}{4n}. \tag{3.20}$$

The unit cost of setting up a bank is denoted by $F$. The optimal number of banks is obtained by minimizing the sum of setup costs and transportation costs:

$$nF + \frac{tD}{4n}.$$

Disregarding indivisibilities (the fact that $n$ is an integer), the minimum of this expression is obtained when its derivative with respect to $n$ vanishes:

$$F - \frac{tD}{4n^2} = 0,$$

which gives

$$n^* = \frac{1}{2}\sqrt{\frac{tD}{F}}. \tag{3.21}$$

How many banks will appear if banking competition is completely free (no entry restrictions, no rate regulations)? To answer this question, consider that $n$ banks enter simultaneously,[13] locate uniformly on the circle, and set deposit rates $r_D^1, \ldots, r_D^n$. To determine the volume $D_i$ of deposits attracted by bank $i$ ($i = 1, \ldots, n$) in this situation, it is necessary to compute the location of the "marginal depositor" who is indifferent about going to bank $i$ or bank $i+1$ (fig. 3.2).

The distance $\hat{x}_i$ between this marginal depositor and bank $i$ is defined by

$$r_D^i - t\hat{x}_i = r_D^{i+1} - t\left(\frac{1}{n} - \hat{x}_i\right). \tag{3.22}$$

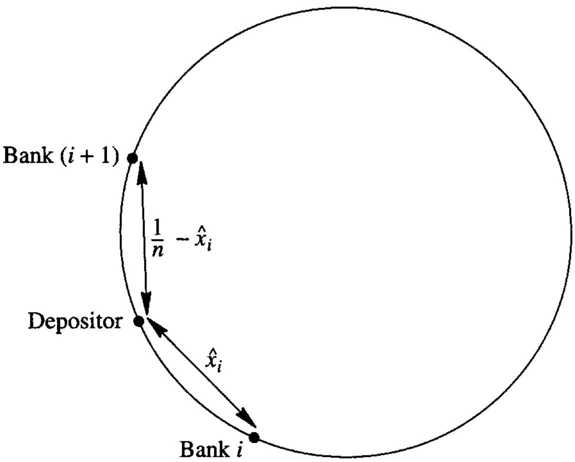



**Figure 3.2**
Locations on Salop circle.

Therefore,

$$\hat{x}_i = \frac{1}{2n} + \frac{r_D^i - r_D^{i+1}}{2t},$$

and the total volume of deposits attracted by bank $i$ is

$$D_i = D\left[\frac{1}{n} + \frac{2r_D^i - r_D^{i+1} - r_D^{i-1}}{2t}\right].$$

Since this example uses a circle, the following conventions are adopted: $r_D^{n+1} = r_D^1$ and $r_D^0 = r_D^n$.

The profit of bank $i$ is thus

$$\pi_i = D(r - r_D^i)\left(\frac{1}{n} + \frac{2r_D^i - r_D^{i+1} - r_D^{i-1}}{2t}\right).$$

The equilibrium is obtained when for all $i$, $r_D^i$ maximizes $\pi_i$ (while other rates are kept constant). This is equivalent to

$$r - r_D^i = \frac{t}{n} + \frac{2r_D^i - r_D^{i+1} - r_D^{i-1}}{2}, \qquad (i = 1, \dots, n). \tag{3.23}$$

This linear system has a unique solution:

$$r_D^i = \cdots = r_D^n = r - \frac{t}{n},$$

which gives the same profit to all the banks:

$$\pi_1 = \cdots = \pi_n = \frac{tD}{n^2}.$$

Since there are no entry restrictions, the equilibrium number of banks (denoted $n_e$) will be obtained when this profit is equal to the setup cost $F$, which gives

$$n_e = \sqrt{\frac{tD}{F}}. \tag{3.24}$$

A comparison with (3.21) shows that free competition leads to too many banks. Consequently, there is potentially some scope for public intervention. The question is now to determine which type of regulation is appropriate. For instance, in such a context, imposing a reserve requirement on deposits is equivalent to decreasing the return rate $r$ on the banks' assets. Equation (3.24) shows that this has no effect on the number of active banks at equilibrium. On the contrary, any measure leading directly (entry, branching restrictions) or indirectly (taxation, chartering fees, capital requirements) to restricting the number of active banks will be welfare-improving as long as the whole market remains served. This can be seen in particular as a justification of branching restrictions, which exist or have existed in many countries.[14] However, the robustness of this result[15] is questionable. Other models of industries with differentiated products actually lead to too *few* products at equilibrium (see Tirole 1988, ch. 7, for a discussion of this issue).

### 3.3.2 The Impact of Deposit Rate Regulation on Credit Rates

Section 3.2, using the Monti-Klein model, concluded that if the markets for deposits and loans are independent, the impact on loan rates of imposing a maximum deposit rate is determined by the properties of the cost function of the bank.[16] In particular, if this cost function is separable between deposits and loans, the pricing of loans is independent of the deposit rates. Chiappori, Perez-Castrillo, and Verdier (1995) have studied the same question in a different context, in which the demands for loan and deposit services originate from the same consumers. They use an extension of the model of section 3.3.1 in which credit activity is introduced. Depositors are also borrowers,[17] with an inelastic credit demand $L$ at the individual level. Assume $L < 1$. The total (net) utility of a typical consumer (depositor-borrower) is therefore

$$U = (1 + r_D) - t_D x_D - (1 + r_L)L - t_L x_L, \tag{3.25}$$

where $x_D$ (resp. $x_L$) is the distance from the bank where the consumer's cash has been deposited (resp. where the consumer's loan has been granted), $r_L$ is the loan rate, and $t_D$ and $t_L$ are the transportation cost parameters for deposits and loans.

Note that transportation costs for loans and deposits may be different (e.g., because the frequencies of these transactions are different) and that the consumer may use different banks for deposits and loans (this issue is discussed later).

A straightforward adaptation of the results of section 3.3.1 shows that if $n$ banks enter, locate symmetrically on the circle, and compete in deposit rates and loan rates, the equilibrium is symmetric. All banks offer the same rates:

$$r_D^e = r - \frac{t_D}{n}, \qquad r_L^e = r + \frac{t_L}{nL}. \tag{3.26}$$

They share the market equally and obtain a profit

$$\pi^e = \frac{D(t_D + t_L)}{n^2}. \tag{3.27}$$

The number of active banks in a free-entry equilibrium is determined by the equality between $\pi^e$ and the entry cost $F$, which gives

$$n^e = \sqrt{\frac{D(t_D + t_L)}{F}}. \tag{3.28}$$

It is easy to see that loans and deposits are independently priced. If deposit rates are regulated (e.g., if $r_D$ is fixed at zero), this has no effect on $r_L$. The only thing that changes is that banks make more profit on deposits, so that more banks enter, which is welfare-decreasing.

However, another pattern appears if banks are allowed to offer tying contracts. Such contracts are defined by the fact that consumers can obtain credit from a bank only if they deposit their cash in the same bank (another possibility is that they get a lower credit rate if they do so). Chiappori, Perez-Castrillo, and Verdier (1995) show that such contracts would never emerge at equilibrium if banks were unregulated. However, if the remuneration of deposits is forbidden, attracting depositors is highly profitable to the banks. Therefore, they are ready to subsidize credit in order to do so.

**Result 3.5** Under deposit rate regulation, banks will offer tying contracts with lower credit rates than in the unregulated case. Therefore, the regulation is effective: it leads to lower credit rates. If deposit rate regulation is maintained, a prohibition of tying contracts is welfare-decreasing.[18]

***Proof*** Recall the expressions of the utility of a typical consumer,

$$U = 1 + r_D - t_D x_D - (1 + r_L)L - t_L x_L,$$

and the profit of a typical bank,

$$\pi = 2D[\hat{x}_D(r - r_D) + L\hat{x}_L(r_L - r)],$$

where $\hat{x}_D$ (resp. $\hat{x}_L$) represents the distance between the bank and its marginal depositor (resp. borrower). For symmetry, this distance is assumed to be the same on both sides of the bank.

If deposit rates are regulated, $r_D = 0$. Moreover, if tying contracts are forbidden, depositors simply go to the closest bank, so that $\hat{x}_D = 1/2n$. The equilibrium loan rates and profits are

$$r_L^0 = r + \frac{t_L}{nL}, \qquad \pi^0 = \frac{D}{n}\left[r + \frac{t_L}{n}\right]. \tag{3.29}$$

Notice that $r_L^0 = r^L$, but $\pi^0 > \pi^e$ (compare with (3.27) and (3.28)).

Suppose now that deposit rates are still regulated but that tying contracts are allowed. The previous situation,

$$r_L = r + \frac{t_L}{nL},$$

is no longer an equilibrium. By offering a tying contract (deposit plus loan) at a slightly lower loan rate, a bank simultaneously gains more customers and a higher profit margin (since each deposit brings a rent $r$). Therefore, all banks use such contracts so that all consumers choose the same bank for deposits and loans: $x_D = x_L$. The distance $\hat{x}$ between a bank and the marginal consumer is determined by

$$1 - (t_L + t_D)\hat{x} - (1 + r_L)L = 1 - (t_L + t_D)\left(\frac{1}{n} - \hat{x}\right) - (1 + r_L')L,$$

where $r_L$ (resp. $r_L'$) denotes the loan rate offered by the bank (resp. its neighbors). The following result is obtained:

$$\hat{x} = \frac{1}{2n} + \frac{L(r_L' - r_L)}{2(t_L + t_D)}.$$

The expression of the bank's profit is

$$\pi = 2D\hat{x}(r + L(r_L - r)). \tag{3.30}$$

The maximization of $\pi$ with respect to $r_L$ ($r_L'$ being fixed) is characterized by

$$\frac{1}{\pi}\frac{\partial \pi}{\partial r_L} = -\frac{L}{2\hat{x}(t_L + t_D)} + \frac{L}{r + L(r_L - r)} = 0,$$

or

$$r_L = r - \frac{r - 2\hat{x}(t_L + t_D)}{L}.$$

By symmetry, at equilibrium $\hat{x} = 1/2n$, so that the new equilibrium loan rate is

$$r_L^1 = r - \frac{1}{L}\left(r - \frac{t_L + t_D}{n}\right),$$

which can also be written

$$r_L^1 = \left(r + \frac{t_L}{Ln}\right) - \frac{1}{L}\left(r - \frac{t_D}{n}\right), \tag{3.31}$$

$$= r_L^e - \frac{r_D^e}{L} < r_L^e. \tag{3.32}$$

This establishes the first part of result 3.5.

The proof of the second part simply results from the remark that (3.30) and (3.31) imply that the equilibrium profit with regulation and tying contracts equals $\pi^e$, the equilibrium profit in the absence of regulation.[19] As already stated, if deposit rate regulation is maintained while tying contracts are prohibited, the equilibrium profit is $\pi^1 > \pi^e$. Therefore, the equilibrium number of banks is higher and welfare is decreased. ∎

### 3.3.3 Bank Network Compatibility

An interesting application of Salop's model is the contribution of Matutes and Padilla (1994).[20] Their model considers a two-stage game in which in the first stage the banks choose whether they want to belong to some ATM network, and in the second one they compete on prices (that is, deposit rates).

Compatibility has no physical cost and yields benefits to the depositors, so full compatibility is welfare-maximizing. Still, full compatibility will never emerge as an equilibrium of the two-stage game. Indeed, the banks know that if they become fully compatible, stronger competition during the second stage will lower their profit, so there is an opportunity cost of compatibility.

For the case of three banks, it is possible to show (see problem 3.8.2) that if equilibrium exists, it can be either with three incompatible networks or with two banks sharing their network and leaving the third bank outside (which implies that an ex ante symmetric situation will yield an asymmetric equilibrium outcome), so the equilibrium is not efficient.

In addition, Matutes and Padilla show that the existence of switching costs tends to reduce the incentives of banks to become compatible. Note that this result may apply as well to other networks (e.g., in the emergence of clearinghouses). The non-efficiency of equilibrium may justify some form of public intervention.

### 3.3.4 Empirical Evidence

A number of contributions have established that the distance from the lending bank to its small-firm customers is a key aspect of lending. This is unexpected because the transportation cost in Salop's model was not initially meant to be interpreted literally but as reflecting product differentiation. But the empirical evidence seems to point out that distance does matter. The phenomenon is also surprising in the context of the improvement in communication technology, in an era when customers can do all their banking by phone or via Internet. In fact, it should be emphasized that distance only matters for some types of loans, those granted to opaque borrowers where lending is based on soft information gathered by the bank branch's loan officer. For hard information–based loans, like consumer loans or mortgages, distance need not be relevant. Because of this difference that depends upon the transparency of the borrower, that is, because of the balance between soft and hard information, the effect of distance on the terms and conditions of lending is interpreted as reflecting monitoring costs.

The importance of distance is emphasized in Degryse and Ongena (2005b), who develop an empirical analysis of spatial discrimination in loan pricing. The issue is directly connected to Salop's model and has interesting implications regarding the nature of competition. If we extend Salop's model to allow for price discrimination, it becomes clear that those borrowers located close to the bank branch will be charged more than those located close to a competitor. This is confirmed by the empirical evidence. A borrower located close to the lender pays a higher interest rate, and in addition this rate is higher when the closest competitor is far away.

Petersen and Rajan (2002) also acknowledge the importance of distance but focus on its evolution. They show that for loans to small firms, the mean (as well as the median or the first quartile) of the distance to the lending bank branch has steadily increased since the 1970s. They argue that this increment cannot be attributed to changes in exogenous variables related to distance, such as banks' density, firms' transparency, or types of loans, but that it is explained by an increase in productivity and in capital intensity of lending, which correspond to the improvement in communication technology.

## 3.4 The Scope of the Banking Firm

An issue directly related to the role of financial intermedaries is the determination of their range of activities. Boot and Thakor (1997) study the relation between bank structure and financial innovation. They show that the incentives to innovate are weaker in a universal bank than under functional separation.

To see this, consider a model where two types of banks coexist: commercial banks and investment banks. Commercial banks grant loans to the firms that do not issue securities. Investment banks underwrite the securities issued by the other firms and in addition produce financial innovations that increase the profits of firms by some amount $S$. Financial innovations occur independently in each bank, with a probability $z$ depending on the level of research activity chosen by the bank. The cost of financial research is denoted $C(z)$, an increasing, convex function of $z$. Assume for simplicity that there are two investment banks and that competition drives their profit down to zero unless one of them (and only one) innovates.

Three situations may arise. With probability $(1-z)^2$ no innovation occurs. With probability $2z(1-z)$ only one of the investment banks innovates and is therefore able to appropriate the whole surplus on its clients. Finally, with probability $z^2$ the two investment banks innovate and the firms share the whole surplus $S$. In equilibrium the marginal expected cost of innovation equals the marginal expected profit from innovation.

It is easy to see that in this context universal banks innovate less than under functional separation. This is so because a universal bank "internalizes the depressing effect that the innovation will have on the loan demand faced by the commercial bank unit" (Boot and Thakor 1997, 1118).

## 3.5 Beyond Price Competition

As it happens in other industries, where firms may compete in dimensions other than price, such as quality or capacity, a more precise modeling of competition in the banking industry requires considering other strategic variables that have an impact on the characteristics of the banks' deposits and loans. For example, banks have to set the level of risk they take on their investments, and the intensity of monitoring and screening of their borrowers.

Nonprice competition is a key issue from the point of view of efficiency of the banking industry as well as from the perspective of banking regulation. In a standard Arrow-Debreu economy, competition always enhances efficiency, and the only causes of concern are fixed costs or increasing returns, which may lead to natural monopoly (or natural oligopoly) situations. In the banking industry, where banks choose the (noncontractible) level of their assets' risk or monitoring effort, price competition might negatively affect the level of these variables, so that the overall effect of competition may be to reduce efficiency. In some cases, restricting entry and leaving a positive charter value to each bank may be the price to pay for bank stability.

### 3.5.1 Risk Taking on Investments

The general consensus is that more competition leads banks to increase their riskiness. This is confirmed by the empirical evidence on the savings and loans crisis

(Keeley 1990) as well as by theoretical models about the effect of deposit insurance. As discussed in chapter 9, deposit insurance can be viewed as a put option. Therefore, absent any cost of bankruptcy (such as the opportunity cost of the loss of future revenues), maximizing the value of this put option leads to choosing maximum risk.[21]

Cordella and Yeyati (2002) establish the robustness of this effect by considering a model where banks are horizontally differentiated. They show that increased competition reduces product differentiation and margins. This gives an incentive for bank managers to take higher risks. Beck et al. (2003) study the effect of bank concentration on the probability of a banking crisis. Using a panel data set with 79 countries on an 18-year period, they show that crises are less likely in a more concentrated banking system, but that entry restrictions increase the probability of a banking crisis.

Keeley (1990), Suarez (1994), and Matutes and Vives (1996) consider the risk-taking decisions of banks in a competitive setting, where depositors are not insured. To illustrate their main findings, consider a bank that is able to choose the level of riskiness, $\sigma$, of its portfolio of loans. Riskless rates are normalized to zero. We assume that when $\sigma$ increases, loans become more risky in the sense of Rothschild and Stiglitz (1970). This means that the gross expected return on loans is constant, $E(r_L|\sigma) \equiv \mu$, but that the dispersion of their returns increases (mean preserving spread) so that $E[h(r_L)|\sigma]$ increases with $\sigma$ for any increasing convex function $h$. The bank's assets consist of a volume $L$ of loans. Its liabilities consist of a volume $D$ of deposits, remunerated at the rate $r_D(\sigma)$, which depending on the assumptions regarding information, may reflect the risk taken by the bank. There is no equity; thus budget balance implies $L = D$. The bank fails whenever the realized return on loans $r_L$ is less than $r_D(\sigma)$. We denote by $p(\sigma) = \Pr(r_L > r_D(\sigma)|\sigma)$ the probability of success.

The charter value $V$ of the bank is equal to the discounted value of its expected future profits. In a perfectly competitive environment, this value equals zero because banks do not need any capital. When there are barriers to entry or restrictions on interest rates, this value can be positive.

**Perfect Information**

Assume that $\sigma$ is observed by depositors, and that $r_D$ is set after $\sigma$ is chosen by the bank. In this case, depositors will require an interest rate on deposits $r_D(\sigma)$ that provides them with (at least) a zero expected return. Given limited liability, depositors anticipate the bank's default whenever $r_L < r_D(\sigma)$, in which case they seize the banks' assets and obtain a return $r_L$ (liquidation costs are neglected). Therefore, depositors participate only if

$$E[\min(r_L, r_D(\sigma)|\sigma] \geq 1.$$

Let $\Pi(\sigma)$ denote the current expectation of the bank's profit given $\sigma$,

$$\Pi(\sigma) = D \cdot E[\max(0, r_L - r_D(\sigma)|\sigma].$$

The bank will choose $\sigma$ so as to maximize $\Pi(\sigma) + p(\sigma)V$ under the participation constraint of depositors, which will be binding at the optimum. But since $\max(0, r_L - r_D) + \min(r_L, r_D) = r_L$, we see, by taking expectations, that

$$\Pi(\sigma) + D = E(r_L)D = \mu D,$$

so that $\Pi(\sigma) = (\mu - 1)D$ is independent of $\sigma$. Thus, whenever $V > 0$, the risk of losing the charter value will be enough to discipline the bank: it will choose the minimal level of riskiness for its loans. In the limit, if the bank faces a competitive setting, $V = 0$, and the level of risk $\sigma$ is indeterminate.[22]

**Imperfect Information**

Assume that $r_D$ is chosen by depositors without knowing $\sigma$. In this case, depositors anticipate that the banks will choose the profit-maximizing level of risk.

Thus depositors anticipate a risk level $\hat{\sigma}$, and require a return $r_D(\hat{\sigma})$, for which they obtain a zero return. Banks set the level of risk $\sigma^*$ that solves

$$Max_\sigma \; \Pi(\sigma) + p(\sigma)V,$$

and rational expectations imply

$$\hat{\sigma} = \sigma^*.$$

But, in contrast to the previous case, the bank's current profit is written as $\Pi(\sigma) = D \cdot E[\max(0, r_L - r_D(\hat{\sigma}))|\sigma]$, which is a convex function and therefore an increasing function of $\sigma$. Thus, in the competitve case ($V = 0$), the banks choose the maximum level of risk, and this is rationally anticipated by depositors. In the more general case, there is a trade-off between short-term and long-term profits, so there exists a threshold $\hat{V}$ such that for $V < \hat{V}$ the bank still chooses a maximum level of risk, whereas for $V > \hat{V}$ it will choose the minimum one.

To summarize, the choice of the level of risk on the bank's assets depends on depositors' information and on the charter value of the bank.

**Result 3.6** When banks are able to choose their level of riskiness without affecting the expected value of their assets, two situations arise:

- If the level of risk is observable by liability holders, a positive charter value (however small) will be sufficient to give the bank an incentive to choose the minimal level of risk.
- If the level of risk is not observable by liability holders, the bank will choose the maximum risk level, except in the case of a sufficiently large charter value.

There is a growing literature that builds on these simple intuitions. The literature on market discipline addresses the issue of the ability of banks' creditors to effectively assess the risk behavior of banks. Another strand considers the link between competition and market efficiency, arguing that perfect competition may lead to excessive risk taking by banks.

It is interesting to emphasize, as Suarez (1994) does, that a bank shareholders' option to recapitalize the bank increases the bank's risk-taking behavior. This may seem surprising because ex post recapitalization appears the simplest way to avoid bank failure. Still, ex ante, when the bank has to choose the risk level for its investment, this possibility fosters risk taking. The reason is simply that the loss of the charter value in case of bankruptcy weights against risk taking in the bank's decision. Since recapitalization allows preserving the charter value in case of losses, the risk of losing the charter value is reduced. This provides incentives for excessive risk taking.

**The Trade-off between Competition and Financial Stability**

Even this simple model is able to shed some light on the possible trade-offs between competition and financial stability that may arise in the banking industry. Indeed, in the absence of banks' equity, perfect competition implies zero future profits and therefore a zero charter value. As a consequence, in a model of asymmetric information, competition increases banks' risk-taking incentives. Three implications can be derived that are of interest from a regulatory point of view.

First, since under perfect information the bank's managers have no incentive to take risks, transparency solves the dilemma. Thus the regulator should foster the disclosure of the bank's risk level.

Second, imposing capital requirements implies that the bank has a charter value. The value of the bank's capital, even under perfect competition, thus provides a second way out of the competition-stability trade-off.[23]

Third, since relationship banking increases a bank's charter value, it also decreases its incentives to gamble.[24]

Allen and Gale (2000, ch. 8) suggest an alternative model to explore the relation between risk and return that does not lead to the simple solutions obtained in Matutes and Vives (1996). It assumes there is a continuous choice of the level of output $y$, and with each level of $y$ is associated a probability of success, $p(y)$. In case of failure the cash flow is zero. To represent an interesting risk-return choice, it is assumed that $p(y)$ is a decreasing concave function. The efficient project choice $y^*$ in a risk-neutral world is the one that maximizes the expected net present value and is therefore characterized by the first-order conditions

$$y\frac{dp(y)}{dy} + p(y) = 0.$$

This will be the choice undertaken by a bank financed through equity as well as by a bank that is paying the market value of its debt. In this case, the bank maximizes the market value of its assets, which is not surprising, because the Modigliani-Miller theorem applies. For a leveraged bank that is not subject to market discipline, so that repayment $R$ does not depend upon $y$, the first-order condition becomes

$$(y - R)\frac{dp(y)}{dy} + p(y) = 0.$$

Implying a bias toward more risky decisions, the value of the objective function for the efficient project $y^*$ is

$$-R\frac{dp(y)}{dy} > 0.$$

The result is that competition on bank funds will increase $R$ and lead to higher risk levels in the banks' choice of assets. This result is in line with the standard view that competition threatens financial stability.

Boyd and De Nicoló (2005) invert the logic of this result by assuming that the risk choice is made at the level of the borrowing firm and increases with the level of the loans rate $r_L$. Banks are then unable to choose their risk level, which is inherited from the firms' behavior. Boyd and De Nicoló then consider the best strategy for the banks' taking into account that its supply of deposits affects the level of both interest rates (on deposits and on loans). They show that when the number of banks increases, the firms' risk level decreases. Their result challenges the conventional view. If firms can control their level of risk, competition enhances financial stability.

These different approaches to the link between banking competition and financial stability illustrate the complexity of the issue. The empirical literature reinforces this point because both concentration and financial stability can be measured in different ways (see Boyd and De Nicoló for a review of the empirical literature). In addition, bank concentration can represent a form of financial barrier to entry in the product market, as established by Cetorelli and Strahan (2006).

### 3.5.2 Monitoring and Incentives in a Financial Conglomerate

The analysis of the level of monitoring can be performed with the same methodology. Assume the level of monitoring is measured by the bank's probability of solvency, $m$, and that the cost of monitoring is $c(m)$.[25] The bank's repayment for the unit of funds it borrows will be $R(m)$ under perfect information. Its income is denoted by $y$. In a competitive risk-neutral, zero-interest-rate market, we will have

$$R(m) = \frac{1}{m}.$$

Consequently the bank's choice of monitoring level will be the solution to

$$\max_m m\left(y - \frac{1}{m}\right) - c(m),$$

yielding the efficient monitoring level characterized by

$$y = c'(m^*).$$

Yet, under imperfect information, the repayment $\hat{R}$ will not depend upon the effective monitoring level of the bank. The level of monitoring will be characterized by

$$y - \hat{R} = c'(\hat{m}),$$

and because of the convexity of the cost function, this implies that the bank will choose too low a level of monitoring, $\hat{m} < m^*$, and in equilibrium

$$\hat{R} = \frac{1}{\hat{m}}$$

with an inefficient monitoring level.

This shows, as is intuitive, that a lower repayment $\hat{R}$ improves banks' monitoring incentives, thus justifying capital requirements.

Boot and Schmeits (2000) apply this framework to the analysis of conglomerates' monitoring decisions by considering a setup where two divisions can either operate independently or jointly as a conglomerate. Each division $i$ ($i = 1, 2$) invests in a project yielding $y$ in case of success (with probability $m_i$) and zero otherwise. Each division manager chooses $m_i$ with a monitoring cost $c(m_i)$.

In the perfect information case, with independent divisions, the efficient level of monitoring is reached.

In the conglomerate case, $y$ is assumed to be large, so the success of one division is enough to avoid the conglomerate's bankruptcy. The probability of debt repayment is

$$m_1(1 - m_2) + m_2(1 - m_1) + m_1 m_2 = m_1 + m_2 - m_1 m_2.$$

As a result, the condition that depositors get a zero expected return implies

$$R_C(m_1, m_2) = \frac{1}{m_1 + m_2 - m_1 m_2}.$$

Thus the conglomerate risk level is jointly generated by the two divisions, and the choice of monitoring level by each of them generates an externality on the other. In spite of perfect information, the conglomerate will not reach the efficient allocation. This comes from the fact that division $i$ chooses $m_i$ solution of

$$\max_{m_i} \; m_i(y - R_C(m_1, m_2)) - c(m_i) = [m_i y - c(m_i)] - m_i R_C(m_1, m_2).$$

Since the derivative of $m_i R_C(m_1, m_2)$ is positive, the level of monitoring chosen by each division is lower than the efficient one.

This effect is reversed in the imperfect information case, where $m_i$ is not known to investors and thus the repayments do not depend upon effort. The conglomerate will choose a higher level of monitoring than the stand-alone divisions. To see why, note that the first-order conditions will be $y - R_i = c'(m_i)$ for each stand-alone division and $y - R_C = c'(m_i)$ for each division that belongs in a conglomerate, where $R_i$ and $R_C$ are the equilibrium repayments given the expected levels of monitoring. Since the conglomerate is safer, we have $R_C < R_i$, and the conglomerate monitoring level will be closer to the efficient one.

To summarize, a division chooses a higher intensity of monitoring in a conglomerate if the investors' information is poor, and a lower intensity if their level of information is high. It therefore appears, as Boot and Schmeits point out, that the consequences of conglomeration involve two phenomena.

First, coinsurance between the two divisions in a conglomerate implies a lower cost of funds for both. In a perfect information setting, as in the Modigliani-Miller one, this has no effect, but in the presence of market imperfections, it will provide higher incentives to monitor because the profit margins are larger.

Second, the free riding of each division on the other decreases the conglomerate's incentives to monitor.

Notice that the level of information available to investors should not be taken as exogenous. Boot and Schmeits relate it to the intensity of market discipline. This raises interesting policy issues regarding disclosure and transparency.

### 3.5.3 Competition and Screening

Broecker (1990) models the consequences of competition between banks on the intensity with which they screen borrowers. He finds that the rejection decision of a bank creates an externality on the other banks when firms are able to sequentially apply for loans. The banks offering lower interest rates will be able to attract the best borrowers, leaving the low-quality borrowers to their competitors, thus generating multiple equilibria and interest rate differentials.

**Modeling Competition among Screening Banks**

Consider a risk-neutral, zero-interest-rate world where a continuum of firms applies for a loan in order to implement a project. The project yields a cash flow $y$ in case of success and zero in case of failure. The population of firms consists of good firms and bad firms. Good firms are in proportion $\psi$ and have a probability of success of $p_G$, and bad firms are in proportion $1 - \psi$ and have a lower probability of success, $p_B$. Assume that $p_G y > 1 > p_B y$, so that it is never profitable to lend to a bad firm.[26]

Assume there are only two banks in the industry. These banks screen the firms and decide whether to grant a loan. The decision to grant a loan is based on the updated probability of success of the project, given the creditworthiness of the population of loan applicants and the power of the screening test.

Note that the population of applicants faced by each of the banks depends upon the interest rates quoted by both banks. The bank offering a low (gross) repayment rate $R_L$ faces the overall population of borrowers. Denote by $\psi(\text{PASS})$ the updated probability of being a good firm that has passed the test. Then the average probability of repayment will be

$$p(1) = \psi(\text{PASS})p_G + (1 - \psi(\text{PASS}))p_B.$$

Suppose the other bank quotes $R_H > R_L$. Only the borrowers rejected from the low-interest-rate bank will apply for a loan at a higher interest rate. As a consequence, the high-interest-rate bank faces a population of firms with lower creditworthiness and lower average probability of repayment:

$$p(2) = \psi(\text{PASS}|\text{FAIL})p_G + (1 - \psi(\text{PASS}|\text{FAIL}))p_B,$$

where $\psi(\text{PASS}|\text{FAIL})$ is the probability of being a good firm conditional on having passed the second test and failed the first.

As is intuitive, the high-interest-rate bank inherits the worse credit risks:

$$\psi(\text{PASS}|\text{FAIL}) < \psi(\text{PASS}) \Rightarrow p(2) < p(1).$$

Broecker (1990) models competition in two ways, using one-stage games and two-stage games. In the one-stage game, the banks decide simultaneously whether they will provide credit and, if so, at which interest rate, without any possibility of revising their decision. In the two-stage game, banks announce first an interest rate but are not committed by their offers and are able to withdraw their credit offer in the second stage. Following Broecker, we show that in the one-stage game a pure strategy equilibrium never exists, whereas in the two-stage game a pure strategy equilibrium may exist.

**One-Stage Screening Game**

In the one-stage game, the banks are committed by the rates they have announced and have to grant credit to all borrowers that have passed the credit test, obtaining a good signal. It is easy to show that no pure strategy equilibrium exists.

Assume bank A sets a repayment rate $R_A$. Define $R(1) = 1/p(1)$ as the rate at which a monopolistic bank would break even. Notice that for $R_A < R(1)$, bank A will make losses, so we can restrict the equilibrium strategies to the interval $[R(1), y]$.

Consider next the strategies where both banks offer credit. Then, if $R_A > R(1)$, by slightly undercutting $R_A$, bank B obtains the expected probability of solvency $p(1)$ while A makes losses. This shows that $R_A$ is not an equilibrium strategy. Finally, if both banks choose $R(1)$, they will both make losses because their expected probability of repayment will be lower than $p(1)$. For example, for an equal splitting of firms between the two banks, it will be $\frac{1}{2}p(1) + \frac{1}{2}p(2)$. Thus there is no pure strategy equilibrium where both banks offer credit.

Proving that there is no equilibrium where only one bank offers credit is equally easy. If only one bank offered credit, it would make positive profits because there is no competition, and except for the very peculiar case where $R(1) = y$, there will be no equilibrium. Also, the case where no bank offers credit is not an equilibrium either, because each bank could then enter as a monopolist and make profits.

**Mixed-Strategy Equilibria**

Broecker shows that a mixed-strategy equilibrium exists, but that it cannot have a finite support because a strategy that would slightly undercut each interest rate in the support would be a profitable deviation.

**Two-Stage Screening Game**

In the two-stage game, the banks are able to walk out of the market without granting any loan. This allows for the existence of a pure strategy equilibrium. Assume that $p(2)y < 1$, so that it is never profitable to quote a high interest rate. Consider both banks offering interest rate $R(1)$ and then one of them walking out of the market. Both banks make zero profit. Deviating from this strategy by quoting a higher interest rate is not profitable. If no bank withdraws from the market, they both make losses. As a consequence, it is also optimal for one bank to withdraw.[27]

The explicit modeling of screening and its impact on credit allocation has interesting implications regarding the effect of the number of banks on the average probability of default. This effect is in stark contrast with the standard view, that a larger number of firms means more competition, which in turn benefits customers by decreasing prices and marginal costs. Here an increase in the number of banks may also be an additional chance for a bad firm to succeed in obtaining credit, thus increasing the banks' cost of funds and decreasing the overall efficiency of the banking industry.

To see this, consider the total output generated in a symmetric equilibrium for an economy with $N$ banks. Denote by $\psi(\text{FAIL}|B)$ and $\psi(\text{FAIL}|G)$, respectively, the equilibrium probability for a bad and a good firm to fail the creditworthiness test. The probability of a bad (resp. good) firm to be granted a loan after applying to $N$ banks will be

$$1 - [\psi(\text{FAIL}|B)]^N \qquad (\text{resp. } 1 - [\psi(\text{FAIL}|G)]^N).$$

Consequently, the total expected output for the economy will be

$$Y(N) = (1-\psi)(1-[\psi(\text{FAIL}|B)]^N)(yp_B - 1) + \psi[1 - \psi(\text{FAIL}|G]^N)(yp_G - 1).$$

To study the effect of an increase in $N$, compute the derivative

$$\frac{dY(N)}{dN} = -(1-\psi)\log(\psi(\text{FAIL}|B)[\psi(\text{FAIL}|B)]^N(yp_B - 1)$$
$$- \psi \log(\psi(\text{FAIL}|G)N[\psi(\text{FAIL}|G)]^N(yp_G - 1). \tag{3.33}$$

Since $yp_B < 1 < yp_G$, the first term is negative and the second is positive. The result will depend on the specific values for the parameters.

We now turn to the cases where the sign of the derivative is easy to compute.

First, consider the case where the test for a bad firm leads always to a fail result. Then increasing the number of banks is always beneficial because this increases the number of good firms that are granted credit. This is so because, in this case, $\psi(\text{FAIL}|B) = 1$ implies that the first term in (3.33) is zero, and the result obtains.

Second, in the general case, it seems natural to think that after a number of rounds almost all good firms will be granted credit, so that increasing the number of banks only serves the purpose of financing bad firms. This can be seen by noticing that

$$\left[\frac{\psi(\text{FAIL}|B)}{\psi(\text{FAIL}|G)}\right]^N$$

tends to infinity with $N$, thus implying that the first term will end up dominating. Consequently, for $N$ larger than some threshold $\hat{N}$, increasing the number of banks decreases the total output.

Finally, in the general case, the proportion of good firms, $\psi$, as well as the loss or gain $yp_K - 1$ $(k = G, B)$ will play a key role. For a low proportion of good firms, $\psi$, the first term will dominate even for a low number of banks $N$, and increasing the number of banks will be welfare-decreasing. The same effect will occur if the losses $yp_B - 1$ are high in comparison to the gains on good borrowers, $yp_G - 1$.

These results, initially pointed out by Broecker (1990), have been studied by Shaffer (1998), Dell'Ariccia (2000) in the context of a multiperiod model of spatial competition, and Marquez (2002) in the context of relationship banking. Shaffer provides an interesting empirical test of these effects. On the one hand, he tests for the effect of the number of (mature) banks on loan charge-off rates and finds a positive significant effect: a higher number of banks implies a higher credit risk.[28] On the other hand, he tests the model's implications on de novo banks that enter the credit market. Those banks will inherit all the borrowers that are rejected by the incumbents, so their charge-off rates should be higher. Shaffer's empirical results show that this is indeed the case and that it takes some ten years for a de novo bank to reach its competitors' charge-off rates.

To summarize, the explicit modeling of the screening process has two types of implications: first on the equilibrium itself because pure strategy equilibria are ruled out; second on the effects of entry. While in traditional pay-versus-delivery markets a larger number of firms increases efficiency, this need not be the case here. A larger number of banks increases the chances of a bad borrower's obtaining credit, which in turn could (but need not) decrease the efficiency of the resulting allocation.

## 3.6 Relationship Banking

Although the existence of long-term contractual relationships between banks and their customers has long been recognized (see, e.g., James's 1987 contribution),[29] the implication of relationship banking on the uniqueness of bank contracts has only been recently identified, providing a major breakthrough in banking theory and its empirical application.

The term *relationship banking* is not rigorously defined in the literature; we use it to refer to the investment in providing financial services that will allow *dealing repeatedly with the same customer* in a more efficient way.

For relationship banking to emerge, two conditions have to be met:

- The bank must be able to provide services to its client at more favorable conditions than its competitors, and this advantage improves with time.
- Contingent long-term contracts are not available.

In what follows these two conditions are assumed to be fulfilled.

The path-breaking contributions on relationship banking are those of Sharpe (1990) and Rajan (1992). Both acknowledge that if monitoring provides better information to the lending bank, this implies that in a multiperiod setting the bank has ex post a monopoly of information while its competitors are uninformed and therefore face a winner's curse,[30] leading to a holdup situation for the borrower.

Rajan's contribution explores, in addition, firms' choice between bank finance and market or arm's-length finance, showing that firms with better-quality projects will choose arm's-length finance. In this way, Rajan justifies the coexistence of banks and financial markets.

### 3.6.1 The Ex Post Monopoly of Information

Consider a two-period, risk-neutral, zero-interest-rate environment. Firms have investment opportunities that require one unit of investment at the beginning of each period and that are successful with a probability $p$ (in which case they yield $y$) and unsuccessful with the complementary probability (in which case they yield zero). Banks have some market power and provide the unit of funding that the firms require only if they obtain a gross expected return $\rho > 1$. $\rho = 1$ corresponds to the competitive situation. Before being able to provide a loan, banks have to incur once and for all a sunk cost $M$ that corresponds to monitoring activities. Once this investment is made, the bank is able to grant future loans to the same firm without any additional cost. As a consequence, maintaining a continued relationship with the same bank avoids duplication of monitoring costs. At the same time, it provides an ex post monopoly power to the incumbent bank.[31]

A firm that is funded by the same bank in both periods will have to repay $R_1$ on its first-period loan and $R_2$ on its second-period loan. The bank will invest $1 + M$ at date $t = 0$ and, if successful, an additional unit, so that the expected outlay is $1 + M + p$. Consequently, ex ante competition implies that all banks obtain the same return $\rho$ on their funds, so that

$$pR_1 + p^2 R_2 = \rho(1 + p + M). \tag{3.34}$$

If we assume competition at the interim stage, $t = 1$, this implies that if the firm switches to another bank, it will have to repay

$$\frac{\rho(1 + M)}{p}$$

at $t = 2$. Therefore,

$$R_2 = \frac{\rho(1 + M)}{p}$$

is also the optimal strategy for the incumbent bank. Regarding the equilibrium level of $R_1$, notice that the bank makes a second loan and obtains an additional profit $\rho M = (1 + M)\rho - \rho$ only if the project is successful, which happens with probability $p$. Competition at date $t = 0$ will drive $R_1$ down to the required expected return $\rho(1 + M)$:

$$pR_1 + p\rho M = \rho(1 + M),$$

and therefore,

$$R_1 = \frac{\rho(1 + M(1 - p))}{p} < R_2 = \frac{\rho(1 + M)}{p}.$$

This result shows that the bank uses its ex post monopoly power during the second period, while competition among banks drives down rates at the initial stage of the relationship. Thus, for instance, in the competitive case, $\rho = 1$, we would observe a profit from the ex post monopolistic situation in the last period and a loss in the first period, with the zero-profit condition holding over the two periods. Thus relationship banking leads to a holdup situation in the last period with an effect on the price structure that is similar to the one obtained in switching cost models.

Although, in general, the existence of a holdup situation generates a cost for the firm, it also implies in this context that banks will finance more risky ventures because they will ask for a lower first repayment at the initial stage of the project. This is important for small firms, as argued by Petersen and Rajan (1995). From that perspective, it is worth noting that the notion of market power has here two different interpretations, which lead to different implications. On the one hand, ex ante market power at the initial stage is reflected by $\rho$; on the other hand, capture power at the interim stage is reflected by $R_2$ and measures a bank's power to capture its borrowers. The difference is striking when we consider the effect of these two forms of market power on the time $t = 1$ marginal firm for which $y = R_1$:

- An increase in ex ante market power, $\rho$, implies an increase in the repayment $R_1$ and will deprive the marginal firm of credit.
- To illustrate the impact of capture power we have to relax the assumption of competition at time $t = 1$, so that through $R_2$ the bank appropriates a fraction of the firm's cash flow. The repayment $R_2$ has an upper bound $y$ but is then only limited by (3.34). An increase in $R_2$ clearly decreases $R_1$, and therefore the marginal firm will be granted credit when capture power is increased. Consequently, an increase in capture power will increase the number of firms the bank will finance. (An increase in $R_2$ with $\rho$ constant implies $pR_1 + (pR_2 - \rho) > \rho(1 + M)$, so it is worthwhile extending credit to riskier firms.)

The existence of an ex post information monopoly may lead to inefficiencies. It is therefore natural to raise the question of how banks can coordinate in order to share information about their clients.

Credit bureaus and public credit registers[32] provide an interesting instrument for this purpose. These are private or public organizations in charge of centralizing and distributing credit information to their members. The type of information they collect can be either negative (black) only or both negative and positive (white). In both cases, access to centralized information about borrowers reduces the extent of adverse selection, both on borrowers' types and on their indebtedness, and allows for more accurate pricing of debt.

The existence of an ex post monopoly of information implies an inefficiency in the allocation of resources. Even if the banking system is competitive and banks get zero profits over the two periods, first-period rates are subsidized while second-period rates for good borrowers are at a markup above the marginal cost. If borrowers have to switch from one bank to another for exogenous reasons, this inefficiency could be particularly harmful. In such a context, Pagano and Jappelli (1993) establish that exchanging information about borrowers may arise endogenously if borrowers can be forced to switch from one bank to another. This justifies the existence of credit bureaus, typically characterized by private voluntary disclosure of information on creditors or public credit registers, where information provision on credit is compulsory. Padilla and Pagano (1997) consider the case of moral hazard on behalf of the entrepreneur and show how credit bureaus, by restraining the banks' own future ability to extract rents, allow the entrepreneur to get a larger share of the surplus and therefore give him a greater incentive to invest in the project.

The justification of credit bureaus or public credit registers goes beyond the issue of the ex post monopoly of information. In fact, the creation of a credit bureau has both a screening effect and an incentive or disciplinary effect. By considering the different types of information exchanges in place around the world, Jappelli and Pagano (2000) show that the existence of credit bureaus is associated with lower credit risk, lower credit spreads (possibly because of better knowledge of the pool of applicants), and broader credit markets. Moreover, they are more likely to arise when creditors' rights are poorly enforced. This is consistent with the theoretical approach that predicts that the disciplinary effect will reduce the incentives to become overindebted. Whether efficient institutions will emerge spontaneously (credit bureaus) or will have to be imposed by the regulatory authorities (public credit registries), and whether it is more efficient to exchange only negative information regarding defaulting borrowers or both negative and positive information, are key issues in the literature and have been addressed in recent contributions (see Jappelli and Pagano 2000 for an excellent survey).

### 3.6.2 Equilibrium with Screening and Relationship Banking

The previous model of sunk cost investment in monitoring allowed us to understand the basic notion of the ex post monopoly of information and the resulting structure of loan interest rates. Still, it does not provide much help in understanding the effect of an improvement in information structure. This is the topic of Hauswald and Marquez (2003).

In their model, *inside*, or informed, banks that have established a relationship with a customer receive a signal on the customer's type, which (borrowing the notation from section 3.5) yields the probabilities $\psi(\text{FAIL}|B)$ and $\psi(\text{PASS}|G)$ for a bad firm to fail the test and for a good firm to pass it.

The *outside* bank does not have access to the screening technology and therefore has to rely on a coarser information structure. Still, the outside bank could access a public signal regarding a specific borrower that provides less informative probabilities $\psi_p(\text{FAIL}|B)$ and $\psi_p(\text{PASS}|G)$.

In this general setup, the ex post monopoly of information depends upon the information gap, that is, the difference between the information conveyed by private and public signals. The private signal depends upon the banks' ability to process information about their customers and is basically soft information; the public signal is obtained at arm's length and is based on hard information such as financial statements, the firm's size, age, and the like.

As a consequence, as argued by Hauswald and Marquez, when studying the effect of an improvement in the quality of information, it is crucial to distinguish between the effects of better public information (or better information dissemination) and the effects of better relationship-based information. Indeed, the effects are opposite.

Once the distinction is drawn, the result is quite intuitive (in spite of the complexity of mixed-strategy equilibrium). First, an increase in the quality of relationship-based information increases the ex post monopoly of information and therefore the expected interest rate. The effect is to reduce market competition.

Second, an increase in the quality of the public signal (information dissemination) will increase competition and decrease the expected interest rate. This is so because it reduces the ex post monopoly of information.

Another important extension of the ex post monopoly of information framework is developed in Hauswald and Marquez (2005), which assumes that a bank's ability to gather information decreases with its distance to the borrower. A bank acquires information for two reasons: "to soften competition and to extend its market share" (967). In this way, Hauswald and Marquez (2005) provide a framework for understanding the geographical aspects of credit markets.

### 3.6.3 Does Competition Threaten Relationship Banking?

It is natural to raise the issue of how competition affects relationship banking. This is clearly a relevant question in the present context of financial intermediation, where traditional banking is progressively replaced by fee-based financial services. This issue is related to the coexistence of financial markets and financial intermediaries (see chapter 2) and more generally to the comparison of market-based versus bank-based financial structures (e.g., U.S.-UK type versus German-Japanese type). How will relationship banking be affected by an increase of competition in the financial markets, and by an increase of competition in the banking industry? As of now, despite the relevance of this issue, the results are inconclusive.

Several contributions have explored this issue. Boot and Thakor (2000) argue that an increase in competition among banks results in more relationship lending, while an increase in competition in financial markets results in less relationship lending. Yafeh and Yosha (2003) obtain similar results in a different framework. Still, Gehrig (1998a; 1998b) and Dell'Ariccia (2000) obtain instead ambiguous results regarding the effects of competition on relationship banking. Also, Dell'Ariccia and Marquez

(2004) show that an increase in competition from outside lenders leads banks to reallocate credit toward more captured borrowers, what they refer to as a "flight to captivity."

### 3.6.4 Intertemporal Insurance

Since relationship banking is a priviledged link between a bank and a firm, it allows the firm to be partly protected from adverse business conditions. This idea, developed by Berlin and Mester (1999), allows relating relationship banking to intertemporal insurance.

The issue of intertemporal smoothing is developed by Allen and Gale (1997) as a justification for financial intermediation (see problem 2.9.6). Their argument is that intertemporal smoothing provides insurance for agents who would otherwise be confronted with random market conditions. Berlin and Mester (1999) develop a somewhat similar argument, focusing on firms rather than consumers and basing the demand for insurance on the existence of liquidity costs.

There are two main reasons why it might be efficient for a bank to provide insurance to its corporate borrowers. First, the bank portfolio is better diversified than the firms', and second, the risk has more impact on the firm than on the bank, which is better equipped to manage it. This may be because the bank has access to markets and financial contracts that the firm cannot use.

The link between insurance and relationship banking comes from the need for repeated lending in order to provide insurance. This is so because insurance premia cannot be used in this context; the firm's credit rating is not observable. Now, the bank will provide insurance in bad states of the world only if it knows it will benefit from the continuation of the relationship. This is why, since the bank cannot commit to lending, the insurance provision will be limited by the bank's option to invest in other assets. The opportunity cost of the bank's not providing insurance is the loss of the ex post monopoly rents. As a consequence, the larger the ex post monopoly rents, the higher the level of intertemporal insurance.

Berlin and Mester consider the bank liquidity insurance provision function in a competitive environment. By so doing, they predict that banks with more core deposits (whose remuneration is close to zero) are able to provide more insurance than those that are funded with liabilities yielding the market interest rate. Berlin and Mester's empirical analysis shows that this is indeed the case.

### 3.6.5 Empirical Tests of Relationship Banking

The theory of relationship banking has a number of implications that can be tested. It is based on the specific investments the two parties have to make to establish the relationship. Theory suggests that a bank will invest in relationship lending if it is able to appropriate a fraction of the benefits it generates. This is related to its capture power. On the other hand, a firm will only invest in relationship banking if it has limited access to direct finance.

Once the two parties have invested in the relationship, theory predicts that the firm will benefit from extended credit availability. In particular, it will be able to borrow at loan rates that differ from those that arm's-length borrowers are confronted with. Of course, the approach takes a simplified view of the world, since hybrid financial modes can be envisaged, mixing market finance and relationship banking (see Berger and Udell 2006).

We now discuss some empirical findings concerning the investment behavior of banks and firms, and the credit conditions faced by borrowing firms.

A straightforward way to test the importance of firms' investment in relationship banking is to measure the cost of destroying such a relationship. This methodology was used by Slovin, Sushka, and Polonchek (1993) to analyze the Continental Illinois distress. They showed that the severance of the relationship created a loss for the firms, because firms that were Continental Illinois creditors had their market value negatively affected by the event.[33]

Obtaining evidence on banks' investment in relationship banking is difficult; one needs a measure of capture power that is not simply a measure of market power.

Petersen and Rajan (1995) made a path-breaking, contribution when they showed that, as the market power of the bank increases, small, riskier businesses are able to obtain better credit conditions. Nevertheless, Elsas (2005) finds the opposite conclusion; he shows that the probability of a firm's having a single bank ("hausbank") is higher, the more competitive the market. So the mixed empirical evidence mirrors the theoretical models, with some linking relationship banking to monopoly power (Sharpe 1990; Rajan 1992) and others (Boot and Thakor 2000) predicting a higher intensity of relationship banking in a more competitive environment.

An indirect way to test for the existence of an implicit relationship banking contract is to analyze the characteristics of firms that rely on a single bank ("hausbank") and compare them with firms that borrow from multiple sources. Theory predicts that less transparent firms should prefer relationship banking, and the empirical results support this. Empirical evidence stresses three important determinants of firms involved in relationship banking: age, size, and type of business. The age and size of a firm are relevant because screening successful projects in younger, smaller firms is harder for the bank. The type of business is also important because the larger the fraction of intangibles in the firms' assets, the harder it is for them to obtain direct finance. Houston and James (2001) examined a sample of 250 publicly traded firms and found that "bank-dependent firms are smaller and less transparent than firms that borrow from multiple lenders" and that "only 3 percent of the firms with a single bank relationship have public debt outstanding." In their results, once a firm is large, old, transparent, and successful enough, it can tap the public debt market or engage in multiple bank relationships to avoid the holdup problem. In the same vein, Elsas (2005) analyzes a data set of German firms and shows that the probability of a firm's borrowing from a main bank is higher, the higher its risk (the lower its rating) and the lower its equity ratio.

Interestingly, this literature has shown that relationship banking is usually offered by smaller banks. As Berger and Udell (2006) summarize it, "Large institutions are found to lend to larger, older SMEs with stronger financial ratios, and small institutions are found to rely more on soft information and lend to SMEs with which they have stronger relationships" (9). The reason could be that large institutions have a comparative advantage in transaction lending, whereas small ones have a comparative advantage in relationship lending. This, in turn, could be explained by agency problems within the institutions related to the bank's organizational structure.

An alternative way of testing relationship banking is to study the credit conditions of firms that obtain their funds from a main bank. This can be done by analyzing either credit availability or credit spreads. The liquidity access approach was initially developed by Hoshi, Kashyap, and Scharfstein (1990) to analyze the Japanese banking industry. Using a sample of Japanese firms, they argued that relationship lending generates a surplus because firms that formed part of a conglomerate (Kereitsu) and engaged in a relationship with the bank of the Kereitsu had easier access to credit or faster recovery in periods of financial distress. This confirms the link between relationship lending and the provision of insurance, previously discussed. Subsequently, Elsas and Krahnen (1998) provided a similar result. But Houston and James (2001) showed that, on the contrary, "publicly traded firms that rely on a single bank are significantly more cash flow constrained than firms that maintain multiple bank relationships or have access to public debt markets."

Empirical studies have also considered the effect of duration and other proxies of relationship banking on the cost of credit for the borrowing firm. This may be complex, in particular if the repayment track record of borrowers is observable, so that the credit characteristics of firms become more easily observable after a number of loan renewals, as in Diamond (1991). The results are inconclusive and seem to differ depending on the country. Whereas in the United States loan prices decrease with duration (Berger and Udell 1995), in Europe they are unaffected (Elsas and Krahnen 1998) or increase (Degryse and Van Cayseele 2000; Degryse and Ongena 2005a). Of course, these studies differ in the way they measure the cost of credit, define the banking relationship, and treat collateral.

Empirical research has also examined the behavior of firms that have entered into a relationship with a main bank and, according to theory, face an information holdup situation.

Working with a data set of small Portuguese firms, Farinha and Santos (2000) found, "The likelihood of a firm substituting a single with multiple relationships increases with the duration of the single relationship.'' They implied that once a firm has proved to be successful for certain number of periods, it develops multiple banking links to avoid the holdup problem. Petersen and Rajan (1994) considered the relationship between age and bank borrowing and obtained a similar result: ''The fraction of bank borrowing declines from 63 percent for firms aged 10 to 19 years to 52 percent for the oldest firms in our sample. This seems to suggest that firms follow a ''pecking order'' for borrowing over time, starting with the closest sources (family) and then progressing to more arm's length sources.'' This evidence tends to indicate that the younger the firm, the larger the degree of asymmetric information, which leaves young firms with the unique option of a single bank relationship. Ongena and Smith (2001) corroborate this finding and establish, in addition, that small, young, and highly leveraged firms maintain shorter relationships.

Finally, and directly related to the empirical evidence on the effect of distance on credit conditions (see section 3.3.4), the literature on relationship banking has addressed the issue of distance by testing whether the effects of distance on loan rates could be the result of asymmetric information. Degryse and Ongena (2005b) and Agarwal and Hauswald (2007) conclude that distance effect on loan rates decreases (or becomes irrelevant) when some measure of relationship banking is introduced into the regression. Degryse and Ongena found the loan rate charged to relationship borrowers is unaffected by distance while the rate charged to arm's-length borrowers decreases with distance. In Agarwal and Hauswald's contribution, the firm-bank distance becomes statistically insignificant when the lending bank's internal rating (the bank's proprietary information) is introduced.

In a similar vein, Berger et al. (2004) find that small banks lend at lower distances because they have a comparative advantage in gathering soft information controlling for firm and market characteristics. Berger et al. show that the distance between a bank and its borrowing firm increases with the size of the bank.

## 3.7 Payment Cards and Two-Sided Markets

Banks provide noncash payment services to their customers in the form of checks, transfers, direct debit systems, and payment cards. The payment card industry has grown substantially in the last decades. Invented in the 1950s in the United States (Evans and Schmalensee 1999), payment cards became incredibly successful, but only after they solved a ''chicken-and-egg'' problem. A payment card is only valuable to customers if it is accepted by sufficiently many retailers, and retailers find it profitable to accept a card only if sufficiently many consumers hold it. The challenge for the payment card platform (which can be an independent entity like American Express or an association of banks like Visa or MasterCard)[34] is to attract the two sides of the market while not losing money overall. Thus the payment card industry is a natural example of a "two-sided industry" (Rochet and Tirole 2003; 2006; Armstrong 2006), where competing platforms provide interdependent services to two (or several) categories of users. Examples of two-sided industries are software (Parker and Van Alstyne 2005), Internet portals (Caillaud and Jullien 2003), media (Anderson and Coate 2005), and intermediaries (Ellison, Fudenberg, and Möbius 2004; Jullien 2005). This section focuses on the payment card industry and shows how the traditional concepts of industrial organization and competition economics (monopoly outcome, Bertrand competition, social welfare) must be adapted to the analysis of indirect externalities in two-sided markets.

### 3.7.1 A Model of the Payment Card Industry

This model was developed by Rochet and Tirole (2003) and extended by Wright (2003), Guthrie and Wright (2003), and Rochet and Tirole (2006). It considers an economy where consumers can pay by cards or by an alternative means of payment (say, cash).

We present here a simplified version of the Rochet and Tirole model where the card network is for-profit and contracts directly with retailers. Cards are issued by competing banks. This fits well the case of a proprietary network such as American Express.[35] Figure 3.3 shows the costs and benefits attached to a card transaction.

All these costs and benefits are measured with respect to a cash payment. For example, $b_B$ represents the cost saved by the buyer when he pays by card instead of cash. $b_B$ is a random variable, realized at the time of purchase and only observed by the buyer. For example, if the buyer has no cash in his pocket (and not enough time to find an ATM), $b_B$ may be high. The demand for card payments (measured by the fraction of purchases settled by card) is $D(p_B) = \Pr(b_B > p_B)$, where $p_B$ is the buyer fee.

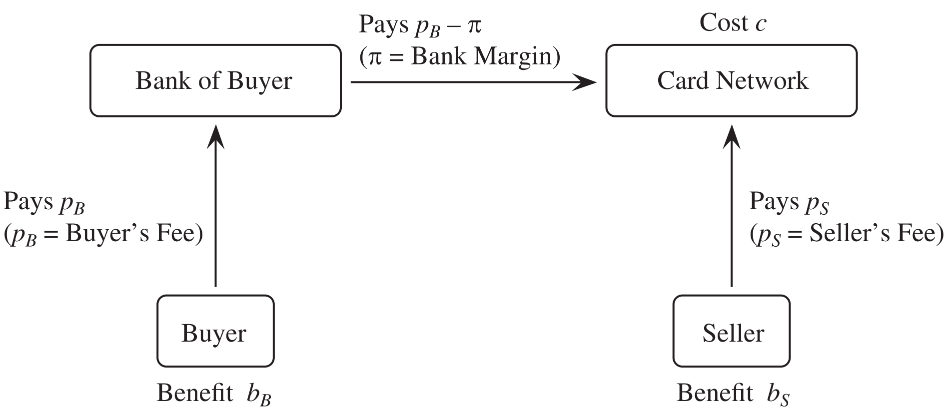



**Figure 3.3**
Costs and benefits of a card transaction.

Similarly, $b_S$ represents the cost saved by the retailer when the payment is made by card instead of cash. We assume that $b_S$ is the same for all transactions and observable by the card network. For simplicity, we neglect the cost incurred by the bank of the customer (the issuer of the card) when the payment is processed. The processing cost is entirely borne by the network. However, to compensate the bank of the buyer from incurring the cost of issuing the card, we assume the issuer collects a fixed margin $\pi$ on each payment. Assuming that the total number of transactions (card + cash) is fixed and normalized to 1, we can compute the expected welfare of the different protagonists:

Consumer surplus $\quad u - p + \int_{p_B}^{\infty} D(s)\,ds,$

where $u$ is the utility of consuming the good, $p$ is the retail price, and the integral represents the option value associated with the possibility of paying by card:

$$\int_{p_B}^{\infty} D(s)\,ds = E[\max(0, b_B - p_B)].$$

Bank profit $\quad \pi D(p_B).$

Network profit $\quad (p_B + p_S - c - \pi)D(p_B).$

Seller profit $\quad p - \gamma + (b_S - p_S)D(p_B),$

where $\gamma$ is the cost of the good for the seller.

### 3.7.2 Card Use

Card use depends on two things: whether retailers accept them, and how often customers want to use them. Retailers' decisions depend in turn on the price $p_S$ they face (the merchant service charge) and on their competitive environment. Rochet and Tirole (2002) study retailers' card acceptance in the Hotelling-Lerner-Salop model, and Wright (2003) considers monopoly and Cournot models. In all cases, there is a maximum value of $p_S$ above which merchants reject cards. We consider for simplicity the case of perfectly competitive retailers. In this case, retailers are segmented: some reject cards (and charge a lower retail price, equal to the marginal cost $\gamma$ of the good); the rest accept cards but charge a higher[36] price (the increment being equal to the expected net cost $(p_S - b_S)D(p_B)$ of card payment). Since consumers are ex ante identical, they will choose the store accepting cards if and only if their option value for card payments exceeds this incremental price:

$$\int_{p_B}^{\infty} D(s)\,ds \geq (p_S - b_S)D(p_B).$$

This condition can be reformulated as

$$\phi \equiv \int_{p_B}^{\infty} D(s)\,ds + (b_S - p_S)D(p_B) \geq 0, \tag{3.35}$$

where $\phi$ represents the total surplus of final users, namely, buyers and sellers.

Condition (3.35) characterizes card use when only one network operates (monopoly). When several identical cards compete (Bertrand competition), only the ones such that $\phi$ is maximum are effectively used. This fundamental result holds true also when retailers have some market power (Rochet and Tirole 2002; Wright 2003).

### 3.7.3 Monopoly Network

When there is only one (for-profit) network, it selects the two prices $p_B$ and $p_S$ that maximize its profit:

$$B = (p_B + p_S - c - \pi)D(p_B) \tag{3.36}$$

under the constraint that cards are used

$$\phi \equiv \int_{p_B}^{\infty} D(s)\,ds + (b_S - p_S)D(p_B) \geq 0. \tag{3.37}$$

The Lagrangian of this problem is

$$L = B + \lambda\phi,$$

and the first-order conditions are

$$\frac{\partial L}{\partial p_B} = D(p_B) + (p_B + p_S - c - \pi)D'(p_B) - \lambda D(p_B) + \lambda(b_S - p_S)D'(p_B) = 0,$$

$$\frac{\partial L}{\partial p_S} = D(p_B) - \lambda D(p_B) = 0.$$

The second condition is equivalent to $\lambda = 1$. Then the first condition gives

$$p_B = c + \pi - b_S.$$

$p_S$ is then determined by the constraint

$$\phi = 0 \Rightarrow p_S = b_S + \frac{\int_{p_B}^{\infty} D(s)\,ds}{D(p_B)}.$$

Thus $p_S > b_S$, which means that card payments increase the retailers' cost. This increase is passed on to consumers through an increase in retail prices. The social surplus generated by the card network is thus shared between the (monopoly) network and the banks.

### 3.7.4 Competing Payment Card Networks

If there are two (or more) networks that offer perfectly substitutable cards (Bertrand competition), only the ones that offer the maximum total user surplus $\phi$ will be used. The outcome of Bertrand competition is therefore characterized by prices $p_B$, $p_S$ that solve

$$\begin{cases} \max \phi = \int_{p_B}^{\infty} D(s)\,ds + (b_S - p_S)D(p_B) \\ \text{under } p_B + p_S \geq c + \pi, \end{cases}$$

Namely, that maximize total user surplus under the break-even constraint of the platform. Since this constraint is clearly binding, we can write

$$p_S = c + \pi - p_B,$$

and thus

$$\phi = \int_{p_B}^{\infty} D(s)\,ds + (b_S - c - \pi + p_B)D(p_B).$$

This is maximum when

$$p_B = c + \pi - b_S,$$

and thus

$$p_S = b_S.$$

### 3.7.5 Welfare Analysis

By adding the welfares of all protagonists, we obtain the expression of social welfare:

$$W = [u - \gamma] + \int_{p_B}^{\infty} D(s)\,ds + (p_B + b_S - c)D(p_B).$$

Note that $W$ does not depend on $p_S$. It is maximum for $p_B = p_B^W \equiv c - b_S$. Comparing with the formulas obtained for the monopoly and competitive cases, we see that

- both in the monopoly and competitive cases, buyer prices are higher than the welfare-maximizing level $p_B^W$;
- network competition leads to lower seller prices (and indirectly to lower retail prices) but does not change card use (and thus social welfare).

Thus in a two-sided industry, perfect competition between platforms does not necessarily lead to a social optimum. This is even more striking in the case where platforms are not-for-profit associations of banks (Rochet and Tirole 2002). In this case, the networks do not make any profit:

$$p_B + p_S = c + \pi,$$

and $p_B$ is chosen to maximize the profit of banks $\pi D(p_B)$ under the constraint that $\phi \geq 0$ (so that cards are used effectively).

In the monopoly case, $p_B$ is then the minimum buyer price such that

$$\phi = \int_{p_B}^{\infty} D(s) + (b_S - c - \pi - p_B)D(p_B) \geq 0,$$

which is lower than the optimal price $p_B^W$ when

$$\int_{p_B^W}^{\infty} D(s)\, ds - \pi D(p_B^W) \geq 0,$$

that is, when the profit margin $\pi$ of issuers is not too large.

The competitive case gives the same outcome as when networks are for-profit: Bertrand competition automatically reduces their profit to zero.

Therefore, not-for-profit associations tend to choose too low buyer prices when they have market power (monopoly) and too high buyer prices when they are in competition.

## 3.8 Problems

### 3.8.1 Extension of the Monti-Klein Model to the Case of Risky Loans

This problem is adapted from Dermine (1986). Modify the model of section 3.2 by allowing borrowers to default. More specifically, suppose that the bank has lent $L$ to a firm that has invested it in a risky technology with a net (unit) return $\tilde{y}$. In the absence of collateral, the net (unit) return to the bank will be $\min(r_L, \tilde{y})$. When $\tilde{y} < r_L$, the firm defaults, and the bank seizes the firm's assets, which are worth $(1 + \tilde{y})L$.

1. Assuming that the bank has no equity (in conformity with the model of section 3.2), show that the bank itself will default if $\tilde{y}$ is below some threshold $y^*$. Compute $y^*$.

2. Assume risk neutrality and limited liability of the bank. The bank chooses the volumes $L^*$ of loans and $D^*$ of deposits that maximize the expectation of the positive

part of its profit. (If this profit is negative, the bank defaults.) Write the first-order conditions that characterize $L^*$ and $D^*$.

3. Show that $D^*$ is characterized by the same condition as in the Monti-Klein model.

4. On the contrary, show that $L^*$ depends in general on what happens on the deposit side, so that the separability property is lost.

### 3.8.2 Compatibility between Banking Networks

This problem is adapted from Matutes and Padilla (1994). Consider a circular banking industry as described in section 3.3.1 with three symmetrically located banks, A, B, and C, each having an ATM. Competition is modeled by a two-stage game in which banks first decide whether their ATMs will be compatible and then compete in deposit rates. There are no costs of compatibility, and fixed costs and management costs are ignored. Let $C_A$ denote the set of banks such that their ATMs are compatible with that of A (by convention $A \in C_A$).

For a depositor located at a distance $x_j$ from bank $j$, the utility obtained by depositing its unit of cash in bank A is

$$r_A + k|C_A| - tx_A,$$

where $r_A$ is A's offered deposit rate, $|C_A|$ is the number of elements of $C_A$, $t$ is the transportation cost parameter for account management (which is necessarily to be done at bank A), and $kn$ represents the benefits derived from the use of a network of $n$ ATMs.

1. Consider first the two symmetric cases in which the ATMs are incompatible or fully compatible, and confirm that each bank's profit is equal at equilibrium:

$$\Pi = \frac{D}{3^2} t.$$

Explain this result.

2. Assume now that A and B have compatible ATMs, whereas C's remains incompatible. Compute the equilibrium profits of the three banks.

3. What do you conclude?

### 3.8.3 Double Bertrand Competition

This problem is adapted from Stahl (1988) and Yanelle (1989). Consider a banking industry characterized by a downward-sloping demand for loans $L(r_L)$ and an upward-sloping supply of deposits $D(r_D)$. Two banks compete à la Bertrand, by quoting interest rates $r_L^i$ and $r_D^i$ $(i = 1, 2)$.

1. If the quotes are simultaneous on the two markets, show that the unique equilibrium is characterized by

$$r_D^i \equiv r_L^i = r^w,$$

where $r^w$ is the Walrasian interest rate, characterized by

$$L(r^w) = D(r^w)$$

(equality of supply and demand).

We suppose from now on that banks compete sequentially. In a first stage they compete for deposits. We denote by $(D_1, D_2)$ the volumes of deposits collected by the two banks. In a second stage they compete for loans. We assume that deposits are their only source of funds, so that bank $i$ cannot lend more than $D_i$. We denote by $\hat{r}$ the interest rate that maximizes $(1 + r)L(r)$.

2. Show that when $\hat{r} > r^w$ the Walrasian equilibrium is not an equilibrium of the two-stage game. (Indication: If bank 2 offers $r_D^2 = r^w$ at the first stage, bank 1 can quote a slightly higher rate $r_D^1 = r^w + \varepsilon$, collect all deposits, and charge $r_L^1 = \hat{r}$ at the second stage.)

### 3.8.4 Deposit Rate Regulation

We use the Monti-Klein model to study the effect of a ceiling $\bar{r}_D$ on deposit rates. Specifically, we want to examine whether such a regulation could decrease interest rates on loans.

Let the profit function of the bank be

$$\pi(r_L, r_D) = (r_L - r)L(r_L) + (r - r_D)D(r_D) - C(D(r_D), L(r_L)).$$

Assume regulation is binding.

1. Computing the first-order conditions and using comparative statics analysis, show that a necessary and sufficient condition for a ceiling on deposit rates to induce a decrease in lending rates is

$$\frac{\partial^2 \pi}{\partial r_L \partial r_D} > 0.$$

2. Show that this condition is equivalent to

$$\frac{\partial^2 C}{\partial D \partial L} > 0.$$

What do you conclude?

## 3.9 Solutions

### 3.9.1 Extension of the Monti-Klein Model to the Case of Risky Loans

1. The profit of the bank becomes

$$\tilde{\pi}(L, D, \tilde{y}) = [\min(r_L(L), \tilde{y}) - r]L + [r - r_D(D)]D.$$

The threshold $y^*$ corresponds to the value of $\tilde{y}$ such that $\tilde{\pi}$ vanishes ($L$ and $D$ being given):

$$y^* = r - [r - r_D(D)]\frac{D}{L}.$$

2. The objective function of the bank is

$$\pi(L, D) = E[\max(0, \tilde{\pi}(L, D, \tilde{y}))].$$

The first-order conditions are

$$\frac{\partial \pi}{\partial D} = E\left[\frac{\partial \tilde{\pi}}{\partial D}(L, D, \tilde{y}) \mathbb{I}_{\tilde{y} > y^*}\right] = 0,$$

$$\frac{\partial \pi}{\partial L} = E\left[\frac{\partial \tilde{\pi}}{\partial L}(L, D, \tilde{y}) \mathbb{I}_{\tilde{y} > y^*}\right] = 0,$$

where $\mathbb{I}_A$ denotes the indicator function of the set A, with

$$\frac{\partial \tilde{\pi}}{\partial D}(L, D, \tilde{y}) = r - r_D - Dr'_D(D)$$

and

$$\frac{\partial \tilde{\pi}}{\partial L} = \begin{cases} r_L - r + Lr'_L(L) & \text{if } \tilde{y} > r_L, \\ \tilde{y} - r & \text{if } \tilde{y} < r_L. \end{cases}$$

3. The expression of

$$\frac{\partial \tilde{\pi}}{\partial D}$$

is independent of $\tilde{y}$. Therefore, $D^*$ is characterized as before by the condition

$$r - r_D(D^*) = D^* r'_D(D^*),$$

which does not depend on what happens on the asset side.

4. The determination of $L^*$ is more complex:

$$0 = \frac{\partial \pi}{\partial L} = (r_L - r + Lr'_L) \quad \text{Proba}(\tilde{y} > r_L) + E[(\tilde{y} - r)\mathbb{I}_{y^* < \tilde{y} < r_L}].$$

The second term, if it does not vanish, introduces a relation between deposits and loans, since $y^*$ depends on $D$.

### 3.9.2 Compatibility between Banking Networks

1. The incompatibility case is exactly the one studied in section 3.3.1 with $n = 3$. In the case of full compatibility, $3k$ is added to all utilities, which does not change the formulas defining marginal depositors. The profit functions are identical, and so is the equilibrium.

2. On the B side of A, the marginal depositor for A is indifferent between depositing in A or in B if he is at a distance $x$ of A (and $\frac{1}{3} - x$ of B) such that

$$r_A + 2k - tx = r_B + 2k - t\left(\frac{1}{3} - x\right).$$

This condition simplifies into

$$x = \frac{r_A - r_B + t/3}{2t}.$$

On the C side of A, the marginal depositor is defined similarly by a distance $y$ such that

$$r_A + 2k - ty = r_C + k - t\left(\frac{1}{3} - y\right),$$

yielding

$$y = \frac{r_A - r_C + k + t/3}{2t}.$$

Therefore, A's supply of deposits equals $D(x + y)$, and its profit is $D(r - r_A)(x + y)$.

The marginal depositor for bank C will be obtained similarly as

$$x' = \frac{(r_C - r_A - k + t/3)}{2t}, \qquad y' = \frac{(r_C - r_B - k + t/3)}{2t},$$

and its profit equals $\Pi_C = D(r - r_C)(x' + y')$.

First-order conditions for A and C yield, respectively,

$$-\left[\frac{r_A - r_B}{2t} + \frac{r_A - r_C + k}{2t} + \frac{1}{3}\right] + \frac{(r - r_A)}{t} = 0$$

and

$$-\left[\frac{2r_C-(r_A+r_B)-2k}{2t}+\frac{1}{3}\right]+\frac{(r-r_C)}{t}=0.$$

By symmetry, $r_A = r_B$. Thus we obtain a system of two equations with two unknowns:

$$3r_A - r_C = 2r - k - \frac{2t}{3},$$

$$-r_A + 2r_C = r + k - \frac{t}{3}.$$

The solution is

$$r_A = r_B = r - \frac{k}{5} - \frac{t}{3}; \qquad r_C = r + \frac{2k}{5} - \frac{t}{3}.$$

The profit levels are

$$\pi_A = \pi_B = Dt\left(\frac{1}{3}+\frac{k}{5t}\right)^2, \qquad \pi_C = Dt\left(\frac{1}{3}-\frac{2k}{5t}\right)^2.$$

3. Banks A and B get a higher profit by organizing compatibility between their ATMs but excluding bank C. The outcome of free competition is thus partial compatibility, whereas full compatibility is more efficient.

### 3.9.3 Double Bertrand Competition

1. The proof uses the traditional undercutting argument of Bertrand. If bank 2 takes a positive margin $(r_L^2 > r_D^2)$, bank 1 can undercut it by quoting $(r_L^1, r_D^1)$ such that $r_L^2 > r_L^1 > r_D^1 > r_D^2$. The only possible equilibrium involves $r_D^i = r_L^i \equiv r$. But if $r \neq r^w$, one side is rationed (supply or demand), and banks lose money. Now, if $r = r^w$, banks make a zero profit but cannot deviate without losing money.

2. Suppose banks quote $r_D^i \equiv r^w$ in the first stage, and $r_L^i \equiv r^w$ in the second. They make zero profit. Suppose bank 1 deviates by quoting $r_D^1 = r^w + \varepsilon$, where $\varepsilon > 0$ is small. All the deposits go to bank 1, which becomes a monopoly in the second-stage game. Since $\hat{r} > r^w$, bank 1 maximizes its profit by quoting $r_L^1 = \hat{r}$ and serving the loan demand $L(\hat{r}) < L(r^w) < D(r^w + \varepsilon)$. For $\varepsilon$ small, the profit is positive since

$$(1+\hat{r})L(\hat{r}) > (1+r^w)L(r^w) = (1+r^w)D(r^w).$$

Thus the Walrasian allocation cannot be an equilibrium of the two-stage Bertrand game.

### 3.9.4 Deposit Rate Regulation

The maximization program of the bank becomes

$$\begin{cases} \max \pi(r_D, r_L) \\ r_D \leq \bar{r}_D. \end{cases}$$

1. If the constraint is binding, the first-order condition is simply

$$\frac{\partial \pi}{\partial r_L}(\bar{r}_D, r_L) = 0,$$

and this yields a solution $\hat{r}_L(\bar{r}_D)$.

The effect of changing $\bar{r}_D$ is given by

$$\frac{\partial^2 \pi}{\partial r_L^2} \frac{d\hat{r}_L}{d\bar{r}_D} + \frac{\partial^2 \pi}{\partial r_L \partial r_D} = 0.$$

Therefore, since

$$\frac{\partial^2 \pi}{\partial r_L^2} < 0,$$

the condition for

$$\frac{d\hat{r}_L}{d\bar{r}_D} > 0$$

is that

$$\frac{\partial^2 \pi}{\partial r_L \partial r_D} > 0.$$

2. Differentiating twice the expression giving $\pi(r_L, r_D)$, we obtain

$$\frac{\partial^2 \pi}{\partial r_L \partial r_D} = -\frac{\partial^2 C}{\partial D \partial L} D'(r_D) L'(r_L).$$

Since $D' > 0$ and $L' < 0$, the condition

$$\frac{\partial^2 \pi}{\partial r_L \partial r_D} > 0$$

is equivalent to

$$\frac{\partial^2 C}{\partial L \partial D} > 0.$$

We have seen that the converse property,

$$\frac{\partial^2 C}{\partial L \partial D} < 0,$$

is more reasonable because it implies scope economies between deposits and loans. But in this case a ceiling on deposit rates provokes an increase in loan rates, the opposite of what was intended.

## Notes

1. Indeed, granting a loan is like buying a security issued by the borrower. Similarly, collecting deposits is like issuing securities. However, the discussion here conforms to the more traditional view of a bank's buying funds from depositors and selling them to borrowers. It therefore speaks of a *demand* for loans by borrowers and a *supply* of deposits by households.

2. Sealey and Lindley (1977) were among the first to use the microeconomic theory of the firm to build a rigorous model of banks' production functions. In their approach, banks can be described as multiunit firms that use labor and physical capital as inputs for producing different financial services for depositors and borrowers. The main specificity of banks (or more generally, depository financial institutions) with respect to industrial firms is that their outputs (namely, these financial services) can be measured only indirectly, through the volumes of deposits $D$ and loans $L$ they generate. The apparent cost function $C(D, L)$ of the bank is obtained by finding the efficient combination of inputs that generates a given vector $(D, L)$ (see also section 3.8).

3. The monetary base is the sum of currency in circulation plus reserves held by the banks at the Central Bank.

4. The symbol $\Delta$ refers to increments in stock variables.

5. In this simple model, an issue of Treasury bills is equivalent to a decrease in the monetary base.

6. As usual with constant returns to scale, equilibrium is possible only when profit margins are zero.

7. The comparison with the standard approach of the credit multiplier is complicated by the fact that the monetary base $M_0$ is no longer considered as a policy instrument. The focus is now on open market operations (changes in $B$) that were equivalent to changes in $M_0$ of the same amount and opposite sign in the simple model put forth in section 3.1.2.

8. The minus sign is only there to ensure that the elasticity $\varepsilon_L$ is positive, which is the more usual and more convenient convention.

9. As usual, we assume that $\varepsilon_L$ is greater than 1; otherwise the bank's problem may not have a solution.

10. The alternative model of price competition à la Bertrand is examined in problem 3.8.3.

11. This assumption is not important; any increasing convex cost function would lead to similar results.

12. This location model can also be interpreted in a more abstract fashion. Suppose depositors have different preferences about the mix of services to be provided by their bank. Each depositor prefers a specific combination of banking services. Transportation costs then correspond to the utility loss associated with consuming the mix of services offered by a bank instead of one's preferred combination.

13. The assumption of sequential entry, although more natural, would enormously complicate the analysis because banks do not always locate uniformly.

14. In this simplistic model, each bank has one branch. However, the example can be easily extended to a case in which banks have several branches. If these branches are not adjacent, the equilibrium deposit rates are unchanged, and the results still hold.

15. It is nothing but the adaptation to the banking context of the "proliferation" result of Salop (1979), who had in mind a different context: industries with differentiated products such as breakfast cereal brands.

16. This point is developed in problem 3.8.4.

17. It may seem strange that a consumer simultaneously borrows (at a high rate) and lends (at a low rate), instead of netting out the position. However, this is very common in practice. The reader is invited to determine why this is the case.

18. Chiappori, Perez-Castrillo, and Verdier (1995) also study the effect of tied-up contracts on the efficiency of monetary policy.

19. In fact, tied-up contracts allow banks to bypass the regulation. The total subsidy to each borrower-depositor equals the forgone interest on this deposit.

20. Bouckaert and Degryse (1995) have also used the Salop model for modeling "phonebanking" (an option that some banks offer their customers to deal with them by phone, which obviously reduces transaction costs). Similarly, Degryse (1996) studies the impact on banking competition of the possibility of offering remote access to banking services.

21. Allen and Gale (2004) point out that the trade-off between competition and financial stability may have been overemphasized. They argue, first, that the competitive outcome is still constrained-efficient even in an incomplete contracts framework, and second, that dynamic considerations may lead to different conclusions: inefficient banks are likely to disappear in the long run, either by failure or by being acquired and restructured by another bank.

22. This is a particular form of the Modigliani-Miller theorem.

23. Still, Hellmann et al. (2000) argue that if capital is very costly, this may in fact reduce the charter value, a point that we develop in chapter 9.

24. However, it is not always the case that increased competition decreases financial stability; see, for example, Allen and Gale (2004).

25. Notice that this framework can be viewed as an extension of Holmström and Tirole (1997).

26. We also make the classical tie-breaking assumption that firms that are indifferent between borrowing or not, choose to apply for a loan.

27. Notice that there is another, less attractive equilibrium where both banks withdraw.

28. Of course, this does not mean that banks' profitability is lower in regions with a higher number of banks. In fact, the opposite is true, and a larger number of banks is associated with a higher growth rate (Shaffer 1998, 382), a result also found in King and Levine (1993) and Rajan and Zingales (1998) at the country level.

29. James (1987) examined the effect of a bank loan and a bond issue on the value of the firm raising funds equity and found the interesting result that receiving a bank loan increased significantly the price of a firm but issuing a bond did not. Since the firm leverage is equally affected by both, the main element responsible for the difference was the monitoring of a bank. Banks had better information than the market, and when a firm was granted a loan, the stock market reacted accordingly by updating the information on the firm's prospect and setting a higher price for the firm's stock. As a refinement of James's findings, Lummer and McConnell (1989) find that it is only loan renewals that affect the firms equity prices, not the initial loan a firm is granted. This would support the idea that monitoring is perfected through time, and therefore that the relationship between the banker and its client implies a progressive transmission of privileged information to the former.

30. As in Rajan (1992), the exact equilibrium of the game, in Sharpe's model, obtained by Von Thadden (2004), is characterized by completely mixed strategies, which is not surprising given Broecker's (1990) results. Consequently, this setup makes it particularly difficult to explore relationship banking. The discussion instead assumes reusability of the banks' information, stating that banks initially pay a firm-specific sunk cost of monitoring and that once they have paid it, the updating cost is much lower. Therefore, the monitoring cost is lower for the bank that already has a lending relationship with the firm.

31. This is akin to assuming the existence of a switching cost. The difference is that the "switching cost" is paid by the monitoring bank.

32. Credit bureaus receive information voluntarily supplied by the banks and under the principle of reciprocity. This allows each bank to have access to all the information collected. Public credit registers are managed by Central Banks, which make compulsory the transmission of information.

33. This type of analysis has been successfully extended to study the effect of the banking relationship on the terms of the borrowing firms' IPOs.

34. MasterCard and Visa have decided to become for-profit and go public.

35. See Rochet and Tirole (2003) for a more complete model in which the card network is run by an association of issuers (consumers' banks) and acquirers (retailers' banks).

36. We assume $p_S \geq b_S$, which holds true in all the cases considered here.

## References


Agarwal, S., and R. Hauswald. 2007. Distance the information asymmetries in lending. Mimeo, American University, Washington, D.C.

Armstrong, M. 2006. Competition in two-sided markets. *Rand Journal of Economics* 37 (3): 668–691.

Allen, F., and D. Gale. 1997. Financial markets, intermediaries, and intertemporal smoothing. *Journal of Political Economy* 105 (3): 523–546.

———. 2000. *Comparing financial systems*. Cambridge, Mass.: MIT Press.

———. 2004. Competition and financial stability. *Journal of Money, Credit and Banking* 36 (3): 453–480.

Anderson, S., and S. Coate. 2005. Market provision of broadcasting: A welfare analysis. *Review of Economic Studies* 72 (4): 947–972.

Baumol, W. J., J. C. Panzar, and R. P. Willig. 1982. *Contestable markets and the theory of industrial structure*. San Diego: Harcourt Brace Jovanovich.

Beck, T., A. Demirgüç-Kunt, and R. E. Levine. 2003. Bank concentration and crises. University of Minnesota, Carlson School of Management Working Paper; and World Bank Policy Research Paper No. 3041.

Bell, F. W., and N. B. Murphy. 1968. Economies of scale and division of labor in commercial banking. *National Banking Review* 5: 131–139.

Benston, G. J. 1965a. Branch banking and economies of scale. *Journal of Finance* 20: 312–331.

———. 1965b. Economies of scale and marginal costs in banking operations. *National Banking Review* 2: 507–549.

Benston, G. J., G. A. Hanweck, and D. Humphrey. 1982. Scale economies in banking. *Journal of Money, Credit and Banking* 14 (1): 435–546.

Berger, A. N. 1999. The "big picture" of relationship finance. In *Business access to capital and credit*, ed. J. L. Blanton, A. Williams, and S. L. Rhine, 390–400. A Federal Reserve System research conference.

Berger, A. N., A. Demirguc-Kunt, R. Levine, and J. G. Haubrich. 2004. Bank concentration and competition: An evolution in the making. *Journal of Money, Credit and Banking* 36 (3): 433–451.

Berger, A. N., and R. DeYoung. 1997. Problem loans and cost efficiency in commercial banks. *Journal of Banking and Finance* 21 (6): 849–870.

Berger, A. N., G. A. Hanweck, and D. B. Humphrey. 1987. Competitive viability in banking: Scale, scope, and product mix economies. *Journal of Monetary Economics* 20 (3): 501–520.

Berger, A. N., and L. Mester. 1997. Inside the black box: What explains differences in the efficiency of financial institutions. *Journal of Banking and Finance* 21 (7): 895–947.

Berger, A. N., and G. Udell. 1992. Some evidence on the empirical significance of credit rationing. *Journal of Political Economy* 100 (5): 1047–1077.

———. 1995. Relationship lending and lines of credit in small firm finance. *Journal of Business* 68 (3): 351–381.

———. 2006. A more complete conceptual framework for SME finance. *Journal of Banking and Finance* 30 (11): 2945–2966.

Berlin, M., and L. Mester. 1999. Deposits and relationship lending. *Review of Financial Studies* 12 (3): 579–607.

Boot, A.W.A. 2000. Relationship banking: What do we know? *Journal of Financial Intermediation* 9 (1): 7–25.

Boot, A.W.A., and A. Schmeits. 2000. Market discipline and incentive problems in conglomerate firms with applications to banking. *Journal of Financial Intermediation* 9 (3): 240–273.

Boot, A.W.A., and A. V. Thakor. 1997. Banking scope and financial innovation. *Review of Financial Studies* 10 (4): 1099–1131.

———. 2000. Can relationship banking survive competition? *Journal of Finance* 55 (2): 679–713.

Bouckaert, J., and H. Degryse. 1995. Phonebanking. *European Economic Review* 39 (2): 229–244.

Boyd, J., and G. De Nicoló. 2005. The theory of bank risk-taking and competition revisited. *Journal of Finance* 60 (3): 1329–1343.

Broecker, T. 1990. Credit worthiness tests and interbank competition. *Econometrica* 58: 429–452.

Caillaud, B., and B. Jullien. 2003. Chicken and egg: Competition among intermediation service providers. *Rand Journal of Economics* 34 (2): 309–328.

Caminal, R., and C. Matutes. 2002. Market power and banking failures. *International Journal of Industrial Organization* 20 (9): 1341–1362.

Carletti, E., and P. Hartmann. 2002. Competition and stability: What's special about banking?. Special paper 140, Financial Markets Group, London School of Economics.

Cecchetti, S. 1999. Legal structure, financial structure and the monetary transmission mechanism. NBER working paper 7151.

Cerasi, V. 1995. A model of retail banking competition. London School of Economics.

Cetorelli, N., and P. Strahan. 2006. Finance as a barrier to entry: Bank competition and industry structure in local U.S. markets. *Journal of Finance* 61 (1): 437–461.

Chamberlin, E. 1933. *Theory of monopolistic competition*. Cambridge, Mass.: Harvard University Press.

Chiappori, P. A., D. Perez-Castrillo, and F. Verdier. 1995. Spatial competition in the banking system, localization, cross-subsidies and the regulation of interest rates. *European Economic Review* 39 (5): 889–919.

Clark, J. A. 1988. Economies of scale and scope at depository financial institutions: A review of the literature. Federal Reserve Bank of Kansas City. *Economic Review*, September, 16–33.

Cordella, T., and E. L. Yeyati. 2002. Financial opening, deposit insurance and risk in a model of banking competition. *European Economic Review* 46 (3): 471–485.

Degryse, H. 1996. On the interaction between vertical and horizontal product differentiation: An application to banking. *Journal of Industrial Economics* 44 (2): 169–186.

Degryse, H., and S. Ongena. 2005a. Competition and regulation in the banking sector: A review of the empirical evidence on the sources of bank rents. Mimeograph.

———. 2005b. Distance, lending relationship and competition. *Journal of Finance* 60 (1): 231–266.

Degryse, H., and P. Van Cayseele. 2000. Relationship lending within a bank-based system: Evidence from European small business data. *Journal of Financial Intermediation* 9 (1): 90–109.

Dell'Ariccia, G. 2000. Learning by lending, competition and screening incentives in the banking industry. Working paper 00-10, Wharton School, University of Pennsylvania, Philadelphia.

Dell'Ariccia, G., and R. Marquez. 2004. Information and bank credit allocation. *Journal of Financial Economics* 72 (1): 185–214.

Dermine, J. 1986. Deposit rates, credit rates and bank capital: The Monti-Klein model revisited. *Journal of Banking and Finance* 10: 99–114.

Diamond, D. W. 1991. Monitoring and reputation: The choice between bank loans and directly placed debt. *Journal of Political Economy* 99: 689–721.

Edwards, F. R. 1964. Concentration in banking and its effects on business loan rates. *Review of Economics and Statistics* 46: 294–300.

Edwards, F. R., and F. Mishkin. 1995. The decline of traditional banking: Implications for financial stability and regulatory policy. Federal Reserve Bank of New York. *Economic Policy Review* 1: 27–45.

Ellison, G., D. Fudenberg, and M. Möbius. 2004. Competing auctions. *Journal of the European Economic Association* 2 (1): 30–66.

Elsas, R. 2005. Empirical determinants of relationship banking. *Journal of Financial Intermediation* 14 (1): 32–57.

Elsas, R., and J. Krahnen. 1998. Is relationship lending special? Evidence from credit-file data in Germany. Mimeograph.

Evans, D., and R. Schmalensee. 1999. *Paying with plastic*. Cambridge, Mass.: MIT Press.

Farinha, L. A., and J.A.C. Santos. 2000. Switching from single to multiple bank lending relationships: Determinants and implications. Working paper 83, Bank for International Settlements. DOI 10.2139/ssm.208268.

Flannery, M. 1996. Financial crises, payment system problems, and discount window lending. *Journal of Money, Credit and Banking* 28 (4): 804–824.

Freixas, X. 2005. Deconstructing relationship banking. *Investigaciones Economicas* 29 (1): 3–31.

Gale, D. 1993. Branch banking, unitary banking and competition. Department of Economics, Boston University. Mimeograph.

Gehrig, T. 1998a. Competing markets. *European Economic Review* 42 (2): 277–310.

Gehrig, T. 1998b. Screening, cross-border banking, and the allocation of credit. *Research in Economics* 52 (4): 387–407.

Gilligan, T. W., and M. L. Smirlock. 1984. An empirical study of joint production and scale economies in commercial banking. *Journal of Banking and Finance* 8 (1): 67–77.

Gilligan, T. W., M. L. Smirlock, and W. Marshall. 1984. Scale and scope economies in the multi-product banking firm. *Journal of Monetary Economics* 13 (3): 393–405.

Greenbaum, S. I., G. Kanatas, and I. Venezia. 1989. Equilibrium loan pricing under the bank-client relationship. *Journal of Banking and Finance* 13: 221–235.

Guthrie, G., and J. Wright. 2003. Competing payments schemes. Mimeo, University of Auckland.

Hancock, D. 1991. *A theory of production for the financial firm*. Norwell, Mass: Kluwer.

Hannah, T., and A. Berger. 1991. The rigidity of prices: Evidence from the banking industry. *American Economic Review* 81: 938–945.

Hauswald, R., and R. Marquez. 2003. Information technology and financial services competition. *Review of Financial Studies* 16 (3): 921–948.

———. 2005. Competition and strategic information acquisition in credit markets. *Review of Finance Studies* 19: 967–1000.

Hellmann, T., K. Murdock, and J. Stiglitz. 2000. Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough? *American Economic Review* 90 (1): 147–165.

Hoshi, T., A. Kashyap, and D. Scharfstein. 1990. The role of banks in reducing the costs of financial distress in Japan. *Journal of Financial Economics* 27 (1): 67–89.

———. 1993. The choice between public and private debt: An analysis of post-deregulation corporate financing in Japan. NBER working paper 4421.

Houston, J. F., and C. M. James. 2001. Do relationships have limits? Banking relationships, financial constraints and investment. *Journal of Business* 74 (3): 347–375.

Hughes, J. P., and L. J. Mester. 1993a. A quality and risk-adjusted cost function for banks: Evidence on the too-big-to-fail doctrine. *Journal of Productivity Analysis* 4 (3): 293–315.

———. 1993b. Accounting for the demand for financial capital and risk-taking in bank cost functions. Working paper 93-17, Federal Reserve Bank of Philadelphia, Economic Research Division.

———. 1994. Bank managers' objectives. Working paper 94-8, Federal Reserve Bank of Philadelphia, Economic Research Division.

Hunter, W. C., S. G. Timme, and W. K. Yang. 1990. An examination of cost subadditivity and multiproduct production in large U.S. banks. *Journal of Money, Credit and Banking* 22: 504–525.

James, C. 1987. Some evidence on the uniqueness of bank loans. *Journal of Financial Economics* 19 (2): 217–235.

Jappelli, T., and M. Pagano. 2000. Information sharing in credit markets: A survey. Working paper no. 36, Center for Studies in Economics and Finance, University of Salerno, Italy.

Jullien, B. 2005. Two-sided markets and electronic intermediaries. *CESifo Economic Studies* 51: 235–262.

Kashyap, A., and J. C. Stein. 1994. Monetary policy and bank lending. In *Monetary Policy*, ed. N. G. Mankiw. Chicago: University of Chicago Press.

Kashyap, A., J. C. Stein, and D. W. Wilcox. 1993. Monetary policy and credit conditions: Evidence from the composition of external finance. *American Economic Review* 83 (1): 78–98.

Keeley, M. 1990. Deposit insurance, risk and market power in banking. *American Economic Review* 80: 1183–1200.

Kim, H. Y. 1986. Economies of scale and economies of scope in multiproduct financial institutions: Further evidence from credit unions. *Journal of Money, Credit and Banking* 18 (2): 220–226.

Kim, M. 1986. Banking technology and the existence of a consistent output aggregate. *Journal of Monetary Economics* 18 (2): 181–195.

King, R. G., and R. Levine. 1993. Finance and growth: Schumpeter might be right. *The Quarterly Journal of Economics* 108 (3): 717–737.

Klein, M. 1971. A theory of the banking firm. *Journal of Money, Credit and Banking* 3: 205–218.

Kolari, J., and A. Zardhooki. 1987. *Bank costs, structure and performance*. Lexington, Mass.: D. C. Heath.

Longhofer, S. D., and J.A.C. Santos. 2000. The importance of bank seniority for relationship lending. *Journal of Financial Intermediation* 9: 57–89.

Lummer, S., and J. McConnell. 1989. Further evidence on the bank lending process and the reaction of the capital market to bank loan agreements. *Journal of Financial Economics* 25: 99–122.

Marquez, R. 2002. Competition, adverse selection, and information dispersion in the banking industry. *Review of Financial Studies* 15 (3): 901–926.

Matutes, C., and J. Padilla. 1994. Shared ATM networks and banking competition. *European Economic Review* 38 (5): 1113–1138.

Matutes, C., and X. Vives. 1996. Competition for deposits, fragility, and insurance. *Journal of Financial Intermediation* 5 (2): 184–216.

McAllister, P. H., and D. A. McManus. 1992. Diversification and risk in banking: Evidence from ex post returns. Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 201, Federal Reserve Board, Washington, D.C.

Mester, L. J. 1987. A multiproduct cost study of savings and loans. *Journal of Finance* 42 (2): 423–445.

———. 1991. Agency costs among savings and loans. *Journal of Financial Intermediation* 1 (3): 257–278.

———. 1992. Traditional and nontraditional banking: An information-theoretic approach. *Journal of Banking and Finance* 16 (3): 545–566.

———. 1993. Efficiency in the savings and loans industry. *Journal of Banking and Finance* 17 (2–3): 267–286.

Monti, M. 1972. Deposit, credit, and interest rate determination under alternative bank objectives. In *Mathematical methods in investment and finance*, ed. G. P. Szego and K. Shell. Amsterdam: North-Holland.

Murray, J. D., and R. W. White. 1983. Economies of scale and economies of scope in multi-product financial institutions: A study of British Columbia credit unions. *Journal of Finance* 38 (3): 887–902.

Neuberger, J., and G. Zimmerman. 1990. Bank pricing of retail deposit accounts and "the California rate mystery." Federal Reserve Bank of San Francisco. *Economic Review*, Spring, 3–16.

Newmark, D., and S. Sharpe. 1992. Market structure and the nature of price rigidity: Evidence from the market for consumer deposits. *Quarterly Journal of Economics* 102: 657–680.

Ongena, S., and D. C. Smith. 2001. The duration of bank relationships. *Journal of Financial Economics* 61: 449–475.

Padilla, A., and M. Pagano. 1997. Endogenous communication among lenders and entrepreneurial incentives. *Review of Financial Studies* 10 (1): 205–236.

———. 2000. Sharing default information as a borrower discipline device. *European Economic Review* 44 (10): 1951–1980.

Pagano, M., and T. Jappelli. 1993. Information sharing in credit markets. *Journal of Finance* 48 (5): 1693–1718.

Parker, G., and M. Van Alstyne. 2005. Mechanism design to promote free market and open source software innovation. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*, 231.2.

Petersen, M., and R. G. Rajan. 1994. The benefits of lending relationships: Evidence from small business data. *Journal of Finance* 49 (1): 3–37.

———. 1995. The effect of credit market competition on lending relationships. *Quarterly Journal of Economics* 110: 407–443.

———. 2002. Does distance still matter? The information revolution in small business lending. *Journal of Finance* 57: 2533–2570.

Rajan, R. G. 1992. Insiders and outsiders: The choice between informed and arm's-length debt. *Journal of Finance* 47 (4): 1367–1400.

Rajan, R., and L. Zingales. 1998. Financial dependence and growth. *American Economic Review* 88 (3): 559–586.

Rochet, J. C., and J. Tirole. 1996. Interbank lending and systemic risk. *Journal of Money, Credit and Banking* 28 (4): 733–762.

———. 2002. Cooperation among competitors: The economics of payment card associations. *Rand Journal of Economics* 33 (4): 1–22.

———. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1 (4): 990–1029.

———. 2006. Two-sided markets: A progress report. *Rand Journal of Economics* 37 (3): 645–667.

Rothschild, M., and J. E. Stiglitz. 1970. Increasing risk: I. A definition. *Journal of Economic Theory* 2: 225–243.

Salop, S. 1979. Monopolistic competition with outside goods. *Bell Journal of Economics* 10 (1): 141–156.

Sealey, C. W., and J. T. Lindley. 1977. Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance* 32: 1251–1266.

Shaffer, S. 1998. The winner's curse in banking. *Journal of Financial Intermediation* 7 (4): 359–392.

Shaked, A., and J. Sutton. 1982. Relaxing price competition through product differentiation. *Review of Economic Studies* 49: 3–13.

Sharkey, W. 1982. *The theory of natural monopoly*. Cambridge: Cambridge University Press.

Sharpe, S. 1990. Asymmetric information, bank lending and implicit contracts: A stylized model of customer relationships. *Journal of Finance* 45 (4): 1069–1087.

Slovin, M. B., M. E. Sushka, and J. A. Polonchek. 1993. The value of bank durability: Borrowers as bank stakeholders. *Journal of Finance* 48: 289–302.

Stahl, D. O. 1988. Bertrand competition for inputs and Walrasian outcomes. *American Economic Review* 78 (1): 189–201.

Stiglitz, J. E., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.

Suarez, J. 1994. Closure rules, market power, and risk-taking in a dynamic model of bank behavior. London School of Economics/Financial Markets Group Paper No. 196.

Tirole, J. 1988. *The theory of industrial organization*. Cambridge, Mass.: MIT Press.

Von Thadden, E. L. 1996. Optimal liquidity provision and dynamic incentive compatibility. Cahiers de Recherches Economiques 9604. Université de Lausanne, Switzerland.

———. 2004. Asymmetric information, bank lending and implicit contracts: The winner's curse. *Finance Research Letters* 1: 11–23.

Wright, J. 2003. Optimal card payment systems. *European Economic Review* 47 (4): 587–612.

Yafeh, J., and O. Yosha. 2003. Large shareholders and banks: Who monitors and how? *The Economic Journal* 113 (484): 128–146.

Yanelle, M. O. 1989. The strategic analysis of intermediation. *European Economic Review* 33 (2-3): 294–301.

Zarutskie, R. 2006. Evidence on the effects of bank competition on firm borrowing and investment. *Journal of Financial Economics* 81 (3): 503–537.

# 4 The Lender-Borrower Relationship

When a bank grants a loan to a borrower, both parties typically sign a contract. Ideally, it would be useful to specify in this contract all the obligations of the two parties in every possible future contingency. Even in the case of a one-period contract, this would mean writing down a complete list of these contingencies (states of nature) at the end of the period and specifying, for each of these contingencies, the amount of the repayment to the lender. In a dynamic (multiperiod) setting, things are even more complicated. A complete contingent contract would have to specify as well, in every state of nature and at every interim date,

1. the amount of repayment or (possibly) the amount of additional loan,
2. the interest rate on the remaining debt,
3. a possible adjustment in the collateral required by the lender,
4. the actions (in particular investment decisions) to be undertaken by the borrower.

In practice, debt contracts are often less complex. In general, repayment obligations (points 1 and 2) and collateral (point 3) are specified for the whole duration of the contract, whereas actions to be taken (point 4) are left to the borrower. Sometimes, however, some covenants are stipulated, specifying when default can be declared, in which case creditors take over the debtor's assets. Therefore, loan contracts are much less flexible than one could expect because writing a complete contingent contract would be prohibitively costly.

These issues are crucial in corporate finance because they explain the use of second-best financial contracts. Harris and Raviv (1991; 1992) provide interesting surveys of these questions, with particular attention to the famous results of Myers and Majluf (1984) and Jensen and Meckling (1976). The objective here is more limited. This chapter discusses only the aspects of the lender-borrower relationship that concern banking, leaving aside the issues related to the financial structure of firms.

Section 4.1 presents the benchmark case of symmetric information, in which the characteristics of the loan contract are determined only by risk-sharing considerations.

The discussion shows that this is not enough to explain all the features of bank loans. Section 4.2 studies one of the most popular paradigms for explaining the lack of flexibility of loan contracts, namely, the costly state verification model of Townsend (1979), further developed by Gale and Hellwig (1985). In this model it is assumed that the lender cannot observe the result $y$ of the investment made by the borrower unless a costly audit is performed. In that case, incentive compatibility conditions imply that, absent auditing, the repayment cannot depend on $y$. Typically, the optimal contract is such that an audit takes place only when the cash flows are so low that the (fixed) agreed repayment is not feasible. This is interpreted as failure, in which case the lender seizes all the cash flows.

Another interesting issue involves the incentives to repay in a dynamic context (section 4.3). The discussion starts with the case of corporate debt, as studied by Bolton and Scharfstein (1990) (section 4.3.1). It proceeds to the study of strategic debt service for a private debtor with inalienable human capital (section 4.3.2), the impact of judicial enforcement (section 4.3.3), and the case of sovereign debt (section 4.3.4).

Section 4.4 is dedicated to the subject of moral hazard, and section 4.5 presents the incomplete contract approach. As a complement, section 4.6 studies the possible use of collateral and loan size as devices for screening heterogeneous borrowers.

## 4.1 Why Risk Sharing Does Not Explain All the Features of Bank Loans

This section presents the simple model of the lender-borrower relationship that is used throughout this chapter and studies the benchmark case of symmetric information. In this case, the analysis focuses on optimal risk sharing between the two parties, the lender (or investor) and the borrower (or entrepreneur). Assume only one good and two dates.

At date 0 the borrower has the possibility of investing some quantity $I$ (assumed to be fixed) of the good, which will produce in return a random quantity $\tilde{y}$ of the same good at date 1. For simplicity, assume that the borrower has no private resources at date 0 and borrows $I$ from the lender. Therefore, $I$ designates the amount of the loan. Assume that both agents consume only at date 1 and that their preferences are characterized by Von Neumann–Morgenstern utility functions $u_L$ (for the lender) and $u_B$ (for the borrower), assumed to be twice continuously differentiable, concave, and strictly increasing.

If result $\tilde{y}$ of the investment is observable by both agents (a situation of symmetric information), these agents can sign a contract specifying in advance how they will share $\tilde{y}$ at date 1. This sharing rule is completely determined once the repayment $R(y)$ to the lender is specified as a function of the realization $y$ of $\tilde{y}$. The borrower then gets $y - R(y)$. The family of optimal debt contracts (under symmetric information) can be obtained parametrically as the solution of the following program $\mathscr{P}_0$:

$$\begin{cases} \max\limits_{R(\cdot)} \; Eu_B(\tilde{y} - R(\tilde{y})) \\ \text{under } Eu_L(R(\tilde{y})) \geq U_L^0. \end{cases} \tag{4.1}$$

where the parameter $U_L^0$ denotes the expected utility demanded by the lender (individual rationality level). Since $u_B$ and $u_L$ are monotonic, (4.1) will always be binding. Note that optimal contracts could be obtained as well by maximizing the expected utility of the lender under an individual rationality constraint for the borrower. Therefore, the lender and the borrower play completely symmetric roles, and the features of optimal contracts will be determined purely by risk-sharing considerations. The solution of $\mathscr{P}_0$ is characterized by the equality of marginal rates of substitution across states for the two agents. For all $y_1$ and $y_2$ in the support of $\tilde{y}$, one must have

$$\frac{u_L'[R(y_1)]}{u_L'[R(y_2)]} = \frac{u_B'[y_1 - R(y_1)]}{u_B'[y_2 - R(y_2)]}, \tag{4.2}$$

or, put in another way, the ratio of the marginal utilities of the two agents is a constant $\mu$ in the support of $\tilde{y}$. For all $y$ in this support,

$$\frac{u_B'(y - R(y))}{u_L'(R(y))} = \mu. \tag{4.3}$$

Of course, $\mu$ depends on the individual rationality level $U_L^0$ demanded by the lender. If the logarithm of (4.3) is differentiated with respect to $y$, the following result is obtained for all $y$ in the support of $\tilde{y}$:

$$\frac{u_B''}{u_B'}(y - R(y))(1 - R'(y)) - \frac{u_L''}{u_L'}(R(y))R'(y) = 0.$$

This gives a relation between $R'(y)$ and the absolute indexes of risk aversion of the two agents, defined by

$$I_B(y) = -\frac{u_B''(y)}{u_B'(y)} \quad \text{and} \quad I_L(y) = -\frac{u_L''(y)}{u_L'(y)}.$$

A classical result is obtained that can be traced back to Wilson (1968):

**Result 4.1** Optimal debt contracts under symmetric information are characterized by the condition

$$R'(y) = \frac{I_B(y - R(y))}{I_B(y - R(y)) + I_L(R(y))}.$$

This result can be easily interpreted.[1] The sensitivity of the repayment $R(y)$ to the result $y$ is high when the borrower is more risk-averse than the lender ($I_B/I_L$ large) and low in the reverse case.

This finding is not very satisfactory in the banking context. Indeed, banks typically have large diversified portfolios, which means that in general they are approximatively neutral vis-à-vis the small risks of individual loans. But result 4.1 suggests that $R'(y)$ should be close to 1, whereas the typical bank loan involves instead a constant repayment ($R(y) \equiv R$).

As has just been shown, although risk-sharing considerations can explain some features of the loan contract (see, e.g., Alm and Follain 1982), it alone cannot explain the widespread use of standard debt contracts. Thus, we abandon the symmetrical treatment of the lender and borrower. Later sections introduce a fundamental asymmetry between them, by considering that the observation of $\tilde{y}$ by the lender is costly (section 4.2) or even impossible (section 4.3).

## 4.2 Costly State Verification

Following Townsend (1979) and Gale and Hellwig (1985), this section modifies the model of section 4.1 by assuming that the realization $y$ of $\tilde{y}$ is not observable by the lender unless the lender undertakes an audit, which costs $\gamma$. The rules of the contract to be signed between lender and borrower are now more complex. The contract must specify when an audit will be undertaken and how its result will affect the payment to the lender. Using the *revelation principle* (see, e.g., Fudenberg and Tirole 1991; Mas Colell, Whinston, and Green 1995), the contract may be described (without loss of generality) by a revelation mechanism in which the borrower is asked to report $y$ and in which the rules of the mechanism are designed in such a way that it is always in the interest of the borrower to report truthfully. Therefore, the contract can be described as

- *a repayment function* $\hat{y} \to R(\hat{y})$, transfer promised by the borrower to the lender, as a function of the report $\hat{y}$ sent by the borrower;
- *an auditing rule*, identified as a set $\mathscr{A}$ of reports of the borrower for which the lender undertakes an audit;
- *a penalty (or reward) function* $P(y, \hat{y})$ specifying a possible additional transfer between the borrower and the lender after the audit, and depending on the result $y$ of the audit and on the report $\hat{y}$ previously sent by the borrower.

This array $(R(\cdot), \mathscr{A}, P(\cdot\,,\cdot))$ specifies a *direct revelation mechanism* in the language of contract theory. This mechanism has to fulfill the incentive compatibility constraints, ensuring that truthful reporting ($\hat{y} = y$) is a dominant strategy. We also require limited liability (or positive consumption for both agents): $0 \le R(y) \le y$ for all $y$. The next section characterizes the set of incentive-compatible mechanisms. It then shows that efficient incentive-compatible contracts are standard debt contracts, characterized by repayment function $R(y) = \min(y, R)$, in which the borrower promises a fixed repayment $R$, and the bank seizes the entire cash flow $y$ when the borrower cannot repay $R$. We also study what happens when borrowers can falsify their reports.

### 4.2.1 Incentive-Compatible Contracts

The optimal penalty function $P(y, \hat{y})$ can be taken as arbitrarily large for $\hat{y} \neq y$ and normalized to zero for $\hat{y} = y$. Thus it is easy to prevent untruthful reporting in the auditing region, and therefore (this is in fact a convention) truthful reporting need not be rewarded.

Preventing nontruthful reporting in the no-audit region (the complement of $\mathscr{A}$) imposes strong constraints. The repayment function is necessarily constant outside the verification region (on the complement of $\mathscr{A}$) because otherwise the borrower could cheat by announcing the message that corresponds to the minimum repayment in the no-audit zone. Denote by $R$ the (constant) value of this repayment function on the complement of $\mathscr{A}$.

Finally, in order to prevent the borrower from fraudulous avoidance of auditing, $R$ cannot be smaller than the maximum repayment possible on $\mathscr{A}$. Otherwise, the borrower would have an interest, for some realizations of $y$ in $\mathscr{A}$, in reporting a message in the no-audit region and paying $R$; then the mechanism would not be incentive-compatible.

**Result 4.2a** A debt contract is incentive-compatible if and only if there exists a constant $R$ such that

$$\begin{cases} \forall y \notin \mathscr{A} & R(y) \equiv R, \\ \forall y \in \mathscr{A} & R(y) \leq R. \end{cases}$$

***Proof*** We have seen that these conditions were necessary for incentive compatibility. Let us now establish the converse. We consider a contract specifying a constant repayment $R$ outside the audit region $\mathscr{A}$, and a repayment $R(y) \leq R$ in the audit region. To show that this contract is incentive-compatible consider two cases:

- $y \in \mathscr{A}$ (the true return is in the audit region). Clearly the borrower has no interest in sending a false message $\hat{y}$ that belongs to $\mathscr{A}$. He would be detected and punished. If he sends a false message $\hat{y}$ that does not belong to $\mathscr{A}$, he has to pay $R \geq R(y)$, which is not in his interest. Truthful behavior is thus guaranteed in the audit region.
- $y \notin \mathscr{A}$ (the true return is outside the audit region). In this case, sending a message $\hat{y} \notin \mathscr{A}$ (outside the audit region) would not change anything (because repayment is constant outside $\mathscr{A}$), whereas by sending a message $\hat{y} \notin \mathscr{A}$ (in the audit region) the borrower would provoke the audit. Thus his manipulation would be detected and punished. ■

### 4.2.2 Efficient Incentive-Compatible Contracts

The next task is to select, among these incentive-compatible debt contracts, those that are efficient. Assume that both agents are risk-neutral, so that risk-sharing considerations are irrelevant. Efficient incentive-compatible debt contracts are then obtained by minimizing the probability of an audit for a fixed expected repayment, or equivalently, by maximizing the expected repayment for a fixed probability of an audit. In view of result 4.2a, for a given expected repayment $E[R(y)]$, an incentive-compatible debt contract will be efficient only if $R(y)$ is maximum in the audit region. We now establish that an incentive-compatible debt contract $(R^*(\cdot), \mathscr{A}^*)$ is efficient if and only if

- $\forall y \in \mathscr{A}^*$, $R^*(y) = \min(y, R^*)$ (maximum repayment in the audit zone, taking into account limited liability and incentive compatibility constraints);
- $\mathscr{A}^* = \{y | y < R^*\}$, (an audit will take place only when reimbursement is less than $R^*$—bankruptcy).

This can be interpreted as a standard debt contract.

**Result 4.2b** If both agents are risk-neutral, any efficient incentive-compatible debt contract is a standard debt contract.

***Proof*** We have to establish that among the incentive-compatible contracts characterized in result 4.2.a, the efficient ones are characterized by a repayment function $R^*(y) = \min(y, R^*)$ with an audit region $\mathscr{A}^* = \{y | y < R^*\}$ (audit if and only if the borrower declares that he cannot repay $R^*$). Take any other incentive-compatible contract $(R(\cdot), \mathscr{A})$ associated with the same probability of audit but with a different audit region $\mathscr{A} \neq \mathscr{A}^*$. By result 4.2.a, this contract is associated with a constant repayment $R$ outside $\mathscr{A}$: $\forall y \notin \mathscr{A}$, $R(y) = R$. Limited liability implies that $y \geq R$ for all $y \notin \mathscr{A}$, which means that $R$ has to be less than $\mathit{inf}\{y, y \notin \mathscr{A}\}$. Since $\mathscr{A}$ has the same probability as $\mathscr{A}^* = \{y | y < R^*\}$ but is different from $\mathscr{A}^*$, there is at least one element $y_0$ of $\mathscr{A}^*$ that does not belong to $\mathscr{A}$. Therefore, $R \leq y_0 < R^*$. From this property, we can now deduce that the expected repayment is lower with this new contract. Indeed result 4.2.a and limited liability imply that $R(y) \leq \min(y, R)$ for all $y$. Since $R < R^*$, $R(y) \leq R^*(y) = \min(y, R^*)$ with strict inequality in the no-audit region. This establishes that any incentive-compatible contract with the same probability of auditing as a standard debt contract $(R^*(y) = \min(y, R^*), \mathscr{A}^* = \{y | y < R^*\})$ gives a lower expected repayment. Thus only standard debt contracts can be efficient among incentive-compatible contracts. ■

Figure 4.1 illustrates this result by comparing incentive-compatible contracts giving the same expected repayment for the borrower. The inefficient contract necessarily has a higher probability of audit: $\mathscr{A} \supset \mathscr{A}^*$. If risk aversion is introduced, optimal

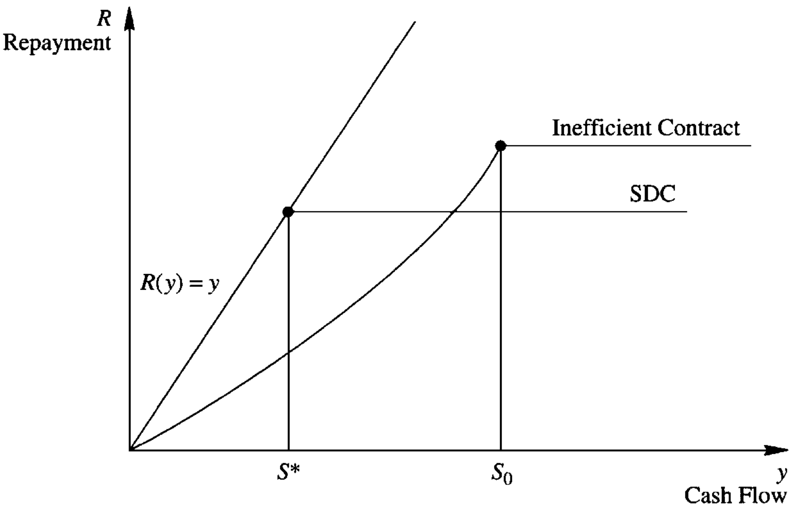



**Figure 4.1**
Optimality of the standard debt contract under costly state verification.

contracts are more complex and do not always correspond to standard debt contracts. Moreover, even if agents are risk-neutral, standard debt contracts can be dominated if the situation allows for stochastic auditing procedures (see problem 4.7.4). Also, it may not be easy for the lender to commit to an audit when the borrower defaults.[2]

### 4.2.3 Efficient Falsification-Proof Contracts

This section briefly addresses the issue of falsification, which arises when the borrower can manipulate the reported cash flow at a certain cost. This model keeps exactly the same framework and notations that have already been introduced but assumes that there is a cost $c(y, \hat{y})$ incurred by the borrower for reporting $\hat{y}$ when $y$ has occurred (with $c(y, y) = 0$, which means that truthful reporting is costless). Lacker and Weinberg (1989) address this problem in a general setting. The aim here is simply to illustrate how falsification may alter the characteristics of the optimal repayment function $R(\hat{y})$.

Assume that the cost of falsification is $c(y, \hat{y}) = \gamma|y - \hat{y}|$, where $\gamma$ is positive but smaller than 1.[3] A borrower who reports $\hat{y}$ after obtaining $y$ obtains the following profit:

$$\pi_B = y - R(\hat{y}) - \gamma|y - \hat{y}|.$$

The mechanism will be falsification-proof if for all $y$, this expression has a maximum for $\hat{y} = y$ (no falsification). Since $\pi_B$ has a kink for $\hat{y} = y$, this is equivalent to requiring that for all $y$,

$$-\gamma \leq R'(y) \leq \gamma. \tag{4.4}$$

Because of limited liability constraints, $R(0)$ necessarily equals 0, which together with the above inequality, implies that for all $y$,

$$R(y) \leq \gamma y.$$

The maximum expected repayment for the lender is therefore $\gamma E(y)$.

The possibility of falsification thus imposes a severe constraint on the projects that can be funded. If $L$ represents the size of the loan and $r$ the interest rate demanded by the bank, a necessary condition for funding is

$$\gamma E(y) \geq (1+r)L. \tag{4.5}$$

If this condition is satisfied, funding is possible, and the characteristics of the optimal repayment function can be investigated. The most realistic case is when the borrower is risk-averse, whereas the lender is risk-neutral. In the absence of falsification possibilities, risk-sharing considerations alone would lead to a repayment function such that $R'(y)$ equals 1 (see result 4.1). This violates the no-falsification constraint (4.4). In fact, the optimal falsification-proof contract is such that $R'(y)$ is as close as possible to 1.[4] The following result is obtained:

$$R(y) = \max(0, \gamma y - \alpha),$$

where $\alpha$ is a positive number determined by the lender participation constraint.

## 4.3 Incentives to Repay

This section considers a more extreme framework than the costly state verification paradigm. Assume that an audit is impossible and that the borrower will repay only when incentives are present. Section 4.3.1 analyzes the Diamond (1984) model of nonpecuniary cost of bankruptcy, and section 4.3.2 examines the general model of Bolton and Scharfstein (1990), who assume that the returns of the borrower's investment are not verifiable by a third party and thus are noncontractible. Section 4.3.3 studies the case in which the borrower can dispute the lender's claim in court (Jappelli, Pagano, and Bianco 2005), thus exploring the effects of judicial enforcement on the credit market.

### 4.3.1 Nonpecuniary Cost of Bankruptcy

A result related to Gale and Hellwig (1985), establishing the optimality of standard debt contracts, was previously obtained by Diamond (1984) (see section 2.4) within a

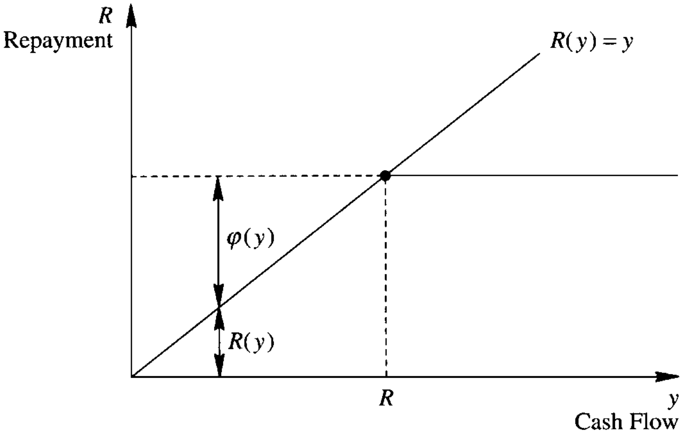



**Figure 4.2**
Optimality of the standard debt contract under nonpecuniary costs of bankruptcy.

similar context of risk neutrality. The objective is also to obtain truthful revelation of the borrower's cash flows $y$. The difference is that in Diamond's model cash flows are not observable (or equivalently, auditing costs are infinite), so that mechanisms have to be defined only for $y \notin \mathscr{A}$. But result 4.2a shows that this implies a constant repayment $R$, which has to be smaller than the smallest possible value of $y$. To go beyond this uninteresting case, Diamond assumes that the contract may also include a nonpecuniary cost $\varphi(y)$ that the lender can inflict on the borrower (e.g., a loss of reputation). This modifies the incentive compatibility condition, which now becomes

$$R(y) + \varphi(y) = R,$$

and is interpreted as the indifference of the borrower to announce any cash flow level, since the total (pecuniary plus nonpecuniary) cost is constant.

Efficient contracts, then, are those that minimize the expected nonpecuniary cost. This leads to minimizing the set on which $\varphi(y) > 0$ and taking the minimum possible value $\varphi(y)$; that is, $\varphi(y) = R - y$. A standard debt contract is thus obtained (fig. 4.2). However, even if nonpecuniary penalties (such as prison for debt) exist in practice, it is hard to believe that they can be so finely tuned. Also, the introduction of a nonpecuniary cost is equivalent to a violation of the limited liability constraint.

### 4.3.2 Threat of Termination

Bolton and Scharfstein (1990) study a repeated borrower-lender relationship in which the threat of termination by the lender provides incentives for the borrower to repay the loan.[5] In its simplest formulation, their model considers an entrepreneur who owns a technology that transforms a fixed amount 1 into a random cash flow $\tilde{y}$.

This technology can be used repeatedly, at discrete dates $t = 0, 1, \ldots$, and cash flows are independently identically distributed (i.i.d.) across dates. The entrepreneur has no resources of his own and can invest at a given date only if a bank grants him a loan. For simplicity, assume risk neutrality and no discounting, and $E(\tilde{y}) > 1$. Therefore, in a world of symmetric information, the investment would be undertaken. However, the realized cash flows are not observable by the bank (or verifiable by a court of justice). This may lead the bank to refuse to grant the loan. For example, Bolton and Scharfstein examine a case in which the cash flows $\tilde{y}_t$ can take only two values, high $(y_H)$ with probability $p_H$ or low $(y_L)$ with probability $p_L = 1 - p_H$, and they assume that $1 > y_L$. Then the borrower can always pretend that $\tilde{y} = y_L$, which is therefore the maximum repayment that the bank can enforce. In a one-shot relationship, there would be no lending because it would lead to a deficit of $1 - y_L$ for the bank.

However, in a two-period relationship, the bank can threaten to terminate the relationship at $t = 1$ if the firm does not repay $R$, but commit to renew (at $t = 2$) the initial loan if the firm repays $R > y_L$ at the end of period 1 (which can only occur if $\tilde{y}_1 = y_H$). Of course, at the end of period 2 the firm will be in the same situation as in a one-shot relationship and will always repay the minimum possible amount. Therefore, the bank knows that it will lose money at $t = 2$ (if the second loan is granted), but the first repayment $R$ can be sufficiently high to compensate for this loss. More specifically, the expected present value of the bank profit is

$$\pi = -1 + p_L y_L + p_H(R - 1 + y_L),$$

which can also be written as

$$\pi = p_H(R - 1) - 1 + y_L.$$

The bank will sign the two-period contract if $\pi$ is non-negative, or if

$$R \geq 1 + \frac{1 - y_L}{p_H}. \tag{4.6}$$

It must be confirmed that the borrower has incentives to repay when $\tilde{y}_1 = y_H$:

$$-R + p_H(y_H - y_L) \geq -y_L,$$

which is equivalent to

$$R \leq E(\tilde{y}). \tag{4.7}$$

Therefore, any repayment $R$ that satisfies both conditions (4.5) and (4.6) will be acceptable to the lender. Such a repayment exists if and only if

$$1 - y_L \leq p_H[E(\tilde{y}) - 1]. \tag{4.8}$$

If this condition is satisfied, the threat of termination provides the incentives to repay in a two-period model. Gromb (1994) provides a detailed analysis of an extension of this model to several periods (see also Dewatripont and Maskin 1995).

It is interesting to remark that this model and Diamond's (1984) share the assumption that lenders cannot observe cash flows. Diamond's solution is based on the existence of a nonpecuniary cost. In Bolton and Scharfstein, the incentives are brought in through the threat not to lend in the future. Haubrich (1989) combines the two types of mechanisms in an infinite horizon model with no discounting. The optimal contract involves a test that detects any agent who systematically cheats (in an infinite horizon model with no discounting, cheating a finite number of times is irrelevant). Honest unlucky borrowers will almost never be punished. Also, punishment need not entail a cessation of credit but may take the form of higher interest rates, as the net present value of these future additional costs may be sufficient to make strategic default unattractive.

### 4.3.3 Impact of Judicial Enforcement

Jappelli, Pagano, and Bianco (2005) explore the impact of the judiciary system on the terms and availability of credit. The cost and length of trials, together with the degree of protection given to the borrower, determine the latter's opportunities to renegotiate the contracted repayment.

Consider a zero-interest, risk-neutral environment where entrepreneurs ask for credit to finance projects. Projects succeed with probability $p$, in which case they yield $y$, and fail with the complementary probability $1 - p$, in which case the return is zero. The success or failure of a project is publicly observable. In order to obtain funding for the project, the entrepreneur may pledge some collateral $C$. The characteristics of judicial enforcement are then given by two parameters: the recovery rate $\emptyset_p$ on the firm's cash flow $y$ and the recovery rate $\emptyset_c$ on the external collateral $C$. These recovery rates are determined by the costs of judicial enforcement, paid to third parties, and the ability of the borrower to retain a fraction of the payment owed.

From the lender's perspective, $\emptyset_p$ and $\emptyset_c$ are sufficient to compute the expected payment on a loan with a contractual repayment $R$ when the level of collateral $C$ is pledged:

$$p \min[R, \emptyset_p y + \emptyset_c C] + (1 - p) \min[R, \emptyset_c C].$$

This expression can be simplified in the case where the borrower prefers to repay when the project succeeds (no strategic default):

$$R \leq \emptyset_p y + \emptyset_c C.$$

If the credit market is competitive, we then have

$$1 = pR + (1 - p) \min(R, \emptyset_c C].$$

Thus there are three cases:

- If the borrower has sufficient collateral,

$$C \geq \frac{1}{\emptyset_c},$$

the loan is fully collateralized, and there is no credit risk for the lender. In this case $R = 1$ (no risk premium).

- If the borrower has too little collateral,

$$C < \frac{1 - p\emptyset_p y}{\emptyset_c},$$

then it is credit-rationed. This is because the maximum nominal repayment that avoids strategic default $(R = \emptyset_p y + \emptyset_c C)$ is not sufficient to make the lender break even (because $pR + (1 - p)\emptyset_c C < 1$).

- Finally, if the borrower has an intermediate level of collateral,

$$\frac{1 - p\emptyset_p y}{\emptyset_c} \leq C < \frac{1}{\emptyset_c},$$

then credit is possible, but for a high nominal repayment $R = \emptyset_p y + \emptyset_c C \geq 1 + (1 - p)\emptyset_p y$.

Consequently, the characteristics of the judiciary system $(\emptyset_p, \emptyset_c)$ determine which projects are financed and which have no access to credit because of insufficient collateral. The efficiency loss comes from "the interaction of judicial inefficiencies and opportunistic behavior" (Jappelli, Pagano, and Bianco 2005). Positive net present value projects, $(py > 1)$, may not be funded because of a low $\emptyset_p$ or insufficient external collateral $C$.

Note that a lower recovery rate on cash flows is here the source of the inefficient investment. For $\emptyset_p = 1$, all positive net present value projects are implemented. For lower recovery rates $\emptyset_p$, only investors with sufficient collateral or retained earnings will be able to implement their projects.

It is worth mentioning that a low level of legal enforcement, in particular regarding collateral, may be the consequence of a high level of borrower protection. The empirical evidence on Italian data support this view. The ratio of loans to GDP is negatively related to the number of pending trials per inhabitant. Finally, it is interesting to emphasize the implications of this model on the determinants of growth and on development policy. As the judiciary system improves, that is, as the coefficients $\emptyset_p$ and $\emptyset_c$ increase, the financial system becomes more efficient. This justifies a link between judicial development and economic growth (see section 6.4).

### 4.3.4 Strategic Debt Repayment: The Case of a Sovereign Debtor

Consider a very simple stationary model inspired by Allen (1983), in which a sovereign country makes an investment funded by a foreign bank loan $L$. This allows the country to produce output $f(L)$, assumed to be sufficient to repay $(1+r)L$. The static profit of the country is

$$\pi = f(L) - (1+r)L \geq 0,$$

where $r$ is the (exogenous) riskless interest rate. The country's demand for capital $L_D$ (obtained by maximizing $\pi$) is given by the usual condition on marginal productivity:

$$f'(L_D) = 1 + r.$$

Assume that the country is able to repudiate its debt at any moment,[6] in which case it cannot obtain a new loan; no renegotiation is possible.[7] Of course, the country could default only once, and live in autarky ever after, by simply reinvesting some fraction of its production. Assume that this is not possible, and take this as the reduced form of a more complex model in which a lag occurs between the moment at which investment is made and the moment at which the nonstorable output is obtained.[8]

The opportunity cost of default equals the present value of forgone profits (for simplicity, consider an infinite time horizon):

$$V(L) = \sum_{t=1}^{\infty} \beta^t (f(L) - (1+r)L) = \frac{\beta}{1-\beta}(f(L) - (1+r)L),$$

where $0 < \beta < 1$.

For the country to repay, it must have incentives to do so. That is, the cost of repayment has to be inferior to the opportunity cost of default:

$$(1+r)L \leq V(L).$$

This is equivalent to

$$(1+r)L \leq \beta f(L).$$

Assuming (as in fig. 4.3) that there are decreasing returns to scale, this inequality holds for $L \leq \hat{L}$, where $\hat{L}$ is the maximum loan size that satisfies this inequality.

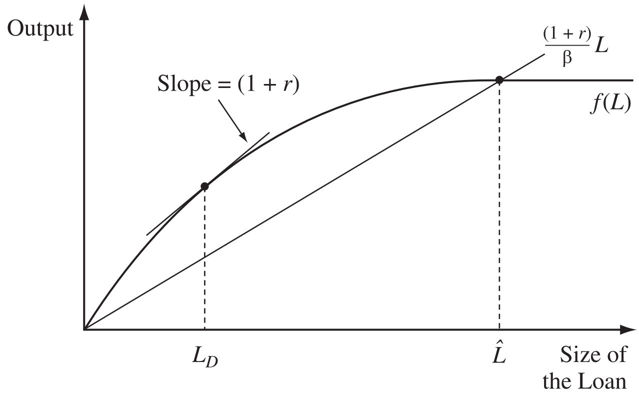



**Figure 4.3**
Underinvestment in the case of a strategic debtor (Allen 1983).

When $\beta$ is large enough, the optimal loan $L_D$ will be feasible (fig. 4.3). But when $\beta$ is small, $\hat{L}$ may be smaller than $L_D$, and the firm will be rationed.

Following Eaton and Gersovitz (1981), we now elaborate on the previous model by studying a more complete infinite-horizon stationary model in which the borrowing country is characterized by an exogenous (stochastic) output process $\tilde{y}_t$ (assumed to be independently and identically distributed on a bounded interval) and by an objective function,

$$U = E\left(\sum_{t=0}^{+\infty} \beta^t u(C_t)\right),$$

where $C_t$ represents absorption ("consumption") at date $t$, and $u$ is a Von Neumann–Morgenstern utility function with the usual properties ($u' > 0$, $u'' < 0$, and $u$ bounded above). Assume that repayment is always due at the next period (short-term borrowing). The lender may set a limit on the borrowing capacity, or more generally, restrict the amount $b$ that is borrowed, which is captured by the condition $b \in B$, where the set $B$ is chosen by the lender. Another assumption is that default is followed by definitive exclusion from future borrowing (no forgiveness).[9] In that case, the continuation payoff for the borrower corresponds to autarky:

$$U_d = E\left(\sum_{t=0}^{+\infty} \beta^t u(\tilde{y}_t)\right) = \frac{Eu(\tilde{y})}{1-\beta},$$

where $U_d$ stands for the utility of defaulting. Assuming that repayment is always feasible, strategic default will occur if and only if

$$u(y) + \beta U_d > u(y - R) + \beta V_r,$$

where $V_r$ represents the continuation payoff associated with repaying the current period loan (and defaulting later if it is optimal to do so). $V_r$ will be computed later. For the moment, notice that the condition for strategic default can also be written as

$$\psi(y, R) = u(y) - u(y - R) > \beta(V_r - U_d).$$

Since $u$ is concave, this condition is satisfied if and only if $y$ is less than some cutoff point, denoted as $\varphi(R)$. Indeed, $u'$ is decreasing, and therefore

$$\frac{\partial \psi}{\partial y}$$

is negative. Since $\psi(\varphi(R), R) = \beta(V_r - U_d)$, $\varphi'$ will have the same sign as

$$-\frac{\partial \psi}{\partial y} \Big/ \frac{\partial \psi}{\partial R}.$$

It is already known that

$$\frac{\partial \psi}{\partial y} < 0.$$

Moreover, the fact that $u$ is increasing implies

$$\frac{\partial \psi}{\partial R} > 0.$$

Therefore, $\varphi$ has to be increasing. In other words, strategic default will occur (as expected) when output is low (bad times) and will be more likely when debt is high. Under the dynamic programming principle (provided default has not occurred previously), for any given level $y$ of current output, the country will choose to borrow the amount $b(y)$ that solves

$$\max_{b \in B} \{u(y + b) + \beta E_{y'}[\max(u(y') + \beta U_d, u(y' - R(b)) + \beta V_t)]\} = V(y), \tag{4.9}$$

where $V(y)$ represents the (optimal) value function for the borrower, $R(\cdot)$ is the repayment function, and $y'$ is the unknown future output. The expression inside the expectation symbol represents the continuation payoff, after strategic debt servicing at the next period.

Assume that the lenders are risk-neutral and have a competitive behavior. An equilibrium in the credit market will then be characterized by a value function $V(y)$, a borrowing decision $b(y)$, a cutoff point $\varphi(R)$ (these three functions together characterize the borrower's behavior), and a repayment function $R(b)$ such that

• condition (4.9) is satisfied, where the maximum is attained for $b = b(y)$ and $V_r$ equals $E[V(\tilde{y})]$ (optimal borrowing decision);

• $y < \varphi(R) \Leftrightarrow u(y) + \beta U_d > u(y - R) + \beta V_r$ and repudiation occurs exactly in that case (optimal repudiation decision);

• for all $b$ in $B$,

$$R(b) = \frac{(1+r)B}{P\{\tilde{y} > \varphi(R(b))\}},$$

where $r$ represents the risk-free rate at which risk-neutral lenders can refinance their loans (zero-profit condition for lenders).

Depending on the parameters, the set $B$ may be bounded or not. Indeed, there is no reason why the function $E(R) = RP\{\tilde{y} > \varphi(R)\}$ should increase up to infinity. If $E(R)$ reaches a maximum, there will be a maximum amount that the country can borrow.

**Result 4.3** At a competitive equilibrium of the credit market, the following propositions hold true:

• Strategic default occurs when current output is low relative to outstanding debt $(y < \varphi(R))$.

• The probability of strategic default increases with the volume of outstanding debt ($\varphi$ is increasing).

• The nominal interest rate $R(b)/b$ increases with $b$.

***Proof*** The only thing not already proved is the last property. Recall that $R(b)$ is defined implicitly by

$$R(b) = \frac{(1+r)b}{P\{\tilde{y} > \varphi(R(b))\}}.$$

Therefore,

$$\frac{R(b)}{b} = \frac{1+r}{P\{\tilde{y} > \varphi(R(b))\}}.$$

$\varphi$ and $R$ are increasing. Therefore, this is also the case for $R(b)/b$. ∎

Another noteworthy property of this equilibrium is that it typically exhibits credit rationing, at least in the following sense. If borrowers were confronted with a linear repayment schedule (in which the interest rate is independent of the size of the loan), they would sometimes want to borrow more than what they get at the equilibrium discussed earlier.[10] This comes from the nonconvexity of preferences due to the repudiation option (see Eaton and Gersovitz 1981).

It is worth noting that although this model is not based on any idea of reputation building, it can be subject to a criticism of Bulow and Rogoff (1989) for reputation models. Indeed, Bulow and Rogoff use a simple arbitrage argument to show that reputation alone is not enough to ensure debt repayment. This argument can be perfectly adapted to the setting used by Eaton and Gersovitz.

The argument of Bulow and Rogoff can be summarized as follows. Two types of contracts are available to the country for smoothing its consumption. A reputation contract is one in which the country receives a loan in exchange for a state-contingent repayment. On the other hand, a cash-in-advance contract is one in which the country makes an initial payment in exchange for a series of state-contingent payments. Since with a cash-in-advance contract the country bears the credit risk, a foreign investor will always accept it and thus will act as an insurance company vis-à-vis the country's macroeconomic risk. Bulow and Rogoff show that if the country's future repayments have a positive expected present discounted value, the country will be better off ceasing payments on its reputation contract and starting a series of cash-in-advance contracts that will replicate the initial contract at a lower cost.[11]

As a consequence, the threat faced by the defaulting country (of being excluded from the credit market) loses its strength. The borrower's choice then is not to repay or to face income fluctuations without any possibility of borrowing, but rather to repay or to pay cash for an insurance mechanism. Bulow and Rogoff show that default dominates repayment. Consequently, absent penalties to the defaulting countries that will exclude them from access to the insurance mechanism, the credit market will be nonexistent.

## 4.4 Moral Hazard

It is characteristic of the banking industry for banks to behave as a sleeping partner in their usual relationship with borrowers.[12] For this reason, it seems natural to assume that banks ignore the actions borrowers are taking in their investment decisions. This is typically a moral hazard setup. The borrower has to take an action that will affect the return to the lender, yet the lender has no control over this action.

Chapter 2 presented a simple model of the credit market with moral hazard. Here we consider a more complex model with continuous returns, inspired by Innes (1990), who uses it to determine the shape of the optimal repayment function. A crucial assumption is the limited liability of the borrower.[13]

Following Innes, consider a static borrower-lender relationship in which the borrower's return $\tilde{y}$ is continuous (instead of binomial, as in chapter 2), and its distribution is influenced by an action $e$ ("effort") undertaken by the borrower and nonobservable by the lender. Assume that both agents are risk-neutral. Given a contract $R(\cdot)$, the borrower will choose the effort level $e^*$ that maximizes his net expected utility:

$$V(R, e) = \int (y - R(y)) f(y, e)\, dy - \psi(e),$$

where $f(y, e)$ is the density function of the return $y$ for a given effort level $e$, and $\psi$ is a convex increasing function representing the pecuniary equivalent of the cost of effort for the borrower. By definition of $e^*$,

$$\forall e \qquad V(R, e) \leq V(R, e^*).$$

Given an individual rationality level $U_L^0$ (minimum expected return demanded by the lender), the optimal contract will be such that it maximizes the utility of the borrower, under the effort constraint and the usual limited liability and individual rationality constraints. Therefore, the program to be solved is

$$\mathscr{P}\begin{cases} \max\ V(R, e^*) \\ 0 \leq R(y) \leq y \qquad \forall y, \\ V(R, e) \leq V(R, e^*) \qquad \forall e, \\ E[R(y)|e^*] \geq U_L^0. \end{cases}$$

**Result 4.4** If for all $e_1 > e_2$, the likelihood ratio

$$\frac{f(y, e_1)}{f(y, e_2)}$$

is an increasing function of $y$ (monotone likelihood ratio (MLR) property),[14] the optimal repayment function is always of the following type:

$$R(y) = \begin{cases} 0 & \text{for } y \geq y^*, \\ y & \text{for } y < y^*. \end{cases}$$

Figure 4.4 shows the shape of the optimal repayment function.

The MLR property (Holmström 1979) means that the result $y$ of the investment is an appropriate signal for inference of the effort level. The higher $y$ is, the higher the relative likelihood that the effort has been high rather than low. Therefore, the intuition behind result 4.4 is that the best way to provide correct incentives for effort is to give the agent maximal reward, $R(y) = 0$, when the result is good, $y \geq y^*$, and maximal penalty, $R(y) = y$, when the result is bad, $y < y^*$ (for a formal proof, see problem 4.7.2). Unfortunately, this type of contract is not frequently seen in practice.

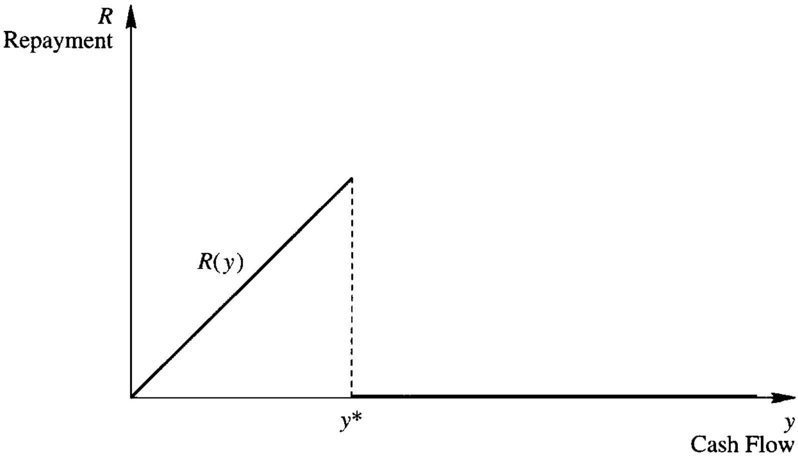



**Figure 4.4**
Optimal contract in Innes (1987) moral hazard model.

Innes (1990) has studied what happens if one restricts the problem further by requiring that the repayment function be nondecreasing in $y$.[15] In that case, Innes shows that the optimal contract is a standard debt contract.

Several articles, such as one by Diamond (1991), have studied moral hazard in a dynamic context (see chapter 2). In such a context, banks can offer multiperiod contracts, possibly involving cross-subsidies across periods. This implies that bank commitment is crucial here because banks would be better off by simply reneging on their promise to renew the loans of successful borrowers.

This important issue should be addressed in a general setting. As emphasized by Boot, Thakor, and Udell (1991), if a firm has to make a nonobservable investment before obtaining the funds to finance a project, lack of commitment on the part of the bank to grant a loan with prespecified terms will reduce the level of nonobservable investment below the efficient one. This is because of the classical holdup problem: the bank can ex post expropriate the firm once the investment cost is sunk.

What makes banks' commitment credible? There are two main reasons a bank might prefer to honor a contract rather than renege on it: (1) the probability of successful legal recourse by the borrower may be sufficiently high (Boot, Thakor, and Udell 1991), or (2) the bank may have built a reputation for honoring its contracts. In the latter case, it is interesting to note that the bank may sometimes choose to "liquefy its reputational capital" (lose its reputation) rather than face an important (capital) loss in financially impaired states, a point emphasized by Boot, Greenbaum, and Thakor (1993).

## 4.5 The Incomplete Contract Approach

Economic theory establishes that writing complete contracts (contracts that are contingent on all future states of nature) can only improve efficiency because it allows for complete risk sharing across these future states. Still, complete contracts are not seen in practice, and even more at odds with this theory, renegotiations often occur after a contract has been signed. A typical example is when a firm goes bankrupt, which triggers a bargaining process involving all the claim holders. An obvious reason might be that it was ex ante too difficult to describe all the events that would lead to bankruptcy and all the actions that the firm should take in each of them.

Incomplete contract theory recognizes this fact and allows modeling of this type of situation. It deals with situations where the states of nature are observable by the two parties to the contract but are not *verifiable*, which means that a third party would not be able to observe the state of nature that has occurred, as happens in the models of Bolton and Scharfstein (1990) and Hart and Moore (1994). Consequently, a contingent contract, if written, would not have any legal value, since no court would be able to determine the contingent obligations of either party.

An incomplete contract will typically involve some delegation and allocation to one of the parties of the power to choose among a predetermined set of actions (investment choice, renewal of a loan, issuing new shares) and will make this power contingent on the realization of a verifiable signal. Thus, for instance, a contract may specify that in case of default, creditors will take over the firm.

In general, these verifiable signals are not perfectly correlated with the nonverifiable states of nature. Typically, the agent in charge will act according to its own objective function and may not choose the most efficient course of action. In such a case, there is scope for renegotiation. Of course, agents will rationally anticipate this from the start.

The objective of this section is not to study all the incomplete contract models[16] that have dealt with the borrower-lender relationship but to briefly present one example of such contributions to give a flavor of how incomplete contracts might improve the understanding of the lender-borrower relationship.

A general conclusion of these models is that the design of contracts should limit the tendency of agents to behave inefficiently. This implies, for instance, a different interpretation of bankruptcy, which plays a role because it allocates control to a different party or because it is a credible threat that gives an incentive to the agent in charge to choose the efficient action.[17] As a simple example in the spirit of Hart and Moore (1995), imagine that a firm's managers are "empire builders" who have a preference for investing independently of the investment's net present value when this net present value is observable but not verifiable. If the firm has available cash flows, it will always invest; if instead it has to make a debt repayment, it will be cash-constrained, and banks will finance only the profitable investment projects. The debt structure will be useful because it restricts the freedom of the managers to choose investments (see problem 4.7.4).

This section studies three examples inspired by Hart and Moore (1994), Myers and Rajan (1998) and Dewatripont and Maskin (1995). It builds on the idea of noncommitment of one of the contractual parties: the entrepreneur, the manager, or the originating bank.

### 4.5.1 Private Debtors and the Inalienability of Human Capital

Hart and Moore (1994) stress as the main characteristic of a debt contract the fact that it cannot impose on the entrepreneur any restriction on the freedom to walk away. This noncommitment for the entrepreneur not to withdraw human capital from the investment project will imply that (1) some profitable projects will not be funded, and (2) the time profile of repayments will be affected by the liquidation value of the project.

To see this, consider a risk-neutral entrepreneur who wants to invest an amount $I$ in a project that yields a certain stream of (discounted) cash flows $y_t$, $t = 1, \ldots, T$. If the entrepreneur is not cash-constrained, she will invest if and only if

$$I \leq \sum y_t,$$

where $\sum y_t$ is the present value of the project (the riskless rate of interest is normalized to zero). The case that will be focused on is the one in which there is a cash constraint and the project has to be funded by a debt contract. If additional subsequent loans are ruled out, the time profile for the repayment $R_t$ $(t = 1, \ldots, T)$ is constrained by the limited liability clause:

$$0 \leq R_t \leq y_t \qquad (t = 1, \ldots, T).$$

Since the entrepreneur cannot commit not to walk away from the project, she will use this possibility in a strategic way. Therefore, the situation must be modeled as a game. At any time the debtor can threaten to end the contract, possibly incurring an opportunity cost, since future cash flows will be lost or reduced. If this threat is credible, then the creditor and the debtor will enter into a bargaining game in which the creditor will obtain at least the liquidation value of the investment project but may obtain more if the project is not liquidated.

A crucial element is how the bargaining game is solved, that is, how the bargaining power is allocated between the two parties (players).[18] We examine the two extreme cases in which all the bargaining power belongs to either the creditor or the debtor. Denote by $V_t$ the value of the project to the creditor if the entrepreneur quits. $V_t$ may represent collateral, the liquidation value of the assets, or the net present value of the project if a new manager has to implement it.

Consider first the case in which the bank has all the bargaining power. This implies that it obtains $\sum_{\tau=t}^{T} R_\tau$ if debt is repudiated at time $t$. Now any contract $(R_1, \dots, R_T$ satisfying the limited liability constraint, $0 \le R_t \le y_t$ for all $t)$ will be repudiation-proof. Among such contracts, the one with maximum net present value of repayments is clearly $R_t = y_t$ for all $t$. Therefore, the maximum amount of debt that a project can raise will be $\sum_{\tau=1}^{T} y_\tau$, and a project will be funded if and only if its net present value is non-negative. In this case, there are no efficiency losses due to incomplete contracts.

Next, assume that the entrepreneur has all the bargaining power. Then, a debt repayment scheme will be repudiation-proof if

$$\sum_{\tau=t}^{T} R_\tau \le V_t \qquad (t = 1, \dots, T).$$

Indeed, in this case repudiation cannot improve the debtor's position because she will never gain from entering the bargaining process. The project will be undertaken only if there is a loan contract such that the present value of repayments exceeds the volume $L$ of the loan, and if the entrepreneur has enough wealth to finance the investment:

$$L \le \sum_{t=1}^{T} R_t \quad \text{and} \quad A + L \ge I,$$

where $R_t$ is the repudiation-proof repayment scheme, and $A$ the wealth of the entrepreneur. Clearly, some profitable investment projects will not be financed because the nonappropriability of human capital reduces the amount of *credible* repayment flows.

As intuition suggests, between these two extreme assumptions on the allocation of bargaining power, repudiation-proof repayment schemes will be obtained that will imply some inefficiency for the credit allocation. More generally,

**Result 4.5** The noncommitment of human capital to a project may generate some inefficiencies because it may limit the borrowing capacity of a firm strictly below the net present value of its future cash flow.

This inefficiency result appears only because of renegotiation possibilities. The model shows why banks may be concerned not only with the net present value of cash flows but also with the projects' collateral. In addition, as mentioned by Hart and Moore (1994, 842), it corresponds to the advice practitioners often give to "lend long if the loan is supported by durable collateral" and to "match assets with liabilities."

### 4.5.2 Liquidity of Assets and Debt Capacity

Myers and Rajan (1998) point out that if firms with more liquid assets (that cannot be pledged as collateral) have more opportunities for asset substitution, this may have a negative impact on their debt capacity.

Their model combines the threat of asset substitution on behalf of the manager (akin to the threat to abandon the project in Hart and Moore 1994) and the threat of early termination (Bolton and Sharfstein 1990), an aspect we disregard here for simplicity.

Consider a manager and an investor in a risk-neutral, zero-interest-rate setting. The manager's project returns $C + d$ at time $t = 1$. $C$ is in cash and $d$ represents the continuation value of assets in place. The cash flow $C$ is certain, observable, and verifiable. The investment is financed by a loan, provided at $t = 0$ by the financier.

The assets can be liquidated for a value $\alpha d$ (with $0 < \alpha < 1$); $\alpha$ is a measure of the liquidity of the assets in place. If the manager does not repay the loan, the maximum value that investors can obtain is $\alpha d$. Assuming the manager has all the bargaining power, this implies that debt repayment $R$ has to satisfy the constraint

$$R \leq \alpha d. \tag{4.10}$$

The manager's second option is to proceed to asset substitution. This could take many forms, from stealing the asset to transforming it into a specific asset with little value in the absence of the manager or transforming it into a risky asset. In that case, investors have a zero return and managers get a return $\alpha_M d$. $\alpha_M$ is also related to the liquidity of the asset. For simplicity, we assume $\alpha_M = \alpha$.

Therefore, the manager will not proceed to asset substitution if

$$C + d - R \geq \alpha d. \tag{4.11}$$

A debt contract with repayment $R$ is feasible if it satisfies both (4.10) and (4.11), that is,

$$R \geq \min(\alpha d, C + d - \alpha d).$$

When we consider different degrees of liquidity, we observe that contrary to the intuition derived from Hart and Moore (1994) and constraint (4.10), liquidity is a two-edged sword.

Indeed, the maximum debt capacity is a nonmonotonic function of the liquidity $\alpha$, with a maximum for

$$\alpha = \frac{C + d}{2d}.$$

A more liquid project will not be able to attract more funding because of the threat of asset substitution. This has important consequences for mergers and financial intermediation.

Define as *illiquid firms* the firms whose debt capacity is limited by constraint (4.10) and as *overly liquid firms* the ones for which constraint (4.11) is binding.

The merger of two firms of the same type will never increase their total debt capacity because of the linearity of the constraint. However, the merger of an illiquid and an overly liquid firm allows increasing their total debt capacity. (Notice that this argument is completely different from the standard diversification justification for a merger that stems from bankruptcy costs.)

The interest of financial intermediation comes from the fact that adding an illiquid activity to an overly liquid firm may increase the debt capacity of the firm. As a consequence, an overly liquid firm as a deposit-taking bank is interested in making loans that are illiquid. This may explain the comparative advantage of banks over nonbanks and the fact that bank loans are illiquid assets.

### 4.5.3 Soft Budget Constraints and Financial Structure

Dewatripont and Maskin (1995) model the problem of soft budget constraints that arises when banks feel obliged to refinance their corporate borrowers under the threat of inefficient liquidation. This phenomenon has been well documented in transition economies where bad loans to formerly state-owned firms have piled up in the banks' balance sheets.

This illustrates a fundamental trade-off between bank finance and market finance. Market investors are typically too much oriented toward short-term returns (which may lead to inefficient liquidations), whereas banks may be inclined to renegotiate their loans too often (thus leading, from an ex ante perspective, to an inefficient choice of projects). This can be related to the banks' initial investment in relationship lending (see section 3.6).

Several features of the model are worth noticing. First, the bank may be willing to continue investing in a project because the previous loans are seen as sunk costs at a given point in time, and therefore the continuation of the project may be the best option. Second, Dewatripont and Maskin explicitly assume that monitoring by the bank adds value to the project. Finally, they assume that this monitoring is not observable, so if the project is sold to a different financier, the original financier, realizing it will not receive the full benefits of its effort, will choose a lower effort level, and the external financier will take this into account.

Consider a population of entrepreneurs perfectly informed about their types $G$ (good) and $B$ (bad), and banks facing an adverse selection problem on the entrepreneurs' types. A $G$ type entrepreneur implements a project yielding $y_G$ with certainty. A $B$ type entrepreneur will require an additional unit of funds to obtain a random return $\tilde{y}$. This return is a lottery between 0 and $\bar{y}$ with probability of success $e$. This probability depends only on the effort $e$ exerted by the original financier. This effort has a cost $\Psi(e)$, where $\Psi(\cdot)$ is increasing and convex.

The optimal level of effort $e^*$ maximizes net expected return:

$$\pi(e) = e \cdot \bar{y} - \Psi(e).$$

It is characterized by the first-order condition

$$\bar{y} = \Psi'(e^*). \tag{4.12}$$

This happens to be also the bank's profits-maximizing level because we assume that in case of bankruptcy the bank appropriates the total surplus of the project. Therefore, it is equivalent to maximizing the total surplus or to maximizing the bank's profit.

We consider two alternative financial structures: a centralized banking structure and a decentralized financial market structure. The difference between the two is that banks have easy access to funds, whereas financial investors can only raise funds by pledging a fraction of the asset they hold to a third party. This apparently slight difference will have important implications for the level of effort, which will now differ from $e^*$.

**Bank Finance**

We assume that the bank learns the firm's type after financing it. The efficient ex post solution for a bank facing a $B$ project will be to continue if the profits from continuation investing an additional unit, $\pi(e^*) - 2$, are larger than the loss due to termination $(-1)$; that is, continuation will take place if $\pi(e^*) > 1$.

Dewatripont and Maskin make the following assumption:

**Condition DM** Profitable continuation of ex ante unprofitable projects:

$$2 > \pi(e^*) > 1.$$

This means that when the efficient level of effort $e^*$ is exerted, it is efficient to refinance bad projects even if it would not be profitable to finance them at their start.

**Market Finance**

Under decentralized market finance, the financier has only one unit of capital, so she will have to sell, pledge, or securitize the loan at time $t = 1$. If the project is to be refinanced, a new, external financier will have to be repaid some amount out of the project's time $t = 2$ cash flow in case of success. Denote this amount by $R_x$, and by $\hat{e}$ the level of effort exerted by the initial financier under decentralized market finance (knowing that she will have to sell claims on her initial investment). Since we are assuming that default at time $t = 1$ leads to the full appropriation of the firm's cash flows in case of success, the initial financier (or originator) will be paid $\bar{y} - R_x$ and the external financier will be paid $R_x$.

The external financier will only agree to finance the project if his individual rationality constraint is satisfied, that is, if $\hat{e}R_x \geq 1$.

Since the effort level is determined by the initial financier at time $t = 0$, under decentralized finance it will be the first investor who will set the level of effort $\hat{e}$. Consequently, $\hat{e}$ results from the maximization of $\pi(e) = e(\bar{y} - R_x) - \Psi(e)$, so the marginal return from effort is now lower, and it is optimal to set $\hat{e} < e^*$. The reason is simply that under decentralized finance, the benefits of a higher effort are partly appropriated by the external financier.

Assume early termination has a cost for the entrepreneur, for instance, a reputational cost, so that early termination will deter bad entrepreneurs from developing the project because they expect no monetary profit.

Depending on the level of $\pi(\hat{e})$ two cases may occur:

- If $2 > \pi(\hat{e}) > 1$, both centralization and decentralization lead to the financing of all projects, but decentralization generates a lower level of effort, thus provoking an inefficiency by increasing the riskiness of the $B$ project.
- If $\pi(\hat{e}) < 1$, then under decentralized finance $B$ entrepreneurs know they will have to face termination at time $t = 1$, since the project is not generating a sufficient expected cash flow, given the level of effort $\hat{e}$ chosen by the bank. In this case, decentralized market finance leads to funding only $G$ firms and therefore to the efficient outcome. By contrast, centralized finance leads to the continuation of $B$ projects, which, in turn, provides incentives for the $B$ entrepreneurs to apply for a loan.

Thus, when condition DM is satisfied, so that it is profitable to continue ex ante unprofitable projects, the decentralization of credit may promote efficient project selection, whereas centralization may lead to continuation of inefficient projects. This conclusion can be reversed when $\pi(e^*) > 2$, that is, when bad entrepreneurs are still worth financing.

This is because it is now efficient to finance bad projects at time $t = 0$, exerting a level of effort $\hat{e}$ is always inferior, and for some values of the parameters only $G$ projects will be financed, which is inefficient. Consequently, centralized banking finance is efficient while decentralized market finance is not. What is particularly striking is that the very same forces are at work: decentralized market finance leads to a lower level of effort on the part of the financier, and that is why the same projects are profitable under bank finance and not profitable under market finance.

The model gives interesting insights on two apparently unrelated issues. When condition DM is satisfied, it explains the phenomenon of transition economies where firms face a soft budget constraint because of the absence of realistic bankruptcy threats while banks continued to accumulate impaired loans. The other side of the coin occurs when condition DM does not hold. In this case, the close links between firms and banks increase efficiency.

## 4.6 Collateral as a Device for Screening Heterogeneous Borrowers

This section assumes the existence of different categories of borrowers, represented by a risk parameter $\theta$. If this parameter $\theta$ were common knowledge, the optimal contract (corresponding to a given individual rationality level $U_L^0$ for the lender) would be obtained, as in section 4.1, by solving program $\mathscr{P}_0$ for each value of $\theta$. In the particular case of exponential utilities (see note 1), the optimal equity participation coefficient $\alpha$ would be constant, but the repayment $R$ would be higher for higher risks.

In fact, it is often more realistic to assume that $\theta$ is observed only by the borrower, in which case the previous contract (with the interest rate conditional on $\theta$) cannot be implemented. Unless other considerations are introduced, all borrowers would claim to be in the lowest risk category in order to pay the minimum interest rate. As a consequence, the lender would be bound to disregard the declaration of the borrower and to charge a uniform interest rate.

This section examines how some flexibility can be reintroduced by offering to the population of borrowers a whole menu of contracts with different provisions.[19] For example, the lender can offer different loan contracts with variable collateral requirements (as in Bester 1985), the interest rate being a decreasing function of the collateral. Another possibility is to offer different loans of variable sizes (as in Freixas and Laffont 1990), the interest rate being an increasing function of the size of the loan. More complex menus can also be offered (as in Besanko and Thakor 1987), and the menus may specify how the terms of the contract depend upon observed variables (Webb 1991). But this discussion concentrates on the first example (Bester 1985) because it clearly illustrates how the lender can obtain a self-selection of heterogeneous borrowers.

We consider the case of binomial risks, where an investment (of a given size) can either fail ($\tilde{y} = 0$) or succeed ($\tilde{y} = y$). The risk parameter $\theta$ represents the probability of failure. Therefore a higher $\theta$ means an increasing risk in the sense of first-order stochastic dominance. For simplicity, the example assumes that there are only two categories of borrowers: low risks $\theta^L$ and high risks $\theta^H$ (with $\theta^L < \theta^H$). The proportions $v^k$ $(k = L, H)$ of borrowers of each type are common knowledge. All agents are risk-neutral.

Assume that borrowers can initially put down some collateral $C$. The lender can thus offer a menu of loan contracts $\{(C^k, R^k), k = L, H\}$, where the repayment $R^k$ in case of success depends on the collateral $C^k$ put down by the borrower. If the project fails ($\tilde{y} = 0$), the lender can liquidate this collateral; the borrower loses $C^k$, whereas the lender gets only $\delta C^k$ (with $\delta < 1$). Thus there is a cost of liquidation,[20]

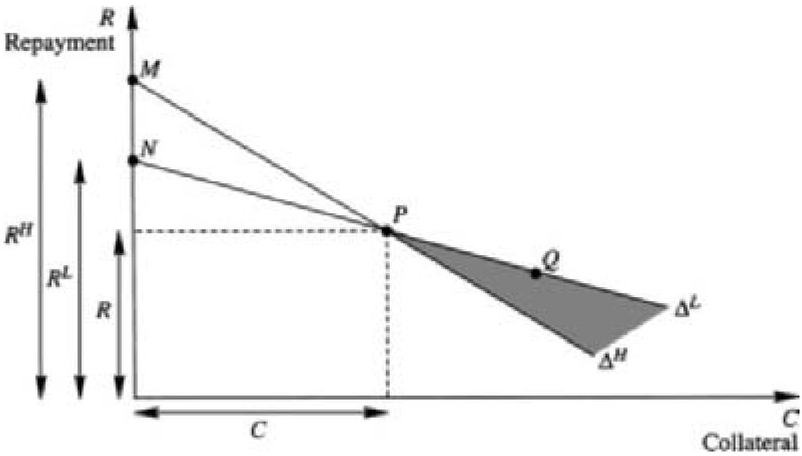



**Figure 4.5**
Borrowers' indifference curves: low risks $\Delta^L$, high risks $\Delta^H$.

$(1-\delta)C^k$, which is assumed to be proportional to the size of the collateral. If, on the other hand, the project succeeds ($\tilde{y} = y$), there is no liquidation; the lender obtains $R^k$ and the borrower gets $(y - R^k)$.

The menu of contracts offered by the lender will depend on the outside opportunities of borrowers (represented by their reservation utilities $U^k$, $k = L, H$) and on the relative bargaining power of the two parties. This example assumes that all bargaining power is concentrated in the hands of the lender. For example, in the benchmark case of symmetric information (when the lender is able to observe $\theta$), the lender will offer contracts such that the individual rationality constraints of each type of borrowers are binding:

$$(1-\theta^k)(y - R^k) - \theta^k C^k = U^k \qquad (k = L, H).$$

The corresponding indifference curves in the $(C, R)$ plane, denoted $\Delta^k$ $(k = L, H)$, are shown in figure 4.5.

The inequality $\theta^H > \theta^L$ implies that $\Delta^H$ is steeper than $\Delta^L$, assuming that the intersection of the two indifference curves $P$ lies in the positive quadrant,[21] which means that

$$\frac{U^L}{1-\theta^L} \geq \frac{U^H}{1-\theta^H}.$$

Since liquidation is costly, the contracts preferred by the lender on each of these lines are, respectively, $M$ and $N$, which both correspond to the absence of collateral $(C = 0)$.

Of course, if $\theta$ is not observable by the lender, and if contracts do not differ in their collateral ($C = 0$), both types of borrowers will claim to be low risks and choose contract $N$. The average expected return to the lender will be $(1 - \bar{\theta})R^L$, where $R^L$ is the maximum repayment that is acceptable to type $L$ borrowers,[22]

$$R^L = y - \frac{U^L}{1 - \theta^L},$$

and $\bar{\theta}$ denotes the average probability of failure in the population of borrowers,

$$\bar{\theta} \stackrel{\text{def}}{=} \nu^L \theta^L + \nu^H \theta^H.$$

In this situation, high risks obtain an informational rent because their expected utility is higher than what they would get if low risks were absent, in which case they would have to repay the higher amount,

$$R^H = y - \frac{U^H}{1 - \theta^H}.$$

A lender who wants high risks to repay $R^H$ must offer simultaneously another contract, designed specifically for low risks and requiring a collateral $C$ and a repayment $R$ such that

$$(1 - \theta^H)(y - R^H) \geq (1 - \theta^H)(y - R) - \theta^H C$$

(high risks prefer contract $M$ to the new contract $P = (C, R)$) and

$$(1 - \theta^L)(y - R) - \theta^L C \geq U^L$$

(low risks accept this new contract).[23] The set of contracts satisfying these two conditions is represented by the shaded area in figure 4.5.

Since collateral is costly, it would be inefficient to design a contract in which both types are required to pledge some collateral. Indeed, the only role of collateral is to allow for self-selection between the two types of risks. Intuitively, the choice between the two types of contracts depends on what the agent would answer to the question, Do you want to bet a collateral $C$ that you will not fail, against a reduction in interest rates? Only low-risk borrowers will take that bet. Figure 4.5 uses indifference curves to examine borrowers' choice between the two contracts (that is, if they would take the bet) for any two points. A menu of (two) contracts will allow discrimination between the two types of borrowers if each of them chooses the contract that he prefers (that is closer to the origin). Thus, for instance, the menu $(M, Q)$ is a discriminating one. Clearly, it is not efficient because low-risk borrowers are offering too much (costly) collateral. Starting from this point, the lender's profit may be increased under these constraints by offering $(M, P)$ where $P$ is the intersection of $\Delta^L$ and $\Delta^H$

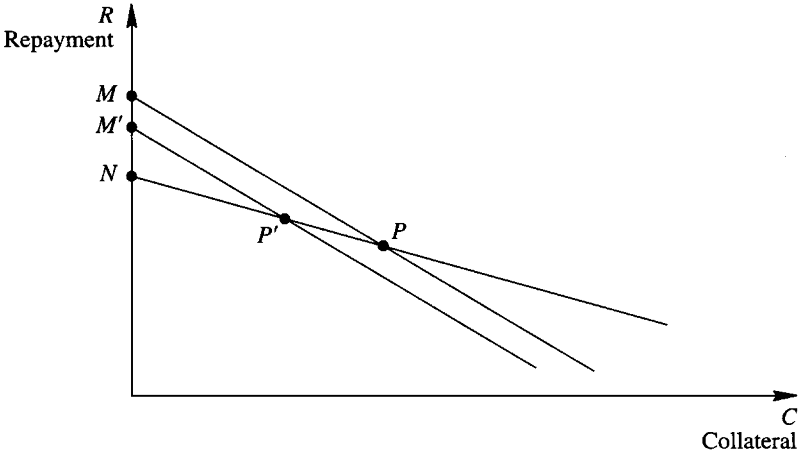



**Figure 4.6**
Optimal menu of loan contracts.

in figure 4.5. To improve on the menu of contracts $(M, P)$, consider the menu $(M', P')$ in figure 4.6.

By accepting that high-risk borrowers receive an informational rent (they repay $M'$ instead of $M$), which implies a loss for the lender, the amount of collateral decreases, and this implies a gain for the lender. The optimal set of contracts will thus be obtained at a pair such as pair $(M', P')$ in figure 4.6, where $P'$ lies between $N$ and $P$. The exact location of $P'$ and $M'$ will be determined by the proportions $\nu^H$ and $\nu^L$. In particular, when $\nu^L$ tends toward 1, the single contract $N$ of figure 4.6 will be offered to both types of borrowers, and when $\nu^L$ tends toward 0, the two offered contracts will be $(M, P)$.

**Result 4.6** The optimal menu of loan contracts combining repayment and collateral is such that

- high risks pay a high interest rate but are not required to put down any collateral (no distortion at the top; high-risk contracts require the efficient level of collateral);
- low risks have to put down some collateral but pay a lower interest rate.

The design of self-selection mechanisms to improve credit allocation in an asymmetric information setting has been widely studied. For instance, Webb (1992) considers an environment in which lenders sequentially invest in two projects, so that truthful reporting of cash flows and costly auditing (depending on the reported cash flows) is necessary in every period. Then it is possible to make the terms of borrowing during the second period depend on the reported first-period cash flows. By so doing, auditing can be reduced, and this implies that long-term lending may dominate short-term lending because the expected cost of auditing is lower.

Another interesting application of self-selection mechanisms is the study of securitization. One of the main characteristics of securitization is that it is associated with credit enhancement (see Greenbaum and Thakor 1987 and problem 4.7.6). In this way it is possible to save on screening costs. The deposit funding mode implies higher screening costs and a lower level of risk, since in case of failure the investors will share the full amount of the bank's capital. Consequently, the choice between the deposit funding mode and securitization will depend on screening costs and on the investors' risk aversion.

## 4.7 Problems

### 4.7.1 Optimal Risk Sharing with Symmetric Information

Using the notation of section 4.1, optimal debt contracts can be obtained by solving the following program:

$$\begin{cases} \max\limits_{R(\cdot)} \; Eu_B(\tilde{y} - R(\tilde{y})) \\ \text{under } Eu_L(R(\tilde{y})) \geq U_L^0. \end{cases}$$

1. If $\mu$ denotes the Lagrange multiplier associated with the individual rationality constraint, show that for all $y$ in the support of $\tilde{y}$, $R(y)$ can be obtained by maximizing $u_B(y - R) + \mu u_L(R)$ with respect to $R$.

2. Prove condition (4.3):

$$\forall \mu \in \text{Supp } \tilde{y} : \frac{u_B'(y - R(y))}{u_L'(R(y))} = \mu.$$

3. Prove condition (4.2):

$$\forall y_1, y_2 \in \text{Supp } \tilde{y} : \frac{u_L'[R(y_1)]}{u_L'[R(y_2)]} = \frac{u_B'[y_1 - R(y_1)]}{u_B'[y_2 - R(y_2)]}.$$

4. When limited liability constraints are introduced ($0 \leq R(y) \leq y$), show that the characterization becomes

$$R(y) = \begin{cases} 0 & \text{if } u_B'(y) \geq \mu u_L'(0), \\ y & \text{if } u_B'(0) \leq \mu u_L'(y), \end{cases}$$

and

$$\frac{u_B'(y - R(y))}{u_L'(R(y))} = \mu$$

in the other cases.

### 4.7.2 Optimal Debt Contracts with Moral Hazard

This problem is adapted from Innes (1990). Recall the notation of section 4.4:

- $f(y, e)$ denotes the density of $\tilde{y}$ when the effort level is $e$.
- $\psi(e)$ represents (the pecuniary equivalent of) the cost of effort for the borrower.
- $V(R, e)$ is the expected utility of the risk-neutral borrower as a function of the repayment schedule $R(\cdot)$ and the effort level $e$:

$$V(R, e) = \int (y - R(y)) f(y, e)\, dy - \psi(e).$$

Given $U_L^0$, the individual rationality level of the lender, the second-best optimal contract $R(\cdot)$ and effort level $e$ will be obtained as the solution of

$$\mathscr{P} \begin{cases} \max\limits_{(R,e)} \; V(R, e) \\ 0 \leq R(y) \leq y \qquad \forall y, \\ V(R, e') \leq V(R, e) \qquad \forall e', \\ \int R(y) f(y, e)\, dy \geq U_L^0. \end{cases}$$

A simpler program is obtained by replacing the incentive compatibility constraint with the first-order condition of the borrower's problem, which determines the borrower's effort choice:

$$V_e(R, e) = \int (y - R(y)) f_e(y, e)\, dy - \psi'(e) = 0.$$

When $V$ is concave in $e$, it is legitimate to use this first-order approach (as proved in Rogerson 1985). The simpler program is equivalent to $\mathscr{P}$, and the optimal contract $R(\cdot)$ can be obtained by maximizing the Lagrangian for each $y$:

$$\max_{0 \leq R(y) \leq y} \mathscr{L}(R(y), y) = [y - R(y)][f(y, e) + \mu f_e(y, e)] + \lambda R(y) f(y, e),$$

where $\mu$ and $\lambda$ denote, respectively, the Lagrange multipliers associated with the first-order condition of the borrower's problem and with the individual rationality constraint of the lender.

1. Show that the optimal contract is such that

$$R(y) = \begin{cases} y & \text{if } (\lambda - 1)f(y,e) > \mu f_e(y,e), \\ 0 & \text{if } (\lambda - 1)f(y,e) < \mu f_e(y,e). \end{cases}$$

2. Show that the monotone likelihood property, which implies that for all $e_1 > e_2$, the function

$$\frac{f(y,e_1)}{f(y,e_2)}$$

is increasing in $y$, implies that

$$y \to \frac{f_e(y,e)}{f(y,e)}$$

is also increasing.

3. Assuming that $\mu > 0$, show that the optimal contract involves a cutoff level $y^*$, with a maximum penalty, $R(y) = y$, when $y$ is less than $y^*$, and a maximum reward, $R(y) = 0$, when $y$ is greater than $y^*$.

### 4.7.3 The Optimality of Stochastic Auditing Schemes

Consider a simple version of the Townsend model (lender-borrower relationship with costly state verification) in which the cash flow $\tilde{y}$ obtained by the borrower during the second period can take only two values: a high value $y_H$ (with probability $p_H$) and a low value $y_L$ (with probability $p_L = 1 - p_H$). The volume of the loan is denoted $I$. The lender and the borrower are risk-neutral. The optimal contracts are found by maximizing the expected repayment to the lender $U_L^0$ (net of auditing costs) under incentive compatibility and individual rationality constraints for the borrower. The status quo utility level of the borrower is denoted $U_B^0$, and the audit cost $\gamma$. Finally, the borrower has limited liability. The maximum penalty that can be inflicted on her if she lies (reports $y_L$ when $y_H$ has occurred) is confiscation of $y_H$.

1. Compute the optimal contract as a function of $U_B^0$. Represent the Pareto frontier in the $(U_B^0, U_L^0)$ plane.

2. Suppose that the lender can credibly commit to a stochastic auditing policy: audit with probability $q \in [0,1]$ when the borrower reports $y_L$. Show that the incentive compatibility constraint is equivalent to

$$q \geq q^* = 1 - \frac{U_B^0}{p_H(y_H - y_L)}.$$

3. Represent the new Pareto frontier. What do you conclude?

### 4.7.4 The Role of Hard Claims in Constraining Management

This problem is adapted from Hart and Moore (1995). Consider a firm whose managers are empire builders in the sense that they always choose to implement investment projects provided they are not cash-constrained. The objective is to show that the firm's indebtedness will help discipline their behavior.

Assume that all agents are risk-neutral, and normalize the interest rate to zero. At $t = 0$, the firm has to finance an investment $I$ that returns deterministic cash flows $y_1$ at $t = 1$, and $y_2$ at $t = 2$. This investment has a positive net present value, $y_1 + y_2 - I > 0$. The firm issues a volume of debt $D \geq I$ in exchange for repayments $D_1$ at $t = 1$, and $D_2$ at $t = 2$. The difference $D - I$ is paid to shareholders in the form of a dividend. At $t = 1$, the firm will have a new, random investment opportunity: by investing $\tilde{I}_1$ it will obtain $\tilde{R}_2$ at $t = 2$. Both $\tilde{I}_1$ and $\tilde{R}_2$ are unknown at $t = 0$ but will be perfectly observable at $t = 1$. Since the managers are empire builders, they will seize the new investment opportunity provided they have the funds, either by self-finance or by borrowing from a bank. The bank loan, if any, is junior to the debt issued at $t = 0$.

1. Show that the firm is able to invest at $t = 1$ in two cases:

$$y_1 \geq D_1 + \tilde{I}_1,$$

or

$$y_1 < D_1 + \tilde{I}_1 \quad \text{and} \quad \tilde{R}_2 + y_2 - D_2 \geq \tilde{I}_1 - y_1 + D_1.$$

2. Show that shareholder value at $t = 0$ is maximized by setting $D_1 = y_1$, and $D_2 = y_2$.

3. Explain the role of debt in this model.

### 4.7.5 Collateral and Rationing

This problem is adapted from Besanko and Thakor (1987).

1. With the notation used in section 4.6, assume $U^L = U^H = U$, and compute the monopoly solution under full information.

2. Compute the monopoly solution under adverse selection. Show that for a low proportion of high-risk borrowers, $\nu^H$, the contract will be designed to attract only low risks. The riskless rate is normalized to zero.

3. Consider a competitive setting in which low-risk borrowers have only a wealth level $W$ to be posted as collateral, and the contracts $(R^H, 0)$ and $(R^L, W)$, which yield a zero profit for the bank, are such that both borrowers prefer $(R^L, W)$. Stochastic mechanisms are allowed. Show that self-selection is possible only if low-risk borrowers can be rationed with a positive probability.

### 4.7.6 Securitization

This problem is adapted from Greenbaum and Thakor (1987). When a firm wants to securitize some of its assets, it typically signs a credit enhancement contract with a bank. In such a contract, the banks promises to insure a fraction $\theta$ of the repayment $R(\theta)$ promised by the firm to the investors who buy the security, in exchange for a fee $Q(\theta)$. This exercise shows how credit enhancement can be used to allow for a self-selection of firms, with better risks buying more credit enhancement.

Consider an economy in which risk-neutral firms have an investment project with a return $X$ in case of success, which occurs with probability $p$, and a zero return in case of failure (probability $1 - p$). The probability $p$ is known to the firms but not to the investors.

Banks offer credit insurance contracts characterized by different levels of credit enhancement $\theta$, where $\theta$ is the fraction of the initially promised repayment $R(\theta)$ that the investor will receive if the firm's project fails. A credit insurance contract will specify the fee paid by the firm $Q(\theta)$ corresponding to the level of $\theta$ and the repayment $R(\theta)$ promised to the investor. Consequently, a contract will define a mechanism $(R(\theta(\hat{p})), Q(\theta(\hat{p})))$ associating it to final investors.

1. Write the first- and second-order conditions that are necessary for the contract to be incentive-compatible.
2. Write the individual rationality (IR) constraint of the bank.
3. Assume the IR constraint holds with equality. By differentiating it, show that the mechanism is such that better risks tend to buy more credit enhancement, and the repayment $R$ decreases with the guarantee $\theta$.

## 4.8 Solutions

### 4.8.1 Optimal Risk Sharing with Symmetric Information

1. The Lagrangian of the problem is simply

$$\mathscr{L} = Eu_B(\tilde{y} - R(\tilde{y})) + \mu(Eu_L(R(\tilde{y})) - U_L^0).$$

The maximization with respect to $R(\cdot)$ can be performed separately for each value of $y$, leading to maximizing $u_B(y - R) + \mu u_L(R)$ with respect to $R$.

2. $R(y)$ is therefore defined implicitly by the first-order condition:

$$-u_B'(y - R(y)) + \mu u_L'(R(y)) = 0,$$

which gives condition (4.3).

3. Condition (4.2) is immediately deduced by applying (4.4) to $y$ and $y_2$.

4. When the constraint $0 \leq R(y) \leq y$ is added, the first-order condition changes only when $R(y) = 0$ or $y$, in which case it becomes

$$\begin{cases} -u'_B(y) + \mu u'_L(0) \leq 0 & \text{if } R(y) = 0, \\ -u'_B(0) + \mu u'_L(y) \geq 0 & \text{if } R(y) = y. \end{cases}$$

### 4.8.2 Optimal Debt Contracts with Moral Hazard

1. To obtain the optimal contract, it is necessary only to maximize the Lagrangian $\mathscr{L}$ with respect to $R(y)$ for given $y$. But $\mathscr{L}$ is linear with respect to $R(y)$. Therefore, if the coefficient affecting $R(y)$ is positive, that is, if

$$(\lambda - 1) f(y, e) - u f_e(y, e) > 0,$$

then $\mathscr{L}$ is increasing and the maximum is obtained for $R(y) = y$. Conversely, if the coefficient is negative, $R(y) = 0$ is obtained.

2. Since

$$\frac{f(y, e_1)}{f(y, e_2)}$$

is increasing in $y$, so is the function

$$\frac{1}{e_1 - e_2}\left(\frac{f(y, e_1)}{f(y, e_2)} - 1\right).$$

When $e_1$ tends toward $e_2$, the limit of this function will also be increasing in $y$. But this limit is

$$\frac{f_e(y, e_2)}{f(y, e_2)}.$$

3. The result obtained in part 1 shows that the optimal function $R(y)$ is characterized by

$$R(y) = \begin{cases} y & \text{for } \dfrac{f_e(y, e)}{f(y, e)} < \dfrac{\lambda - 1}{\mu}, \\ 0 & \text{for } \dfrac{f_e(y, e)}{f(y, e)} > \dfrac{\lambda - 1}{\mu}. \end{cases}$$

Since $f_e/f$ is continuous and increasing in $y$, there is a unique $y^*$ such that

$$\frac{f_e(y^*, e)}{f(y^*, e)} \equiv \frac{\lambda - 1}{\mu}$$

and therefore the solution will be given by

$$R(y) = \begin{cases} y & y < y^*, \\ 0 & y > y^*. \end{cases}$$

### 4.8.3 The Optimality of Stochastic Auditing Schemes

1. The optimal contract is a standard debt contract with nominal repayment $R$. Two cases are possible:

- When $R \leq y_L$, debt is riskless and there is never any audit. Expected utilities are

$$U_L^0 = R - I; \qquad U_B^0 = p_H y_H + p_L y_L - R.$$

- When $y_L < R \leq y_H$, $R_H, R, R_L = y_L$ and audit takes place in state $L$. Expected utilities are

$$U_L^0 = p_H R + p_L(y_L - \gamma) - I; \qquad U_B^0 = p_H(y_H - R).$$

Taking $R > y_H$ is inefficient because audit would also take place in state $H$, which is clearly dominated by the debt contract with $R = y_H$. The Pareto frontier thus has two parts:

$$U_L^0 = \begin{cases} (p_H y_H + p_L y_L - I) - U_B^0 & \text{when } U_B^0 \geq p_H(y_H - y_L), \\ (p_H y_H + p_L y_L - I) - U_B^0 - p_L \gamma & \text{when } U_B^0 < p_H(y_H - y_L). \end{cases}$$

Note that this Pareto frontier is discontinuous at $U_B^0 = p_H(y_H - y_L)$.

2. With stochastic auditing in state $L$, the incentive compatibility constraint is satisfied if

$$U_B^0 = p_H(y_H - R) \geq p_H(1 - q)(y_H - y_L).$$

The Pareto-optimal contract is thus characterized by

$$q = 1 - \frac{y_H - R}{y_H - y_L} = \frac{R - y_L}{y_H - y_L}.$$

Expected utilities are

$$U_B^0 = p_H(1 - q)(y_H - y_L); \qquad U_L^0 = (p_H y_H + p_L y_L - I) - U_B^0 - p_L q \gamma.$$

3. The equation of the new Pareto frontier is obtained by eliminating $q$:

$$q = 1 - \frac{U_B^0}{p_H(y_H - y_L)}.$$

Thus

$$U_L^0 = p_H y_H + p_L y_L - I - U_B^0 - p_L \gamma + \frac{p_L \gamma U_B^0}{p_H(y_H - y_L)},$$

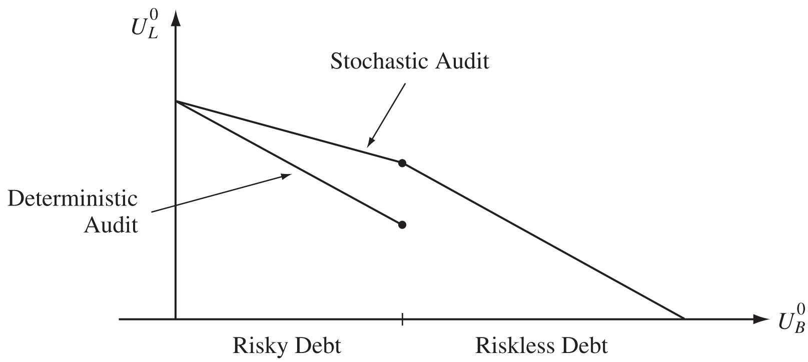



**Figure 4.7**
Pareto frontiers with deterministic and stochastic audits.

in the region where $q \geq 0$, that is, $U_B^0 < p_H(y_H - y_L)$, which corresponds to the region where debt is risky. Thus whenever debt is risky, stochastic auditing dominates deterministic auditing. The two Pareto frontiers (with deterministic and stochastic auditing) are shown in figure 4.7.

### 4.8.4 The Role of Hard Claims in Constraining Management

1. First, if $y_1 - D_1 \geq \tilde{I}_1$, the firm is able to self-finance the new investment. Second, if $y_1 - D_1 < \tilde{I}_1$, but the firm can borrow $(\tilde{I}_1 - y_1 + D_1)$ from the bank in exchange for a repayment of $y_2 - D_2 + \tilde{R}_2$; the bank accepts if

$$y_2 - D_2 + \tilde{R}_2 \geq \tilde{I}_1 - y_1 + D_1.$$

2. Shareholder value is maximized when the investment policy is optimal. When $D_1 = y_1$, the firm cannot self-finance the new investment. When on top of that $D_2 = y_2$, the bank will finance it if $\tilde{R}_2 \geq I_1$, which corresponds to the optimal investment policy.

3. In this model, debt disciplines managers by preventing them from using the cash holdings of the firm for financing inefficient investments, either by self-finance or by pledging future cash flows to a bank.

### 4.8.5 Collateral and Rationing

1. First compute the full-information solution. If each borrower is at its reservation level, then

$$(1 - \theta^L)(y - R^L) - \theta^L C^L = U,$$

$$(1 - \theta^H)(y - R^H) - \theta^H C^H = U.$$

The first best contract is denoted by $(\hat{R}^L, 0)$ and $(\hat{R}^H, 0)$. With adverse selection, the monopoly does not want to discriminate between the two types of borrowers using the level of collateral because starting from $(\hat{R}^H, 0)$ for the two types, any contract with a higher collateral for low risks would be separating but would yield the monopoly a lower return.

2. If there is adverse selection, the monopoly will prefer a contract $(\hat{R}_H, 0)$ designed to attract both types of borrowers over a contract $\hat{R}^L$ designed to attract only the low-risk ones, if

$$(1 - \bar{\theta})\hat{R}^H - 1 > \nu^L[(1 - \theta^L)\hat{R}^L - 1],$$

where $\bar{\theta} = \nu^L\theta^L + \nu^H\theta^H$. This is equivalent to

$$\nu^H(1 - \theta^H)(\hat{R}^H - 1) - \nu^L(1 - \theta^L)(\hat{R}^L - \hat{R}^H) > 0,$$

that is, the gain obtained on the $H$-type borrowers has to be higher than the opportunity cost of quoting a lower rate of interest $\hat{R}_H$. For a low $\nu^H$, this expression will be negative, and the monopoly will lend only to the low-risk borrowers at the high interest rate.

3. If we allow for stochastic mechanisms with probability $\pi^k$ of rationing borrowers of type $k = H, L$, the IC constraint for the high-risk borrowers becomes

$$(1 - \pi^H)(1 - \theta^H)(y - R^H) \geq (1 - \pi^L)\{(1 - \theta^H)(y - R^L) - \theta^H W\}.$$

But $W$ is too low to allow for separation, that is,

$$(1 - \theta^H)(y - R^H) < (1 - \theta^H)(y - R^L) - \theta^H W.$$

The first inequality can be obtained only if $1 - \pi^L < 1 - \pi^H$, that is, $\pi^L > \pi^H$. Thus, $\pi^L > 0$.

### 4.8.6 Securitization

1. The firm's objective function is

$$p[X - R(\theta(\hat{p}))] - Q(\theta(\hat{p})).$$

Consequently, maximization with respect to $\hat{p}$ yields at point $\hat{p} = p$ the following first-order condition:

$$(-pR'(\theta(p)) - Q'(\theta(p)))\frac{d\theta}{dp} = 0.$$

The contract is incentive-compatible.

The second-order condition is

$$-[pR''(\theta(p)) + Q''(\theta(p))]\frac{d^2\theta}{dp} \leq 0.$$

Since the first-order condition holds for every $p$, it may be differentiated and replaced in the second-order condition, which yields

$$-R'(\theta(p))\frac{d\theta}{dp} \geq 0.$$

2. The fee has to equal the expected value of the repayments made to the investor in case of failure, so that

$$Q(\theta(p)) = (1-p)\theta(p)R(\theta(p)),$$

the IR constraint of the bank.

3. Differentiation of IR implies

$$[Q'(\theta(p)) - (1-p)(R(\theta(p)) + \theta(p)R'(\theta(p)))]\frac{d\theta}{dp} = -\theta(p)R(\theta(p)).$$

Replacing the first-order conditions (incentive-compatible),

$$(1-p)R(\theta(p))\frac{d\theta}{dp} = \theta(p)R(\theta(p)) - [pR'(\theta(p) + (1-p)R'(\theta(p))\theta(p)]\frac{d\theta}{dp}.$$

Because of the second-order conditions, the right side of this equality is positive, implying

$$\frac{d\theta}{dp} > 0,$$

so better risks tend to buy more credit enhancement. Using this in the second-order condition yields the result $R'(\theta) < 0$, that is, repayments decrease with the guarantee $\theta$.

## Notes

1. When utilities are exponential ($u_i(x) = -e^{-\rho_i x}, i = B, L$), the absolute indexes of risk aversion are constant ($I_i \equiv \rho_i$). In this case, the optimal repayment function has the simple form $R(y) = \alpha y + \beta$ with

$$\alpha = \frac{\rho_B}{\rho_B + \rho_L},$$

which can be obtained as a combination of a standard debt contract (with nominal debt $R$) and an equity participation (the bank gets a fraction $\alpha$ of the shares). The total repayment is then

$R(y) = R + \alpha(y - R).$

However, this interpretation neglects the control rights associated with an equity participation.

2. Khalil and Parigi (1998) argue that the size of a loan is an important determinant of the incentive to audit in a costly state verification framework. When banks cannot credibly commit to auditing a defaulting loan, it can be useful for them to increase that loan's size and use this as a commitment device.

3. Lacker and Weinberg (1989) take an asymmetric cost function that they justify by arguing that only understating the true value of cash flows can be of interest. The slightly more general formulation discussed here allows proving that even if overstating the return is possible, it is never in the interest of the borrower.

4. This is the case at least when the limited liability constraint of the lender is not binding.

5. Another interesting article that addresses the issue of termination is Stiglitz and Weiss (1983), in which the possibility of renegotiation is introduced.

6. Of course, lenders can retaliate to some extent when they are backed by their home country. Thus they may be able to impose commercial sanctions on the defaulting country.

7. See Gale and Hellwig (1989) for a game-theoretic approach to this issue, and Eaton, Gersovitz, and Stiglitz (1986) for a general overview.

8. Nevertheless, this implies that if the country receives no loan, production is impossible, an assumption that is not needed in the model of Eaton and Gersovitz (1981), which is examined later.

9. Innocuous as it may seem, this is a restriction on the equilibrium concept used for solving the game.

10. In fact, we argue in chapter 5 that this should not be called credit rationing.

11. Nevertheless, as noted by Kletzer and Wright (2000), for instance, they implicitly assumed that the country will receive the contingent repayments of its "cash in advance," which would require external enforcement.

12. Regulation may even give incentives so that banks do not interfere with the choice of investment projects by the firms.

13. As is well known, in a principal-agent relationship in which the agent is risk-neutral and has unlimited liability, moral hazard problems become trivial and are solved by making the agent (borrower) pay a fixed amount to the principal (lender) and become a residual claimant in the project.

14. To understand the MLR property, consider the case of two effort levels $e_1$ and $e_2$, with $e_1 > e_2$, and suppose that output $y$ is observed, but not the effort level. Under Bayes' formula the (posterior) probability of a high effort level ($e_1$) conditional on $y$ is

$$P[e_1|y] = \frac{P(e_1)f(y|e_1)}{P(e_1)f(y|e_1) + P(e_2)f(y|e_2)},$$

or

$$P[e_1|y] = \frac{1}{1 + \frac{P(e_2)}{P(e_1)} \cdot \frac{f(y|e_2)}{f(y|e_1)}}.$$

Under the MLR property, this function increases with $y$: a higher return indicates a greater likelihood of high effort.

15. This is justified if the entrepreneur has access to an alternative source of (short-term) borrowing. Indeed, if $R(y)$ is not increasing (as in result 4.4), the entrepreneur with a return $y$ below the threshold $y^*$ could borrow $(y^* - y)$ and declare a result $y^*$. Thus the borrower would have nothing to repay except the short-term loan $y^* - y$.

16. An interesting synthesis of the incomplete contracts literature can be found in Tirole (1999) or in Bolton and Dewatripont (2005).

17. A seminal article on this question is Aghion and Bolton (1992).

18. The importance of the negotiation that follows default on debt structure is emphasized by Bolton and Scharfstein (1996), among others.

19. See Myerson (1979) for the general analysis of this class of problem.

20. Liquidation costs are important in explaining why loans are not always 100 percent collateralized.

21. The alternative case is studied in Besanko and Thakor (1987) and in problem 4.7.5.

22. It is assumed that lenders will be satisfied with this return. If they are not, and increase the lending rate, they could obtain $(1-\theta^H)R^H$, but this can be still worse and could result in no lending at all. This is related to the lemon problem of Akerlof (1970), and in the banking literature to the credit rationing article of Stiglitz and Weiss (1981) (see also chapter 5).

23. The other self-selection constraint,

$$(1-\theta^L)(y-R)-\theta^L C \geq (1-\theta^L)(y-R^H),$$

is trivially satisfied, since low risks obtain utility $U^L$ with the contract $(R^L, 0)$. Therefore $(R^H, 0)$ is not individually rational.

## References


Aghion, P., and P. Bolton. 1992. An incomplete contract approach to financial contracting. *Review of Economic Studies* 59: 473–494.

Akerlof, G. A. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84 (3): 488–500.

Allen, F. 1983. Credit rationing and payment incentives. *Review of Economic Studies* 50 (4): 639–646.

Alm, J., and J. R. Follain, Jr. 1982. Alternative mortgage instruments, the tilt problem and consumer welfare. *Journal of Financial and Quantitative Analysis* 19 (March): 113–126.

Berlin, M., and L. Mester. 1992. Debt covenants and renegotiation. *Journal of Financial Intermediation* 2: 95–135.

Besanko, D., and A. V. Thakor. 1987. Collateral and rationing: Sorting equilibria in monopolistic and competitive credit markets. *International Economic Review* 28: 671–690.

Bester, H. 1985. Screening vs rationing in credit markets with imperfect information. *American Economic Review* 75 (4): 850–855.

———. 1994. The role of collateral in a model of debt renegotiation. *Journal of Money, Credit and Banking* 26 (1): 72–86.

Bolton, P., and M. Dewatripont. 2005. *Contract theory*. Cambridge, Mass.: MIT Press.

Bolton, P., and D. Scharfstein. 1990. A theory of predation based on agency problems in financial contracting. *American Economic Review* 80 (1): 93–106.

———. 1996. Optimal debt structure and the number of creditors. *Journal of Political Economy* 104 (1): 1–25.

Boot, A., S. I. Greenbaum, and A. V. Thakor. 1993. Reputation and discretion in financial contracting. *American Economic Review* 83 (5): 1165–1183.

Boot, A., A. V. Thakor, and G. Udell. 1991. Credible commitments, contract enforcement problems and banks: Intermediation as credible insurance. *Journal of Banking and Finance* 15: 605–632.

Bulow, J., and K. Rogoff. 1989. Sovereign debt: Is to forgive to forget? *American Economic Review* 97 (1): 43–50.

Chang, C. 1990. The dynamic structure of optimal debt contracts. *Journal of Economic Theory* 52: 68–86.

Chemmanur, T. J., and P. Fulghieri. 1994. Reputation, renegotiation, and the choice between bank loans and publicly traded debt. *Review of Financial Studies* 7 (3): 475–506.

Dewatripont, M., and E. Maskin. 1995. Credit and efficiency in centralized and decentralized economies. *Review of Economic Studies* 62 (4): 541–556.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51: 393–414.

———. 1991. Monitoring and reputation: The choice between bank loans and directly placed debt. *Journal of Political Economy* 99: 689–721.

———. 1992. Seniority and maturity structure of debt contracts. *Journal of Financial Economics* 33: 341–368.

Eaton, J., and M. Gersovitz. 1981. Debt with potential repudiation: Theoretical and empirical analysis. *Review of Economic Studies* 48: 289–309.

Eaton, J., M. Gersovitz, and J. Stiglitz. 1986. On the pure theory of country risk. *European Economic Review* 30 (3): 481–514.

Elston, J. A. 1995. Banks, finance, and investment in Germany: A review article. *Small Business Economics* 7 (6): 475–479.

Freixas, X., and J. J. Laffont. 1990. Optimal banking contracts. In *Essays in honor of Edmond Malinvaud*. Vol. 2: *Macroeconomics*, ed. P. Champsaur et al. Cambridge, Mass.: MIT Press.

Fried, J., and P. Howitt. 1980. Credit rationing and implicit contract theory. *Journal of Money, Credit and Banking* 12: 471–487.

Fudenberg, D., and J. Tirole. 1991. *Game theory.* Cambridge, Mass.: MIT Press.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

———. 1989. Repudiation and renegotiation: The case of sovereign debt. *International Economic Review* 30 (1): 3–31.

Gorton, G., and J. Kahn. 1993. The design of bank loan contracts, collateral, and renegotiation. NBER working paper 4273.

Greenbaum, S. I., and A. V. Thakor. 1987. Bank funding modes: Securitization vs deposits. *Journal of Banking and Finance* 11: 379–401.

Gromb, D. 1994. Renegotiation in debt contracts. Ph.D. diss., Laboratoire d'Econométrie de l'Ecole Polytechnique, Paris.

Harris, M., and A. Raviv. 1991. The theory of capital structure. *Journal of Finance* 46 (1): 297–355.

———. 1992. Financial contracting theory. In *Advances in economic theory: Sixth world congress*, ed. J. J. Laffont, vol. 2, 64–150. Econometric Society Monograph 21. Cambridge: Cambridge University Press.

Hart, O. 1995. *Firms, contracts and financial structure.* Oxford: Oxford University Press.

Hart, O., and J. Moore. 1994. A theory of debt based on the inalienability of human capital. *Quarterly Journal of Economics* 109: 841–879.

———. 1995. Debt and seniority: An analysis of the role of hard claims in constraining management. *American Economics Review* 85: 567–585.

———. 1998. Default and renegotiation: A dynamic model of debt. *Quarterly Journal of Economics* 113 (1): 1–41.

Haubrich, J. 1989. Financial intermediation: Delegated monitoring and long-term relationships. *Journal of Banking and Finance* 13 (1): 9–20.

Holmström, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Hoshi, T., A. Kashyap, and D. Scharfstein. 1993. The choice between public and private debt: An analysis of post-deregulation corporate financing in Japan. NBER working paper 4421.

Innes, R. D. 1990. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory* 52 (1): 45–67.

Jappelli, T., P. Pagano, and M. Bianco. 2005. Courts and banks: Effects of judicial enforcement on credit markets. *Journal of Money, Credit and Banking* 37 (2): 223–244.

Jensen, M., and N. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3 (4): 305–360.

Khalil, F., and B. Parigi. 1998. The loan size as a commitment device. *International Economic Review* 39 (1): 135–150.

Kletzer, K., and B. Wright. 2000. Sovereign debt as intertemporal barter. *American Economic Review* 90 (3): 621–639.

Lacker, J., and J. Weinberg. 1989. Optimal contracts under costly state falsification. *Journal of Political Economy* 97 (6): 1345–1363.

Mas Colell, A., M. D. Whinston, and J. Green. 1995. *Microeconomic theory.* Oxford: Oxford University Press.

Myers, S. C., and N. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.

Myers, S. C., and R. G. Rajan. 1998. The paradox of liquidity. *Quarterly Journal of Economics* 113 (3): 733–771.

Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47 (1): 61–73.

Rogerson, W. 1985. The first-order approach to principal-agent problems. *Econometrica* 53 (6): 1357–1367.

Stiglitz, J. E., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.

———. 1983. Incentive effects of terminations: Applications to the credit and labor markets. *American Economic Review* 73 (5): 912–927.

Tirole, J. 1999. Incomplete contracts: Where do we stand? *Econometrica* 67 (4): 741–781.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21 (2): 265–293.

Webb, D. C. 1991. Long-term financial contracts can mitigate the adverse selection problem in project financing. *International Economic Review* 32 (2): 305–320.

———. 1992. Two-period financial contracts with private information and costly state verification. *Quarterly Journal of Economics* 107 (3): 1113–1123.

Wilson, R. 1968. On the theory of syndicates. *Econometrica* 36: 119–132.

# 5 Equilibrium in the Credit Market and Its Macroeconomic Implications

The preceding chapter extensively analyzed the characteristics of a loan contract as a complex relationship between a borrower and a lender. This chapter turns to the credit market to examine the formation of equilibrium interest rates when multiple borrowers and lenders compete.

It is important to notice that even if a partial equilibrium framework is adopted, the usual graphical analysis of supply and demand does not work in the context of the credit market. The reason is that the credit supply function may well be backward-bending for high levels of the interest rate. As a consequence, demand and supply curves may not intersect, which means that a new equilibrium concept (less demanding than the usual market-clearing condition) has to be designed to describe the outcome of a competitive credit market. Typically it involves a situation of *credit rationing* (the demand for credit exceeds supply at the prevailing interest rate).

Credit rationing has been the subject of an extensive literature. For instance, it has been taken as a postulate in the *availability doctrine* developed in the early 1950s.[1] However, before the useful contributions of Baltensperger (1978), Keeton (1979), and De Meza and Webb (1992), there was no clear-cut definition of equilibrium credit rationing. This situation resulted in some confusion. Therefore, section 5.1 defines credit rationing and explains the exact circumstances in which it may occur. Section 5.2 then pinpoints the reason behind this phenomenon, namely, the backward-bending credit supply curve. Section 5.3 shows that this backward-bending supply curve can be explained by adverse selection (Stiglitz and Weiss 1981), costly state verification (Williamson 1987), or moral hazard (Jaffee and Russell 1976; Bester and Hellwig 1987). However, section 5.4 shows that in the case of adverse selection, whenever collateral can be used as a screening device, credit rationing disappears (Bester 1985).

## 5.1 Definition of Equilibrium Credit Rationing

Following Baltensperger (1978), we define *equilibrium credit rationing* as occurring whenever *some borrower's demand for credit is turned down, even if this borrower is willing to pay all the price and nonprice elements of the loan contract*.[2]

"Price elements of the loan contract" means the interest rate charged by the bank, which is assumed to be unconstrained by the government. Of course, if there is a ceiling on credit rates, rationing can occur, but this is hardly surprising and not specific to the credit market. This discussion considers situations in which the demand for credit exceeds supply even though the banks are free to increase interest rates.

However, loan contracts are not only characterized by interest rates but also, as emphasized by Baltensperger (1978), by "nonprice elements" such as collateral requirements. If a borrower is turned down because he does not have enough collateral, this cannot be denominated as credit rationing. Similarly, it is important to understand that credit is not a perfectly divisible good. The fact that a borrower would be ready to borrow more at a given interest rate does not necessarily mean that he is rationed. Lending more to an individual borrower may increase the risk for the bank, and therefore the equilibrium interest rate may be a nonlinear function of the loan size. As a consequence, if the price of loans did not depend on the amount lent, firms would not take into account the marginal cost of their loans, and this would result in inefficiency (see problem 5.5.2).

The difference between the rejection of borrowers who do not meet these nonprice elements and credit rationing may appear more clearly in the context of complete contingent markets. In such a context, credit rationing is impossible because any borrower (say, a firm) can borrow up to the net present value of all the future cash flows he can generate in the future. For instance, Freimer and Gordon (1965) study a situation in which these future cash flows depend on the size of the investment. When returns to scale are decreasing, there is a maximum amount that the bank is ready to lend at a given interest rate; this should not be called credit rationing.

Another common use of the term *credit rationing* is when some categories of borrowers are totally excluded from the credit market. This phenomenon, known as red-lining, occurs because these borrowers do not have enough future cash flows or collateral to match their demand for credit. Again, this is not equilibrium credit rationing.

Finally, any institutional restrictions that can prevent lenders from offering differentiated conditions to heterogeneous borrowers, such as ceilings on interest rates or discriminatory pricing, may lead to disequilibrium credit rationing. For instance, Smith (1972) shows that rationing may be Pareto-improving when firms have different equity-to-assets ratios and banks have to demand the same interest rate from all of them. Similarly, Jaffee and Modigliani (1969) show that a monopolistic bank that cannot use price discrimination (because of regulation) will typically ration those borrowers for which it would set higher rates in the absence of regulation. Using a related model, Blackwell and Santomero (1982) emphasize the fact that rationing concerns essentially the firms with a higher demand elasticity. Therefore, the model predicts that larger firms, which have access to alternative sources of financing, will more likely be rationed, which seems contradicted by casual empiricism. Finally, Cukierman (1978) uses a similar model to examine the macroeconomic implications of credit rationing (see Devinney 1986 and Jaffee and Stiglitz 1990 for an overview).

## 5.2 The Backward-Bending Supply of Credit

This section shows how equilibrium credit rationing can appear as soon as the expected return on a bank loan (for a given category of borrowers) is not a monotonic function of the nominal rate of this loan (fig. 5.1).

For the moment, we take this property as given (it is explained in section 5.3) and explore its consequences on banks' behavior. Consider the credit market for a homogeneous category of borrowers, and examine the type of competition that prevails.

A monopolistic bank facing the return schedule of figure 5.1 will never offer an interest rate above $R^*$. This explains why a monopolistic bank may prefer to ration credit applicants. To understand why a competitive equilibrium of the banking sector

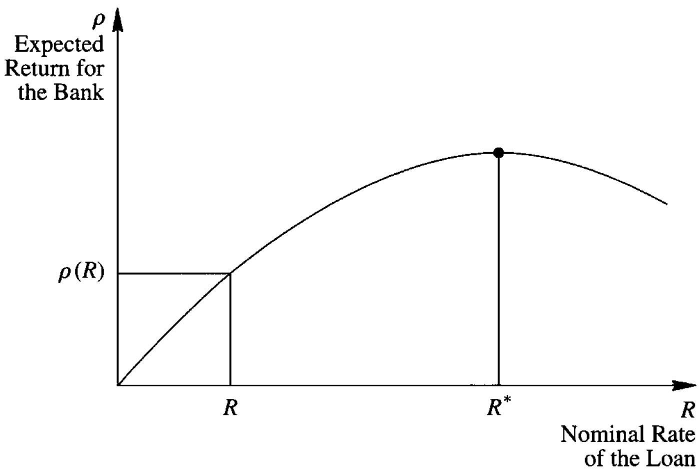



**Figure 5.1**
Expected return to the bank as a function of nominal rate of loan.

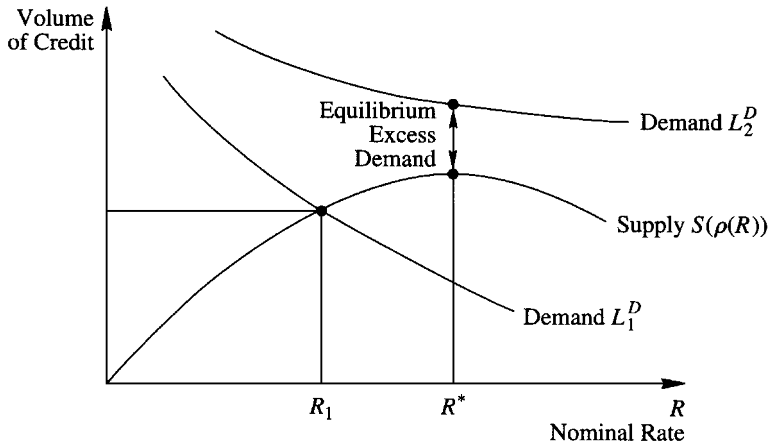



**Figure 5.2**
Equilibrium credit rationing.

may also lead to credit rationing, the aggregate demand and supply of loans must be examined. The aggregate demand analysis is straightforward; it is a decreasing function of the interest rate. The aggregate supply depends on the cost of financing for banks, say, through deposits. In a competitive equilibrium the banks' expected rate of return $\rho$ equals the cost of financing (zero-profit condition). Assuming that the supply of deposits increases with the interest rate paid by banks, the supply of loans by banks will be backward-bending, as can be seen by inverting the axes in figure 5.1 (see Hodgman 1960).

Figure 5.2 shows how credit rationing may occur. If the demand schedule is $L_1^D$, a competitive equilibrium exists, characterized by the equality of supply and demand, so that the nominal rate $R_1$ clears the market. On the other hand, if the demand schedule is $L_2^D$, the supply and demand curves do not intersect. An equilibrium with credit rationing will then occur, characterized by the interest rate $R^*$ and zero profit for the banks.

Although competition between banks is not explicitly modeled by Stiglitz and Weiss (1981) in game theory terms, the implicit rules of the game are that banks are price setters on the credit market and quantity setters on the deposit market. In other words, they simultaneously choose a capacity (demand for deposits) and a nominal loan rate in such a way that their profit is maximized, taking as given the return demanded by depositors and the loan rates set by other banks.

The equilibrium that prevails in this case is characterized by type II credit rationing (see note 2), that is, only some randomly selected applicants will obtain the loan they demand. This is due to the assumption of indivisibility of the investment projects. If the projects were divisible, the type of credit rationing that would prevail would depend on the technology of borrowers: under decreasing returns to scale it would be of type I, and under increasing returns to scale it would be of type II. Notice that supply and demand could intersect at an interest rate $\hat{R}$ larger than $R^*$. In that case, the intersection point is not an equilibrium because any bank can increase its profits by decreasing its interest rates if it is not bound to serve all credit applicants. The market-clearing level $\hat{R}$ is not sustainable.

## 5.3 Equilibrium Credit Rationing

So far, this discussion has taken as a postulate that the credit supply function could be backward-bending for high levels of interest rates. More precisely, it has been assumed that the expected return $\rho$ on a loan is not always a monotonic function of the nominal rate $R$ of this loan. This section shows how this result can be explained by asymmetric information, due to adverse selection (Stiglitz and Weiss 1981), costly state verification, or moral hazard.

### 5.3.1 Adverse Selection

The basic assumption of Stiglitz and Weiss (1981) is that borrowers differ by a risk parameter $\theta$, which is privately observed. The bank knows only the statistical distribution of $\theta$ among the population of potential borrowers. The crucial ingredient is that the characteristics of the loan offered by the bank will affect the composition of the population of firms that actually apply for the loan. In the model of Stiglitz and Weiss, all firms are assumed to bring the same amount of collateral $C$, which can therefore not be used as a screening device. Being unable to observe $\theta$, the banks cannot discriminate among firms. They offer the same standard debt contract, in which all firms have to repay a fixed amount $R$ (if they can) or their cash flow will be seized by the bank. Concerning unsecured loans, each firm will obtain a profit $\pi$ that is related to its cash flow $y$ by the familiar expression

$$\pi(y) = \max(0, y - R).$$

More generally, if a collateral $C$ is introduced, the profit function becomes

$$\pi(y) = \max(-C, y - R).$$

A crucial property needed by Stiglitz and Weiss is that $E[\pi(y) \mid \theta]$ be an increasing function of $\theta$. Since the profit function is convex (fig. 5.3), this property is satisfied if higher $\theta$s indicate riskier distributions of cash flows (in the sense of Rothschild and Stiglitz 1970). Notice that the convexity of the profit function comes from the rules of the standard debt contract, which are here exogenously given.[3]

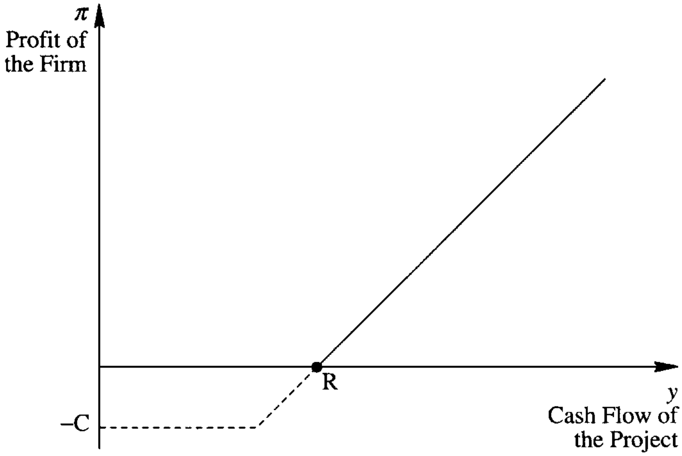



**Figure 5.3**
Profit to the firm as a function of cash flow from project.

**Risk Characteristics of the Demand for Loans**

Assume that firms have a reservation level $\bar{\pi}$ for their expected profits, so that below that level they will not be interested in developing the project financed by the bank loan. For instance, $\bar{\pi}$ could be the level of profits the firm can obtain with another source of funds or another project. Because projects are indivisible, the total demand for loans is given by the number of firms with expected profits higher than $\bar{\pi}$. Since $E(\pi(y) \mid \theta)$ is increasing in $\theta$, there is at most one value $\theta^*$ that satisfies

$$E(\pi(y) \mid \theta^*) = \bar{\pi}.$$

Focus on the case where this value exists (disregarding both cases where either all firms demand credit or none of them does), so that the demand for loans is determined by the population of firms with values of $\theta$ in the interval $[\theta^*, \bar{\theta}]$.

Consider now the banks' expected profits. They depend on the amount of the repayment $R$ and the distribution of the cash flows of the firms applying for a loan. The effect of an increase in interest rates on the banks' expected profit is therefore twofold:

- It increases the profit the bank makes on any individual loan granted to a given firm $\theta$.
- It decreases the expected profit $E(\pi(y) \mid \theta)$ for every $\theta$. Thus the number of applicants decreases ($\theta^*$ increases), so the population of firms that demand a loan becomes more risky.

Thus an increase in the interest rate decreases the demand for loans, but it is the less risky firms that drop out of the market. As a consequence, an increase in the interest rate *need not* necessarily increase the banks' expected profits. This will depend on which of the two effects dominates: the direct effect of the interest rate increase for a given population of borrowing firms, or the indirect effect of changing the risk of this population. The distribution of $\theta$ will play an important role. For some of these distributions the banks' expected return on loans will be single-peaked, with a maximum for repayment $R^*$, and this results in the nonmonotonic profile shown in figure 5.1.

**The Role of the Different Assumptions**

We now revisit the previous assumptions that are crucial to obtaining equilibrium credit rationing.

Recall that banks cannot a priori distinguish between firms. Yet, in general, the banks will try to find devices in order to sort out firms. If banks find a way to distinguish between different classes of risk, the peaks of the expected return functions will occur at different levels in each of these classes, and therefore credit rationing will occur only in (at most) one of them. This has led Riley (1987) to think that rationing as explained by the Stiglitz-Weiss model would not be observed frequently. Still, Stiglitz and Weiss never claimed that the type of credit rationing their model described was frequent, nor that it was likely, but only that it *could* occur in a *competitive* framework.

A second assumption is that the parameter $\theta$ ranks firms by increasing risk.[4] If, for instance, the probability distribution for the firms' cash flows is instead ranked according to first-order stochastic dominance, an increase in interest rates would *decrease* the average risk of the population of borrowers, and consequently credit rationing would never occur at the equilibrium interest rate (see, e.g., De Meza and Webb 1987).

Finally, it has been assumed that the function relating the banks' expected return to the quoted interest rate was single-peaked. This is not a consequence of previous assumptions but only a possibility. If this function is increasing, the equilibrium will be without credit rationing.

Perhaps the main criticism that can be made of the Stiglitz-Weiss model is that the debt contracts used are exogenously given and do not allow for any sorting mechanism. Contributions presented in section 5.4 explore this direction.

### 5.3.2 Costly State Verification

Williamson (1987) offers an alternative theoretical explanation of credit rationing that is based on the costly state verification paradigm of Townsend (1979) and Gale and Hellwig (1985) (see chapter 4). This theoretical explanation has two merits: it justifies the use of the standard debt contract (which is optimal in this context; see chapter 4), and it does not require additional assumptions on the distribution of returns.

Let $\tilde{y}$ denote the random return of the borrower's project, assumed to be unobservable by the lender except when the lender performs an audit that costs $\gamma$. If $R$ denotes the nominal unit repayment (1 plus the nominal interest rate) of the debt contract, and if $\tilde{y}$ has a density $f(y)$, continuous and positive on its support $[\underline{y}, \bar{y}]$, then the return to the lender (as a function of $R$) has the following expression:

$$\rho(R) = \int_{\underline{y}}^{R} (y - \gamma) f(y)\, dy + \int_{R}^{\bar{y}} R f(y)\, dy.$$

Since $f$ is continuous, $\rho$ is (continuously) differentiable, and

$$\frac{d\rho}{dR} = (R - \gamma) f(R) + \int_{R}^{\bar{y}} f(y)\, dy.$$

For $R$ close enough to $\bar{y}$, this is negative (since $f(\bar{y}) > 0$). Therefore, $\rho$ has an interior maximum, and equilibrium credit rationing may arise. To summarize Williamson's simple argument, when failure is costly to the lender, an increase in the nominal rate of a loan may decrease the net return to the bank, since it increases the probability of failure of the borrower.

### 5.3.3 Moral Hazard

In general, lenders will not participate in the management of the projects they finance. This may come from a self-imposed policy decision of financial institutions, aimed at preserving their reputation, or because the law may penalize such a behavior by lowering the rank of the bank in the creditors line in case of bankruptcy, if it is proved to have been involved in the management of the bankrupt firm.

Consequently, it is not always easy for the lender to enforce a particular use for the credit granted to the firm.[5] Nor is it easy to ascertain whether the firm has the capacity to repay. This is the main source of moral hazard problems in credit activities.

These moral hazard problems may lead to credit rationing exactly in the same way as adverse selection does. Moral hazard may generate a nonmonotonic relation between quoted interest rates and expected rates of return, as in the Stiglitz-Weiss model, and therefore lead to equilibrium credit rationing.

Consider, as in chapter 2, a firm that has a choice between a good technology, which produces $G$ (for a unit investment) with probability $\pi_G$ (and zero otherwise), and a bad technology, which produces $B$ with probability $\pi_B$.

Assume that the good technology has a higher expected return,

$\pi_G G > \pi_B B,$

but the cash flow in case of success is higher for the bad technology,

$B > G,$

which implies that $\pi_B < \pi_G$. Therefore, the bad technology is riskier than the good one.

The loan contract specifies the amount $R$ to be repaid by the firm in case of success. Since the size of the loan is normalized to 1, $R$ can be interpreted as (1 plus) the interest rate of the loan.

The technology choice by the firm is then straightforward. Recall from equations (2.25) and (2.26) that the good technology will be chosen if and only if

$\pi_G(G - R) \geq \pi_B(B - R).$

Defining $\hat{R} = (\pi_G G - \pi_B B)/(\pi_G - \pi_B)$, this is equivalent to

$R \leq \hat{R}.$

We can therefore determine the expected return on the loan for the bank as a function of the repayment required (fig. 5.4). For values of $R$ lower than $\hat{R}$, expected repayment is $\pi_G R$, and for values of $R$ higher than $\hat{R}$, it is equal to $\pi_B R$. The region

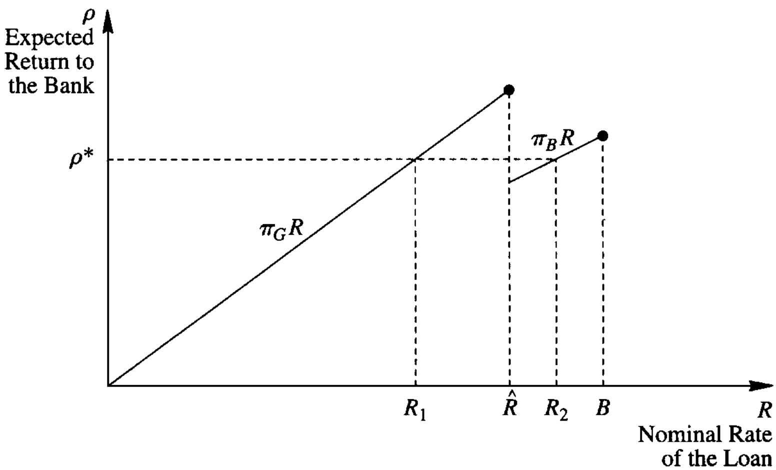



**Figure 5.4**
Expected return to the bank as a function of $R$ in Bester-Hellwig (1987) model: Case 1.

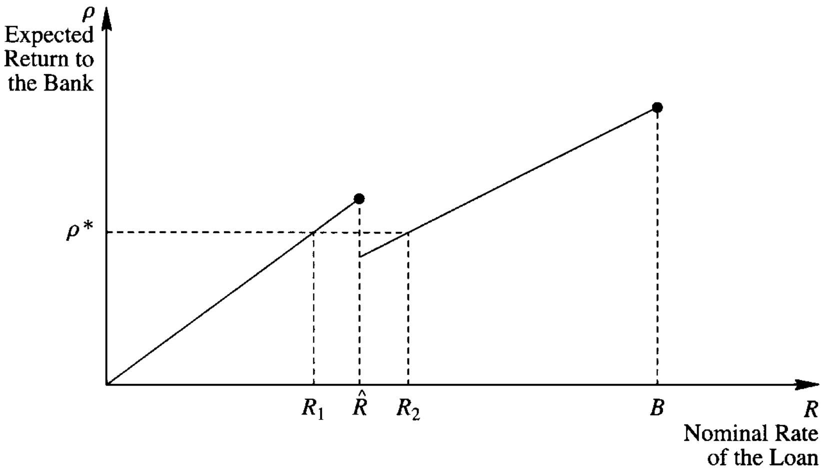



**Figure 5.5**
Expected return to the bank as a function of $R$ in Bester-Hellwig (1987) model: Case 2.

$R > B$ is not interesting, since the repayment cannot exceed $B$, and therefore expected repayment is constant in this region and equal to $\pi_B B$.

As before, the supply of credit will be obtained as a function of expected repayment $\rho$. The simplest specification corresponds to an infinitely elastic supply of funds (when $\rho$ is equal to some constant $\rho^*$). In that case, there may be two equilibria (when $\pi_B \hat{R} < \rho^* < \pi_B B$), as shown in figures 5.4 and 5.5. Both $R_1$ and $R_2$ are interest rates at which the credit market clears. This result is crucially related to the assumption of price-taking behavior by banks, because if lenders were price setters, $R_2$ would not be an equilibrium interest rate. By offering a loan rate $R_1 + \varepsilon$ just above $R_1$ ($\varepsilon$ is positive and small), a bank could attract all borrowers and make a positive profit. In any case, equilibrium credit rationing cannot occur when the supply of funds is infinitely elastic, since markets will clear.

Assuming, as usual, that the deposit supply function $S(\rho)$ is not infinitely elastic, the function $S(\rho(R))$ is not increasing but reaches a global maximum at point $\hat{R}$ for the Bester-Hellwig (1987) model case 1 (see fig. 5.4) and a local maximum at point $\hat{R}$ for case 2 (see fig. 5.5). The credit market may then clear or not, exactly as it happens in the Stiglitz-Weiss model (see fig. 5.2).

Rationing will occur for a supply function that is strictly increasing in the expected return $\rho$ if

$$D > S(\rho(\hat{R})),$$

where $D$ is the (inelastic) demand for credit for a quoted interest rate equal to $\hat{R}$.

This simple model gives the main intuition on how moral hazard may lead to credit rationing. It may be extended to include collateral in the specification of the loan contract. This results in a modification of the incentives for choosing between the two investment projects. It is easily shown that the previous result holds true provided that $\hat{R}$ is replaced by $\hat{R} + C$, where $C$ is the value of the collateral.

## 5.4 Equilibrium with a Broader Class of Contracts

A banker facing a heterogeneous distribution of potential borrowers may benefit from discriminating among them. The fact that the banker is unable to identify the borrowers will lead him to consider sorting devices constructed in such a way that each type of borrower will choose a specific type of contract. Self-selection of clients will result from product differentiation. This idea has been explored, for instance, by Mussa and Rosen (1978) in the case of a monopoly for a durable good and by Rothschild and Stiglitz (1976) in the case of a competitive insurance market.

A natural way to model this strategy of the bank in the credit market is to consider a menu of contracts[6] $\gamma_i = (R_i, C_i)_{i \in I}$ specifying, together with an interest rate $R_i$, a collateral requirement $C_i$. This idea has been explored by Wette (1983), Bester (1985; 1987) and Chan and Kanatas (1985). We follow Bester (1985), who uses a model with only two values for the risk parameter $\theta$ ($\theta \in \{\theta_L, \theta_H\}$), where $\theta_H$ is a higher risk than $\theta_L$ in the sense of Rothschild and Stiglitz (1970).[7]

In Bester's model, the wealth constraint is not binding and collateral has a cost, so the perfectly secured loan solution is inefficient. The strategy of each bank is to offer two contracts $\gamma_L$, $\gamma_H$ (one for each type of borrowers). When $\gamma_L \neq \gamma_H$, we have separating contracts. When $\gamma_L = \gamma_H$, we talk of a pooling situation. Competition on each of these contracts implies that expected profit is zero for each of them, so that

$$\rho_L(\gamma_L) = \rho_H(\gamma_H) = \rho_0,$$

where $\rho_L$ is the expected return computed with $L$'s cash flow distribution (and similarly for $\rho_H$), and $\rho_0$ is the banks' cost of funds.

A *separating equilibrium* is defined as a pair of (distinct) contracts $(\gamma_L^*, \gamma_H^*)$ such that

1. $\gamma_L^*$ is preferred by low-risk firms, and $\gamma_H^*$ is preferred by high-risk firms (self-selection constraints);
2. no bank is able to offer another contract on which it obtains an expected rate of return higher than $\rho_0$;
3. $\rho_L(\gamma_L^*) = \rho_H(\gamma_H^*) = \rho_0$.

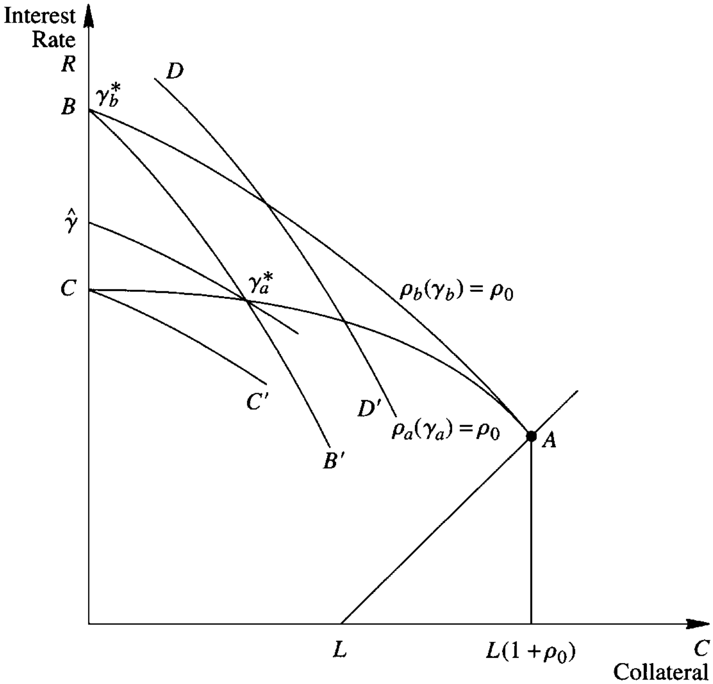



**Figure 5.6**
Separating equilibrium in Bester (1985) model: The only candidate is $(\gamma_L^*, \gamma_H^*)$.

A *pooling equilibrium* would be defined in the same way when $\gamma_L^* = \gamma_H^* = \gamma^*$ (both types of firms choose the same contract), so that the expected return of the contract $\gamma^*$, denoted $\bar{\rho}(\gamma^*)$, is estimated with the whole population of firms. The analogue to condition 3 is written as $\bar{\rho}(\gamma^*) = \rho_0$.

Bester establishes that if an equilibrium exists, it entails no credit rationing. This point can be proved by using a figure in the $(C, R)$ plane (fig. 5.6). Banks will prefer contracts with higher collateral $C$ and interest rates $R$, whereas firms will prefer contracts with lower collateral and interest rates. Notice that the isoprofit of the bank and the borrower curves are different because the existence of a cost for pledging (or monitoring) the collateral implies that this is not a constant sum game.

In figure 5.6 the $AB$ curve (resp. $AC$) represents the locus of all contracts that would entail zero expected profit for the bank if they were chosen only by the type $H$ (resp. type $L$) borrowers. The curves $BB'$ and $DD'$ are isoprofit curves when borrowers are of type $H$. Note that an additional unit of collateral costs more to the firm than what the bank will obtain from it. Therefore, the decrease in interest rate that compensates for a unit increase in collateral will be greater for the firm than for the bank. This explains why the $BB'$ curve is steeper than the $AB$ curve. (If there were no such pledging cost, the $AB$ and $BB'$ curves would merge.)

Isoprofit curves corresponding to type $H$ firms are above and to the right of those for the type $L$ firms, since the former are riskier and the firms' profits are a convex function of the cash flow they obtain, as in the Stiglitz-Weiss model.

To establish that the contracts $\gamma_L^*$ and $\gamma_H^*$ define a separating equilibrium requires confirming that $\gamma_L^*$ is preferred by type $L$ firms and $\gamma_H^*$ by type $H$ firms. This is clearly the case because the two contracts are indifferent from $H$'s point of view, and type $L$ firms strictly prefer $\gamma_L^*$ given their lower risk. On the other hand, condition 3 is satisfied because each contract is on the bank's zero-profit curve.

Finally, it is necessary to confirm that condition 2 of the definition is satisfied, both for separating and for pooling contracts. First, no separating pair of contracts dominates $(\gamma_L^*, \gamma_H^*)$. Indeed, $\gamma_H^*$ is the contract preferred by $H$ types on the zero-profit curve; no other contract $\gamma_H$ can attract them and make a positive profit. On the other hand, for a contract $\gamma_L$ to be preferred by $L$ and make zero profit, it should be to the left of $\gamma_L^*$ and on the $AC$ locus. Such a contract would also attract all type $H$ borrowers, and the bank would suffer a loss.

Second, the zero-profit condition for a pooling contract $(\bar{\rho}(\gamma) = \rho_0)$ defines a curve $A\tilde{\gamma}$ (figs. 5.7a and 5.7b). On this curve both $L$ and $H$ prefer $\tilde{\gamma}$, so $\tilde{\gamma}$ is the only candidate for a pooling contract. If the indifference curve for type $L$ borrowers that goes through $\gamma_L^*$ intersects the vertical axis at a point $\hat{\gamma}$ below $\tilde{\gamma}$ (see fig. 5.7a), then $L$ will stick to $\gamma_L^*$ and the contract $(\gamma_L^*, \gamma_B^*)$ is the only separating equilibrium. If, on the contrary, $\hat{\gamma}$ is above $\tilde{\gamma}$ (see fig. 5.7b), then the equilibrium does not exist because it is possible to design profitable contracts $\gamma_L$ that will attract only type $L$ borrowers and make a profit. The reasons that the equilibrium fails to exist are exactly the same as in Rothschild and Stiglitz (1976): $\tilde{\gamma}$ is not an equilibrium because it can be destabilized by a separating contract, yet the separating equilibrium is itself dominated by the pooling contract $\tilde{\gamma}$.

To summarize, Bester's model shows that if the equilibrium exists, no credit rationing will occur because collateral is used to sort out the different (nonobservable) types of borrowers. Thus, again, enlarging the class of loan contracts eliminates credit rationing.

In some cases, the amount of collateral needed for the equilibrium to be separating may exceed the agent's wealth. This does not imply a return to the Stiglitz-Weiss case, as has been pointed out by Besanko and Thakor (1987). Besanko and Thakor use a slightly different framework, in which agent $L$'s cash flow distribution is less risky in the sense of first-order stochastic dominance. Figures 5.7a and 5.7b can still be used to show that a contract $(R^L, C^L)$ to the left of $\gamma_L^*$ on the $AC$ curve will also be preferred by agent $H$. Still, the class of mechanisms (contracts) may be enriched

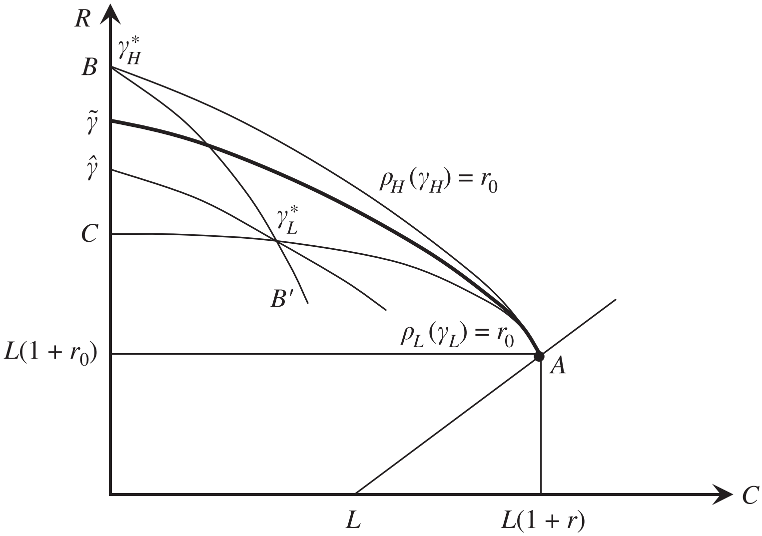



**Figure 5.7a**
Separating equilibrium in Bester (1985) model: Equilibrium exists.

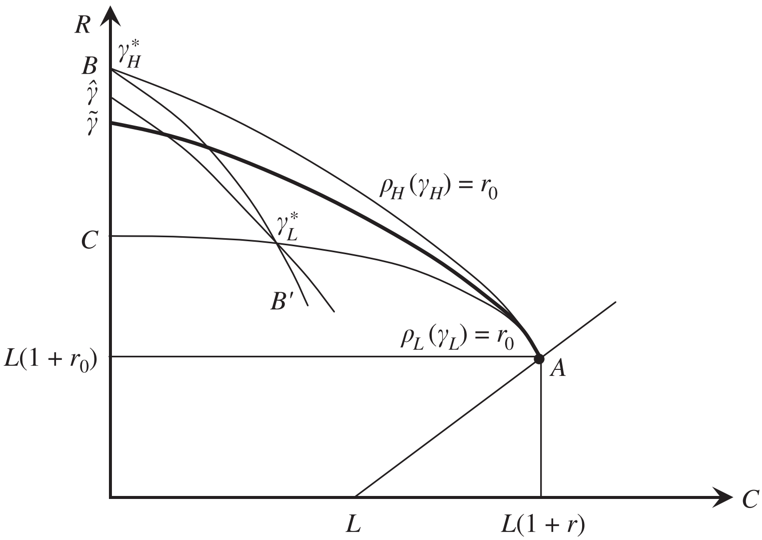



**Figure 5.7b**
Separating equilibrium in Bester (1985) model: Equilibrium does not exist.

by introducing the possibility of stochastic rationing. If the agents demanding contract $(R^L, C^L)$ are rationed, this might have a stronger effect on type $H$ borrowers (who may stick to contract $\gamma_H^*$) than on the type $L$ borrowers, to whom contract $\gamma_H^*$ is not appealing. In this way, separation is restored (provided $W$ is not too low), and a competitive equilibrium is obtained in which the less risky agents are rationed, a paradoxical result.

It may be argued that the game-theoretic formulation of competition in contracts, as in Bester's model, is not completely satisfactory, in particular because equilibrium may fail to exist (at least in pure strategies). This has led Hellwig (1987) to examine more complex games in which the banks can reject some applicants after having observed all the contracts offered and the choices of borrowers. Hellwig shows that an equilibrium always exists, and more important, it may be a pooling equilibrium, which reintroduces the possibility of credit rationing.[8]

## 5.5 Problems

### 5.5.1 The Model of Mankiw

Mankiw's (1986) model considers an economy à la Stiglitz and Weiss in which each firm $\theta$ has an investment technology where one unit investment returns $X_\theta$ with probability $1-\theta$ and zero with probability $\theta$. Firms invest only if they have a strictly positive expected profit. The parameter $\theta$ follows a uniform distribution on [0, 1].

1. Assume $(1-\theta)X_\theta = \overline{X}$, so that all the projects have the same expected return, and higher $\theta$s indicate riskier projects, as in Stiglitz and Weiss. Compute the expected return for each repayment level $R$ fixed in the loan contract, assuming that there is no collateral.

2. Assume that investors are able to obtain an exogenous riskless return $r$. Characterize the different types of equilibria that can be obtained, depending on the level of $r$.

### 5.5.2 Efficient Credit Rationing

This problem is adapted from De Meza and Webb (1992). Consider an economy with risk-neutral agents in which firms (assumed to have no internal source of funds) develop projects that succeed in the state of nature $S_G$ (good) and fail in the state of nature $S_B$ (bad), in which case the investment returns less than its cost. Let $p_j$, $j = G, B$, denote the probability of these two events. The expected return on the project when a loan of size $k$ is obtained is

$$p_G f(k, S_G) + p_B f(k, S_B),$$

where $f$ is the production function, conditional on the state of nature. Competitive risk-neutral banks fund the project, provided they obtain at least the exogenous rate of return $r$. Let $r_L$ denote the interest rate due on the loans when the firm is successful. In the bad state of nature, the bank seizes all the output.

1. Show that if firms act as price takers in the credit market and if $f(k, S_B)$ is not linear in $k$, the allocation will not be efficient.

2. Show that if the banks competed in nonlinear prices (interest rates) on loans, efficiency would be restored.

3. If the equilibrium described in part 2 prevails, when is there apparent credit rationing?

### 5.5.3 Too Much Investment

This problem is adapted from De Meza and Webb (1987). Consider an economy in which a continuum of risk-neutral agents endowed with the same wealth $W$ have access to an investment project that yields $y$ with probability $p$ and zero with probability $1 - p$. The agents differ in their probability of success $p$, which ranges in the interval $[0, 1]$ and has a distribution with density $f(p)$. The investment requires a capital $I$ superior to $W$ so that agents will have to obtain a loan $L = I - W$. The supply of loanable funds (deposits) is an increasing function $S(r)$ of the money market interest rate $r$, which is assumed to be equal to the riskless rate. Assume perfect information and that funding is done through a standard debt contract.

1. Write the agent's individual rationality constraint, given that the agent has the choice between investing $I$ or depositing $W$.

2. Write the equations determining the equilibrium money market rate $r$, the marginal investor, and the zero (marginal) profit for the bank.

3. Confirm that only projects with positive net present value will be implemented.

4. Assuming now that $p$ is not observable, so that $R$ cannot depend on $p$, compute the equilibrium interest rate $\hat{r}$ and the probability of success $\hat{p}$ of the marginal investor. Show that there is overinvestment.

## 5.6 Solutions

### 5.6.1 The Model of Mankiw

1. Firm $\theta$ will demand a loan if $X_\theta - R > 0$ (that is, if $\theta > 1 - \overline{X}/R$). Only the risky firms in $[1 - \overline{X}/R, 1]$ ask for a loan, and the total amount lent equals $\overline{X}/R$. When $\overline{X} > R$, all the firms obtain a loan. The expected return for the bank is thus

$$\rho = \frac{R}{\overline{X}}\left\{\int_{1-\overline{X}/R}^{1} (1-\theta)R\,d\theta\right\} = \frac{\overline{X}}{2},$$

if $\overline{X} \le R$, and

$$\rho = \frac{R}{2} \quad \text{if } \overline{X} > R.$$

2. If $r \le \overline{X}/2$, the equilibrium (nominal) rate is $R = 2r$. However, if $r > \overline{X}/2$, the market for credit collapses: the unique equilibrium involves no trade.

### 5.6.2 Efficient Credit Rationing

1. The banks' zero-profit condition can be written as

$$p_G(1+r_L)k + p_B f(k, S_B) = (1+r)k. \tag{5.1}$$

The firm maximizes its expected profit, given that it will not get anything in state $S_B$ and taking $r_L$ as exogenously given:

$$\max_k \; p_G[f(k, S_G) - (1+r_L)k],$$

which leads to the first-order condition

$$\frac{\partial f}{\partial k}(k, S_G) = 1 + r_L. \tag{5.2}$$

However, the efficient amount of capital is the one for which the following equality is obtained:

$$p_G \frac{\partial f}{\partial k}(k, S_G) + p_B \frac{\partial f}{\partial k}(k, S_B) = 1 + r. \tag{5.3}$$

Replacing $r_L$ given by (5.2) in (5.1) and dividing by $k$,

$$p_G \frac{\partial f}{\partial k}(k, S_G) + \frac{p_B}{k} f(k, S_B) = 1 + r,$$

so that (5.3) does not hold except when $f(k, S_B)$ is linear in $k$.

2. Assume that banks can compete in nonlinear prices. Bertrand competition implies that $r_L(k)$ will be set in such a way that the bank's zero-profit condition (5.1) is satisfied for all $k$. Then the firm's profit is just equal to the total surplus

$$p_G f(k, S_G) + p_B f(k, S_B) - (1+r)k,$$

and efficiency of equilibrium is warranted.

3. Apparent rationing means that, at the equilibrium loan rate $r_L(k^*)$, firms would like to borrow more. Given (5.2), this is satisfied if and only if

$$\frac{\partial f}{\partial k}(k^*, S_G) > 1 + r_L(k^*).$$

Using the equations that define $k^*$ and $r_L(k^*)$,

$$\begin{cases} p_G \dfrac{\partial f}{\partial k}(k^*, S_G) + p_B \dfrac{\partial f}{\partial k}(k^*, S_B) = 1 + r, \\ p_G(1 + r_L(k^*)) + p_B \dfrac{f(k^*, S_B)}{k^*} = 1 + r. \end{cases}$$

Therefore, apparent rationing occurs exactly when

$$\frac{\partial f}{\partial k}(k^*, S_B) < \frac{f(k^*, S_B)}{k^*}.$$

However, as proved in part 2, this credit rationing is efficient.

### 5.6.3 Too Much Investment

1. The individual rationality constraint is

$$p(y - R(p)) \geq (1 + r)W. \tag{5.4}$$

2. For the marginal investor, (5.4) holds with equality, and in addition the bank's zero-profit condition (for each $p$) implies

$$pR(p) = (1 + r)L. \tag{5.5}$$

Therefore (5.4) is satisfied when

$$p \geq \bar{p}(r) \stackrel{\text{def}}{=} \frac{(1 + r)}{p}(W + L),$$

and the demand for credit is

$$D(r) = \int_{\bar{p}(r)}^{1} f(p)\, dp.$$

The market clearing condition is

$$\int_{\bar{p}(r)}^{1} f(p)\, dp = S(r). \tag{5.6}$$

3. Note that adding (5.4) and (5.5) yields

$$py \geq (1 + r)I,$$

so that only profitable projects are funded.

4. $\hat{p}$ and $\hat{r}$ are jointly defined by

$$\hat{p}(y - \hat{r}) = (1 + r)W,$$

and

$$\hat{r}\frac{\int_{\hat{p}}^{1} pf(p)\,dp}{1 - F(\hat{p})} = (1 + r)L.$$

Adding these two equalities yields

$$\hat{p}y + \hat{r}\frac{\int_{\hat{p}}^{1} (p - \hat{p})f(p)\,dp}{1 - F(\hat{p})} = (1 + r)I.$$

The integral is clearly positive. Therefore $\hat{p}y < (1 + r)I$, which means that there is overinvestment.

## Notes

1. According to this doctrine, banks are limited by the availability of the funds they can attract. Therefore, credit is always rationed: the credit market equilibrium is purely determined by the supply conditions. In such a context, monetary policy would be very effective. Changes in the money supply would have direct effects on credit instead of indirect effects channeled via changes in interest rates. However, this theory suffers from a major drawback. It does not explain why banks cannot increase their interest rates to equate demand with supply and make more profit. For a discussion of the availability doctrine, see Baltensperger and Devinney (1985) or the introduction of Clemenz (1986).

2. Following Keeton (1979), one can distinguish two types of rationing:

- *Type I rationing* occurs when there is partial or complete rationing of all the borrowers within a given group.
- *Type II rationing* occurs within a group that is homogeneous from the lender's standpoint, so that some randomly selected borrowers of this group obtain the full amount of the loan they demand while others are rationed.

To see the difference between type I and type II rationing, assume that $2N$ borrowers with demand equal to 1 face a supply of $N$. Type I rationing would imply that each borrower obtains half a unit. Type II would imply that only $N$ borrowers randomly selected out of the $2N$ potential ones obtain one unit.

3. Different justifications for this type of contract were presented in chapter 4.

4. Recall that Rothschild and Stiglitz (1970) defined this notion as follows. Let $\tilde{y}_1$ and $\tilde{y}_2$ be two random variables. $\tilde{y}_1$ is more risky than $\tilde{y}_2$ if and only if for all concave functions $u(\cdot)$, $Eu(\tilde{y}_1) \leq Eu(\tilde{y}_2)$. In economic terms, this means that any risk-averse investor would prefer the random return $\tilde{y}_2$ to $\tilde{y}_1$.

5. Noteworthy exceptions are mortgage loans and project financing. But even inventory financing (which, in principle, can be easily monitored by banks) has, in practice, an important record of fraud on the part of the borrowing firms.

6. This was shown in chapter 4, as was the fact that there are other possible menus of contracts, in particular those linking repayments to the loan size. These types of contracts, which lead to similar results, have been explored by Milde and Riley (1988).

7. The extension to $n$ types of risk is modeled in Bester (1987). The problems in modeling loans that are backed by collateral are that

- collateral may be limited by the entrepreneur's wealth;
- if there is no such limit, the optimal contracts may involve a 100 percent collateral, so that imperfect information becomes irrelevant. The solution is to introduce a cost of collateral, as in Bester (1985; 1987).

8. Since Hellwig's game is a sequential game under asymmetric information, he must use the concept of perfect Bayesian equilibrium and some sophisticated refinement criteria. For a clear presentation of these concepts, see, for instance, Fudenberg and Tirole (1991). Interesting discussions of game theory modeling of the credit market may be found in Clemenz (1986) and Clemenz and Ritthaler (1992).

## References

Baltensperger, E. 1976. The borrower-lender relationship, competitive equilibrium and the theory of hedonic prices. *American Economic Review* 66 (3): 401–405.

———. 1978. Credit rationing: Issues and questions. *Journal of Money, Credit and Banking* 10 (2): 170–183.

Baltensperger, E., and T. Devinney. 1985. Credit rationing theory: A survey and synthesis. *Zeitschrift für die gesamte Staatswissenschaft* 141 (4): 475–502.

Besanko, D., and A. V. Thakor. 1987. Collateral and rationing: Sorting equilibria in monopolistic and competitive credit markets. *International Economic Review* 28 (3): 671–689.

Bester, H. 1985. Screening vs rationing in credit markets with imperfect information. *American Economic Review* 75 (4): 850–855.

———. 1987. The role of collateral in credit markets with imperfect information. *European Economic Review* 31 (4): 887–899.

Bester, H., and M. Hellwig. 1987. Moral hazard and equilibrium credit rationing. In *Agency theory, information and incentives*, ed. G. Bamberg and K. Spremann. Heidelberg: Springer.

Blackwell, N., and A. Santomero. 1982. Bank credit rationing and the customer relation. *Journal of Monetary Economics* 9 (1): 121–129.

Chan, Y., and G. Kanatas. 1985. Asymmetric valuations and the role of collateral in loan agreements. *Journal of Money, Credit and Banking* 17 (1): 84–95.

Clemenz, G. 1986. *Credit markets with asymmetric information*. Berlin: Springer.

Clemenz, G., and M. Ritthaler. 1992. Credit markets with asymmetric information: A survey. *Finnish Economic Papers* 5 (1): 12–26.

Cukierman, A. 1978. The horizontal integration of the banking firm, credit rationing and monetary policy. *Review of Economic Studies* 45 (1): 165–178.

De Meza, D., and D. Webb. 1987. Too much investment: A problem of asymmetric information. *Quarterly Journal of Economics* 102 (2): 281–292.

———. 1992. Efficient credit rationing. *European Economic Review* 36 (6): 1277–1290.

Devinney, T. 1986. *Rationing in a theory of the banking firm*. Berlin: Springer.

Freimer, M., and M. Gordon. 1965. Why bankers ration credit. *Quarterly Journal of Economics* 79: 397–410.

Fudenberg, D., and J. Tirole. 1991. *Game theory*. Cambridge, Mass.: MIT Press.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

Hellwig, M. 1987. Some recent developments in the theory of competition in markets with adverse selection. *European Economic Review* 31 (1-2): 319–325.

Hodgman, D. 1960. Credit risk and credit rationing. *Quarterly Journal of Economics* 74: 258–278.

Jaffee, D., and F. Modigliani. 1969. A theory and test of credit rationing. *American Economic Review* 59: 850–872.

Jaffee, D., and T. Russell. 1976. Imperfect information, uncertainty and credit rationing. *Quarterly Journal of Economics* 90: 651–666.

Jaffee, D., and J. Stiglitz. 1990. Credit rationing. In *Handbook of monetary economics*, ed. B. M. Friedman and F. H. Hahn, vol. 2, ch. 16. Amsterdam: Elsevier.

Keeton, W. 1979. *Equilibrium credit rationing*. New York: Garland Press.

Mankiw, G. 1986. The allocation of credit and financial collapse. *Quarterly Journal of Economics* 101 (3): 455–470.

Milde, H., and J. Riley. 1988. Signaling in credit markets. *Quarterly Journal of Economics* 103 (1): 101–130.

Mussa, M., and S. Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory* 18 (2): 301–317.

Riley, J. 1987. Credit rationing: A further remark. *American Economic Review* 77 (1): 224–227.

Rothschild, M., and J. Stiglitz. 1970. Increasing risk: A definition. *Journal of Economic Theory* 2 (3): 225–243.

———. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90 (4): 629–649.

Smith, V. L. 1972. A theory and test of credit rationing: Some generalizations. *American Economic Review* 62: 66–76.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.

———. 1983. Incentive effects of terminations: Applications to the credit and labor markets. *American Economic Review* 73 (5): 912–927.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21: 265–293.

Wette, H. 1983. Collateral in credit rationing in markets with imperfect information: A note. *American Economic Review* 73 (3): 442–445.

Williamson, S. D. 1987. Costly monitoring, loan contracts, and equilibrium credit rationing. *Quarterly Journal of Economics* 102 (1): 135–145.

# 6 The Macroeconomic Consequences of Financial Imperfections

It may seem odd to devote a chapter of a microeconomics book to macroeconomic issues. But in the last decade this traditional view has been partly revised. There have been promising developments in the theoretical literature on the macroeconomic implications of the *same financial imperfections* that are studied in detail in this book and that have been used to explain the role of banks and financial intermediaries (see Gertler 1988 for a first overview). Although this theoretical literature has not yet stabilized because it has not been convincingly supported by empirical evidence, it is important to be aware of some of its results.

In order to reduce the number of markets, macroeconomic modeling aggregates the different type of financial claims into a *well-functioning* and *unique* market for loanable funds. Consequently, the financial structure is not explicitly taken into account in the hope that it will be irrelevant at a macroeconomic level. What really matters is the aggregate amount of funds available to the economy.

There are two reasons that this approach may run into difficulties. First, the market may not be well functioning; it might face all sorts of market imperfections once we leave the Modigliani-Miller, Arrow-Debreu, complete contingent markets paradigm, where infinite-lived agents face a unique intertemporal budget constraint. Incomplete markets can take many forms, covering the overlapping generations, rational expectations, cash-in-advance models as well as asymmetric information models. Second, the market for loanable funds comprising equity, bonds, and bank loans may aggregate markets that react in completely different ways to economic shocks. Supply and demand in the aggregate loanable funds market will then not behave according to the standard model because what we call supply, demand, and price have not been well defined.

The effect of ignoring the financial structure by using a unique loanable funds market will, of course, be more prominent for some macroeconomic issues than for others. Indeed, the existence of financial market imperfections implies that the perfect insurance characteristic of the Arrow-Debreu model vanishes. We therefore expect that shocks to the economy will have a stronger effect than predicted by the Arrow-Debreu paradigm. This is particularly relevant in two areas: business cycles and monetary policy transmission.

The effect of ignoring the differences among the various types of financial claims, and disregarding the role of monitoring and relationship banking in creating value to the firm, should be relevant in the study of growth as well as in the analysis of the transmission channels of monetary policy.

**Business Cycles and Financial Structure**

Financial structure may have an impact on business cycles. This could come from the existence of a *financial accelerator* or a financial propagation effect. The effect would consist in an amplification of business cycles triggered by financial market imperfections, and in the limit, endogenous fluctuations could be generated by financial movements only. This type of effect was already put forward by Fisher (1933), with his theory of debt deflation as a possible explanation of the depth and length of the Great Depression.

Empirical evidence on the existence of financial market imperfections is pervasive, but one type of financial market imperfection is recurrently found, the *external finance premium*. This premium reflects the existence of a wedge between a firm's own opportunity cost of funds and the cost of funds obtained from an external source. Because of informational asymmetries, external funds may only be available at a premium. This fact is well known to macroeconomists. It implies that a firm's inventory and investment decisions will be sensitive to its level of cash reserves, which could contribute to increasing the magnitude of business cycles.

**Monetary Policy Transmission and Financial Structure**

The empirical analysis of the transmission channels of monetary policy has fostered a debate opposing the so-called money view as well as a set of alternative theories referred to as the broad credit channel. These theories are presented in section 6.2. Suffice to say here that empirical evidence has shown (1) that credit variables could help explain business cycles, (2) that a magnitude puzzle exists positing a large effect of monetary policy on output (Bernanke and Gertler 1995), and (3) that banks' liquidity position is essential in order to understand how banks react to a monetary contraction/expansion (Kashyap and Stein 2000).

After providing a short historical perspective on the macroeconomic consequences of financial markets' imperfections (section 6.1), this chapter studies

- the transmission channels of monetary policy (section 6.2);
- the fragility of the financial system (section 6.3);
- the effect of financial intermediation on growth (section 6.4).

## 6.1 A Short Historical Perspective

In the first issue of *Econometrica*, Irving Fisher (1933) argued that the severity of the economic downturn during the Great Depression resulted from the poor performance of financial markets. He defined the concept of *debt deflation*: when borrowers (firms) are highly leveraged, a small shock that affects their productivity or their net wealth can trigger a series of bankruptcies, which generate a decrease in investment, and thus in demand for intermediate goods, and as a consequence, in prices. This aggravates the real indebtedness of the productive sector, which may provoke a further series of failures, with a cumulative effect.

This viewpoint was later reinforced by the Gurley-Shaw (1955) theory, according to which *financial intermediaries* play a critical role in facilitating the circulation of loanable funds between savers and borrowers. Also in line with this view is the finding by Goldsmith (1969) that a positive correlation exists between economic growth and the degree of sophistication and development of the financial sector.

Following the publication by Friedman and Schwartz (1963) of their monetary history of the United States, the idea that *money supply* was the key financial aggregate gained wide support. Friedman and Schwartz found a high positive correlation between money supply and output, especially during the Great Depression. They argued therefore that banks did matter insofar as they create money. This is in line with the conclusions of simple IS/LM macro models in which the money supply is assumed to be completely controlled by the Central Bank—an oversimplifying assumption. In fact, even if the Central Bank can control the money base, the other components of the money supply adjust to changes in interest rates. As a consequence, the stock of money is in fact less important for macroeconomic performance than the *financial capacity* of the economy, defined as the aggregate volume of credit that lenders are ready to grant to borrowers. Therefore, in response to Friedman and Schwartz's money view, the alternative position was to emphasize the credit view.

After the 1960s, and following Modigliani and Miller's (1958) contribution, the view that "finance is a veil" became widely accepted. If the financial structure of firms is irrelevant, and if financial intermediaries are redundant, then monetary policy can have only a transitory effect on real variables, through unanticipated changes in the money supply. In all the *real business cycle models* that were developed subsequently, finance does not play any role.

The comeback of financial aspects in macroeconomic models started in the early 1980s. Following an earlier study by Mishkin (1978), Bernanke (1983) analyzed the relative importance of monetary versus financial factors in the Great Depression. His central conclusion was that monetary forces alone were quantitatively insufficient to explain the Depression's depth and persistence, and that the collapse of the financial system (half of U.S. banks failed between 1930 and 1933, and the financial markets crashed worldwide) was an important factor. Therefore, the decline in the money stock seems in fact to have been less important than argued by Friedman and Schwartz. Bernanke tested the two explanations (the breakdown in banking having affected borrowers who did not have access to security markets versus the decline in money supply) and concluded in favor of the first. Thus, this piece of empirical evidence gave support to the credit view, which argued that financial markets were imperfect, so the Modigliani-Miller assumptions did not hold, and finance did matter.

The following sections study the different theoretical arguments that support this view, starting with those that aim to explain the transmission channels of monetary policy.

## 6.2 The Transmission Channels of Monetary Policy

Contemporary monetary theory acknowledges that monetary policy produces effects only in the short run and is neutral in the long run. Thus monetary policy can be used to smooth out fluctuations in economic activity but not to increase the rate of growth of GDP.

The basic approach to monetary policy assumes short-run price rigidities, an assumption we maintain throughout this discussion. This assumption is needed to understand both the effects of monetary policy and why these effects taper off once prices start to adjust.

Still, the way in which monetary policy acts upon real variables is not yet completely understood. In spite of the spectacular progress in this field since the days of the Keynes versus Friedman debate, different channels through which transmission could occur have been conjectured, but there is no general agreement on which ones are dominant and how they interact. In the debates on the monetary transmission mechanisms, one issue has dominated the scene: How relevant should the role of the banking system be?

That the banking system has a key role in the transmission of the monetary policy looks like a truism. Traditional theory has acknowledged this but has assumed that explicit modeling of banking activity was unnecessary. This is the case once we view the main role of banks as money creation. From that perspective, only a money or loanable funds market is required in order to model the effects of monetary policy.

This traditional view, which we call the *money view*, was confronted with important puzzles that consistently emerged from the empirical analysis but were left unexplained. Bernanke and Gertler (1995) report three main puzzles that make the money view an unsatisfactory explanation:

• The *magnitude puzzle* is concerned with the fact that the real economy is highly affected by policy innovations, whereas the effect on interest rate is relatively small. This is confirmed by computing the interest rate elasticities of the different components of GDP. The total effect of monetary policy innovations predicted by adding up the different elasticities is smaller than the actual one.

• The *timing puzzle* must also be explained. The money view predicts that since the interest rate is the leading force behind the real effects, once interest rates go back to normal their effect should stop. Yet empirical evidence shows that there is an important lag and that "some important components of spending do not begin to react until after most of the interest rate effect is past" (Bernanke and Gertler 1995, 34).

• The *composition puzzle* states that the changes in the structure of spending do not correspond to the money view predictions. This is the case because monetary policy has its effect on short-term rates, whereas the main effect (the most rapid and the larger percentage) is on real estate investment, which should rather react to long-term real interest rates.

As a consequence, when Bernanke (1983) established that the Great Depression was better explained when the banking system was introduced in addition to the money aggregate, this led to a new perspective on the channels of monetary policy by suggesting the existence of a *lending view*. Since then, the debate has been enriched by the introduction of the *balance sheet channel*, which considers the effect of monetary policy on the value of the firms' assets and therefore on their capacity to borrow from banks.

In any case, the debate has put the banking system in the foreground of monetary policy analysis. This is why we now turn to examine the effects of financial imperfections on the transmission of monetary policy.

For the sake of completeness, the different channels of monetary policy are defined first. Then a framework for the analysis of investment financing is introduced. Next, the implications of the different effects at the macroeconomic level are derived, and finally, we conclude with a review of some empirical findings that back up the different views.

### 6.2.1 The Different Channels

A number of channels through which monetary policy operates have been identified at the theoretical level and tested.

**Interest Rate Channel**

This channel affects consumption and investment through five different macroeconomic variables:

- Income (through net interest payments)
- Wealth (through capital gains or losses on assets)
- Consumption
- Cost of capital and investment
- Exchange rate

**Broad Lending Channel**

This channel takes into account the role of the banking system in a context of asymmetric information. Thus it considers the external finance premium, defined as the wedge between the cost of funds raised externally and the opportunity cost of internal funds, as an essential key in the understanding of the transmission mechanism. The existence of an external finance premium has been consistently confirmed in the empirical literature on firms' investment. The existence of such a premium may be the result of imperfect information, of credit rationing resulting from the scarcity of bank capital (Holmström and Tirole 1997), or adverse selection in the capital markets (Bolton and Freixas 2000). The external liquidity premium will be affected by monetary policy, amplifying the effect on interest rates and generating the so-called financial accelerator effect.

The broad lending channel operates through two subchannels:

- *The lending channel*, because banks with liquidity shortages will lend less
- *The balance sheet channel*, because banks react to a decrease in the value of firms collateral by cutting down the amount lent to firms

Two necessary conditions have to be satisfied for the lending channel to operate. First, bank loans and market finance (e.g., commercial paper, bonds) have to be imperfect substitutes, a point that is central to our book and that we examined in chapter 2. Otherwise, the lack of bank lending would be compensated by firms through the issue of securities in financial markets.

Second, banks have to react to liquidity shortages by cutting down their lending. In other words, if banks react to a restrictive monetary policy simply by issuing additional certificates of deposit, the bank lending channel cannot operate.

### 6.2.2 A Simple Model

Since our objective here is not to consider the short- and long-term macroeconomic effects of monetary policy, but to show the impact of asymmetric information, we use the simple static model of Bernanke and Gertler (1995) to illustrate the effect of a collateral-driven credit constraint. This is by no means the most general way to model the effect of asymmetric information on financial markets. Nevertheless, it conveys the main intuitions in a simplified way.

Consider a firm with a production function $f(\cdot)$, which transforms input $x$ into output $f(x)$. For simplicity, we take the prices of the input and the output equal to 1.

The firm requires external finance in the form of a loan $L$ in order to buy the input. Denoting by $W$ the firm's wealth, the amount of input the firm is able to buy is $x = L + W$.

Consider the case of perfect financial markets. Given the interest rate on loan $r$, the firm will choose $L$ to maximize its profit:

$$\max_L \; f(L+W) - (1+r)L, \tag{6.1}$$

leading to the first-order condition

$$f'(L+W) = 1 + r. \tag{6.2}$$

In this case, investment $L + W$ and output $f(L + W)$ only depend on the interest rate.

Consider now the case of strong informational asymmetries, where the firm is only able to obtain a fully collateralized loan. In this case, the value of the loan $L$ has to satisfy the constraint $(1 + r)L \leq qK$, where $K$ denotes the volume of the firm's assets and $q$ their price. This inequality expresses the constraint that the collateral has to guarantee the principal of the loan plus the interest.

As a consequence, the firm's decision problem becomes

$$\max_L \; f(L) - (1+r)L, \tag{6.3}$$

$$L \leq \frac{qK}{1+r}. \tag{6.4}$$

When the constraint on collateral is binding, the first-order condition becomes

$$f'(L+W) = 1 + r + \lambda, \tag{6.5}$$

$$L = \frac{qK}{1+r}, \tag{6.6}$$

$$\lambda \geq 0. \tag{6.7}$$

Comparing (6.2) with and (6.5), (6.6), and (6.7) allows discussion of the implications of the broad credit channel.

The first-order condition (6.2) reflects the money view. Interest rates equal the firm's marginal cost of capital, and there is no external finance premium. Changes in $q$ and $K$ leave unaffected the firm's investment and output. A decrease in $W$ is matched one-to-one by an increase in $L$, and consequently it does not affect either the firm's investment $L + W$ or its output $f(L + W)$.

When the firm faces a collateral constraint, the results are strikingly different. First, notice that

$$f'\left(\frac{qK}{1+r} + W\right) > (1+r),$$

which means that there is an external finance premium $\lambda > 0$. A reduction in the value $qK$ of the firm's assets decreases the amount of the loan and thus increases the external finance premium. The effect of an increase in the interest rate is more complex than in the money view. In addition to its direct effect of increasing the cost of capital, it is likely to decrease the price of the firm's asset, $q$, thus decreasing further the borrowing capacity of the firm and its investment.

Second, the effect of a decrease in the firm's wealth is to decrease the firm's input expenditure by the same amount. This increases the external finance premium and decreases the firm's output $f(L+W)$.

Since the wealth $W$ of the firm results from previous-period conditions, there are important feedback effects: an increase on last-period interest rates decreases today's wealth, thus leading to a persisting effect.

As already mentioned, the fully collateralized loan assumption is neither the unique nor the most general way to model asymmetric information in financial markets. The effect of an external finance premium is basically the same whether we use this reduced model or the more complex models of Bernanke and Gertler (1990), Holmström and Tirole (1997), or Bolton and Freixas (2000).

### 6.2.3 Credit View versus Money View: Justification of the Assumptions and Empirical Evidence

The assumptions implicit in the money view are as follows:

A1: Prices do not adjust instantaneously to offset changes in the (nominal) quantity of money.

A2: The Central Bank can directly influence the volume of *money* by adjusting bank reserves.

A3: Loans and securities are *perfect* substitutes for borrowers and for banks.

The assumptions implicit in the credit view are as follows:

A1: Prices do not adjust instantaneously to offset changes in the (nominal) quantity of money.

A′2: The Central Bank can directly influence the volume of *credit* by adjusting bank reserves.

A′3: Loans and securities are *imperfect* substitutes for borrowers and for banks.

Assumption A1 is common to both views of monetary policy and is not discussed further.

The early justifications of A′2 (the so-called availability doctrine of Roosa 1951) relied on credit rationing, but this was not fully satisfactory because it left unexplained why banks themselves, with their privileged access to liquid markets, had to restrict their supply of funds as a response to a contractionnary monetary policy operation. This, of course, need not be explained in the money view, since it is the total amount of loanable funds (bank loans and securities issued), not their distribution, that matters. Modern arguments (related to imperfect information and the monitoring role of banks) explain assumption A′2 without the credit rationing assumption.

Nevertheless, even if one accepts the view that loans and securities are not perfect substitutes for borrowers, banks could have liabilities that are perfect substitutes. If this were the case, banks could react to, say, a shock on their deposits by issuing another type of liability and leaving their level of lending unchanged. For instance, Romer and Romer (1990) argue that the supply of large-denomination certificates of deposit (CDs) (which are not subject to reserves) addressed to any given bank is perfectly elastic at the current market rate. However, this is not confirmed by casual empiricism: a non-negligible interest spread exists between CDs and Treasury bills (T-bills) of the same characteristics. Moreover, this spread appears to be strongly dependent on banks' ratings given by the rating agencies. Cook and Rowe (1986; quoted by Kashyap and Stein 1993) give the example of Continental Illinois, which experienced a large increase in its CD rates before going bankrupt. The theoretical answer to Romer and Romer's criticism is that a bank's liabilities themselves are not perfect substitutes, so a decrease in fully insured deposits cannot be matched, say, by CDs or by issuing new equity. Stein (1998) provides a well-articulated model of limited liability substitution under adverse selection, where good banks signal themselves in equilibrium by having a limited access to the CD market. As a consequence, in the unique separating equilibrium, good banks lend less and are more sensitive to the availability of insured deposits. When faced with a monetary contraction, good banks prefer to cut down on their supply of credit rather than raise risk-sensitive funds, as predicted by the credit view.

In the presence of banks' capital regulation, the binding capital constraint is another reason why the supply of bank credit will react to monetary policy. Van den Heuvel (2002) assumes simply a maturity mismatch between banks' assets and liabilities, so that the maturity of the liability is shorter than that of the assets, a standard assumption justified by Diamond and Dybvig's (1983) classical model, among others. When this is the case, if it is costly to issue additional equity, a capital channel appears, working through the following sequence:

1. Monetary policy increases (resp. decreases) interest rates.
2. Because of the banks' maturity mismatch this generates a loss (resp. a profit).
3. This, in turn, produces a capital decrease (resp. increase).
4. The binding capital requirement constraint leads to a reduction (resp. increase) in bank lending.

Assumption A′3 has been strongly justified in this book. Firms having insufficient capital or reputation cannot issue direct debt, and therefore they rely on credit from financial intermediaries that establish strong relationships with them (see chapter 2). This is confirmed by empirical evidence. James (1987) shows that bank credit is more expensive than direct debt, which indicates that there is something special in bank services to borrowers. Similarly, Hoshi, Kashyap, and Scharfstein (1991) show that Japanese firms with close banking ties are less likely to be liquidity-constrained. However, as pointed out by Kashyap and Stein (1993), financial intermediaries other than banks (such as finance companies, which do not hold Central Bank reserves) could in principle provide the same services. Even if these nonbank intermediaries do not appear to have a large share of the credit market, they could effectively be the marginal lenders in the economy and supply credit to the economy when the Central Bank restricts liquidity, thus undermining assumption A′2. This does not seem to be the case, probably because of the lock-in effect of relationship banking (Sharpe 1990; Rajan 1992; Slovin, Sushka, and Polonchek 1993). Because of informational asymmetries, borrowers cannot switch from one lender to another without cost.

Notice, though, that assuming imperfect substitution does allow for some degree of substitution on behalf of the firms. This is clearly illustrated in models of heterogeneous firms like Holmström and Tirole (1997) or Bolton and Freixas (2000). Monetary policy shocks will marginally affect the population of firms that have access (or prefer) a bank loan and those that go for market funds.

The empirical evidence, provided by the analysis of Kashyap, Stein, and Wilcox (1993), establishes that this substitution of liabilities at the firm level does indeed take place. Their results show that a tightening of the monetary policy by the Federal Reserve raises the issue of commercial paper and lowers the amount of loans.

Additional empirical evidence on the imperfect substitutability between loans and securities (here essentially T-bills), documented by Bernanke and Gertler (1987), is that banks use T-bills as a buffer against liquidity shocks. This is confirmed by the fact that large banks hold significantly fewer T-bills than the average (Kashyap and Stein 1993).

### 6.2.4 Empirical Evidence on the Credit View

We conclude with a look at some empirical work on the issue of credit view versus money view. First, favoring the credit view, strong evidence seems to exist in favor of a high correlation between credit supply and economic activity. As already mentioned, Bernanke's (1983) influential study of the Great Depression in the United States attributes the depth and persistence of the Depression to the crisis experienced by the U.S. banking sector at that time. Similar conclusions are obtained for other periods and other countries (Bernanke and James 1991; Schreft 1990; Bernanke 1986).

A second important empirical issue is whether monetary policy really influences credit supply (assumption A′2) more than money supply (assumption A2). Several studies (King 1986; Romer and Romer 1990; Ramey 1992) have found that loans adjust gradually to changes in monetary policy, but that money changes more rapidly and is correlatively a better predictor of output. Kashyap, Stein, and Wilcox (1993) go further and examine the impact of monetary policy on the composition of firms' borrowing (its allocation between bank loans and commercial paper). They find that an episode of restrictive monetary policy is typically followed by a raise in commercial paper issuance and a decline in bank loans. Using cross-sectional data, Gertler and Gilchrist (1992) show that in such a case, external financing of large firms actually *increases* (both commercial paper and bank loans) and that it is the small firms that pay the burden, since they experience a large decrease in bank loans, which are essentially their only source of external finance.

Finally, cross-country comparisons also provide interesting insights in the money view versus credit view debate. As stated by Mihov (2001, 382), one of the main implications of the credit view would be that in more bank-dependent economies, where firms and banks have fewer possibilities of restructuring their liabilities in response to a business cycle or monetary policy shock (as argued by Kashyap, Stein, and Wilcox for firms and by Romer and Romer for banks), the effect of monetary policy will be larger. Mihov's empirical results, although based on a limited number of countries, confirm this idea. Cumulative deviations of output generated by the same monetary policy shock are larger for countries with a larger ratio of bank loans to total liabilities.

## 6.3 Financial Fragility and Economic Performance

The traditional view of the macroeconomic neutrality of the financial sector has also come under attack from another front. Indeed, asymmetric information generates imperfections in financial markets and therefore an inefficient channeling of resources to investment. This inefficiency may also have real effects on aggregate macroeconomic activity. This is studied by Bernanke and Gertler (1990) in a model related to Boyd and Prescott (1986). They consider a one-good, one-period general equilibrium model with an infinite number of risk-neutral agents. Agents can be entrepreneurs (in proportion $\mu$) or households (in proportion $1-\mu$). All of them have access to a riskless technology that yields $1+r$ units at time $t=1$ out of one unit at time $t=0$.

The average initial endowment is normalized to 1, but that of entrepreneurs ($\omega_e$) is less than 1. Since investment requires one unit of the good, they will have to borrow from households.

Each entrepreneur owns a risky technology: one unit of the good at $t=0$ yields $y$ units (at $t=1$) with probability $p$, and zero with probability $1-p$. The initial wealth $\omega$ of each entrepreneur (with $E(\omega)=\omega_e$) is publicly observable, but $p$ may be unknown a priori. It is only privately observable by the entrepreneur after an evaluation (screening) that costs him an effort whose monetary equivalent is $C$. After the screening has been made, the entrepreneur decides to undertake the project (invest 1) if $p$ is large enough. This discussion assumes that $E(p)y<1+r$, so that absent screening, it is efficient to use the storage technology.

**First Best Allocation (Perfect Information)**

We consider as a first benchmark the best allocation that would obtain if the result of the entrepreneur screening (the value of $p$) could be publicly observed. Note that screening a project (which costs the entrepreneur $C$) gives an option on future investment.

If the firm screens, its expected profit is $E_p[\max py,(1+r)]$ to be compared with a sure return $1+r$ if it does not screen. Consequently, the value of screening is $V=E[\max(py-1-r,0)]$, which indeed looks like an option value. Therefore, screening will take place (for all projects) if and only if

$$C<V\overset{\text{def}}{=}E_p[\max(0,py-1-r)], \tag{6.8}$$

which is assumed to hold true. As a consequence, all projects will be screened.

Among the projects that are screened, only those with a positive expected excess return will be undertaken ($py>1+r$). This gives the cutoff probability $p^*$ under which projects are not financed:

$$p^*=\frac{1+r}{y}. \tag{6.9}$$

Clearly, which projects are undertaken in the best allocation does not depend on the distribution of initial endowments.

We introduce the following notation: $h(p)$ denotes the density function of $p$ on $[0,1]$, and $H(p)$ its cumulative distribution. $A(p_0)=E[p|p\geq p_0]$ denotes the average success probability conditionally on $p\geq p_0$.

In the first best allocation, aggregate economic variables are given by

$$I^*=\mu(1-H(p^*)) \quad \text{(investment)},$$

$$q^*=1+r+\mu(V-C) \quad \text{(output, net of screening costs)}.$$

The investment of firms is financed in part by their own wealth $\mu\omega_e(1 - H(p^*))$ and in part by households' savings $S^*$, which equal the firms' demand for funds:

$$S^* = \mu(1 - \omega_e)(1 - H(p^*)).$$

The rest of the initial endowment is stored: $\mu\omega_e H(p^*)$ by firms and $\{1 - \mu(1 - H(p^*))\}$ by households. If the entrepreneurs have sufficient endowments $\omega_e \geq 1$, the first best allocation is reached even if $p$ is only privately observed. This is so because in this case the entrepreneurs do not need to borrow.

**Credit Constraints and Limited Liability**

Assume now that $p$ is only privately observable and that (at least some fraction of) entrepreneurs cannot self-finance their projects ($\omega < 1$). These entrepreneurs have to find a lender (a household) to fund the remaining part $(1 - \omega)$ and sign with the lender a contract specifying the repayment $R$ in case of success.[1] The lender knows that because of limited liability there will be no repayment in case of failure. It is assumed that the contract is signed after screening takes place, that is, after the borrower observes $p$, but borrowers cannot credibly communicate the value of $p$. The timing is therefore as shown in figure 6.1.

In this setup, a loan contract is completely described by the amount $(1 - \omega)$ to be lent and the amount $R$ (which will depend on $\omega$) to be repaid in case of success. The characteristics of the contract determine the minimum cutoff probability $\hat{p}(\omega)$ required by the entrepreneur to implement the project. It is given by

$$(y - R(\omega))\hat{p}(\omega) = (1 + r)\omega. \tag{6.10}$$

The equilibrium contract is jointly determined by (6.10) and by the zero-profit condition for lenders:[2]

$$A(\hat{p}(\omega))R(\omega) = (1 + r)(1 - \omega). \tag{6.11}$$

The option value of the screening technology becomes

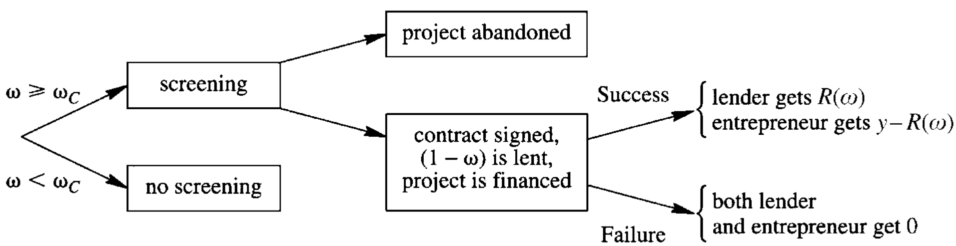



**Figure 6.1**
Timing in Bernanke-Gertler (1990) model.

$$V(\omega) = E_p[\max(0, p(y - R(\omega)) - (1+r)\omega)]$$
$$= \int_{\hat{p}(\omega)}^{1} \{(p(y - R(\omega) - (1+r)\omega)\} h(p)\, dp.$$

Using (6.11), $R(\omega)$ can be eliminated, and $V(\omega)$ appears to be equal to the expected surplus for $p$ above the cutoff point $\hat{p}(\omega)$:

$$V(\omega) = E_p[(py - (1+r))\mathbb{I}_{\{p > \hat{p}(\omega)\}}] = \int_{\hat{p}(\omega)}^{1} (py - (1+r))h(p)\, dp. \tag{6.12}$$

**Result 6.1** The consequences of moral hazard and limited liability are as follows:

1. Entrepreneurs take too much risk:

$$\hat{p}(\omega) \le p^* = \frac{1+r}{y}$$

with equality when $\omega = 1$.

2. There is a critical value $\omega_C$ of firms' wealth under which projects are not screened. $\omega_C$ is defined implicitly by $V(\omega_C) = C$.[3] Firms with $\omega < \omega_C$ will prefer to invest in the storage technology.

3. The nominal interest rate

$$r(\omega) \stackrel{\text{def}}{=} \frac{R(\omega)}{1-\omega} - 1$$

is a decreasing function of $\omega$.

***Proof*** When $\omega = 1$, equation (6.11) implies $R(1) = 0$, and (6.10) gives that $\hat{p}(1) = p^*$. All the properties stated in result 6.1 are consequence of the fact that $\hat{p}(\omega)$ is a nondecreasing function of $\omega$, a property that is proved later.

1. $\hat{p}(\omega) \le \hat{p}(1) = p^*$, or equivalently, $\hat{p}(\omega)y - 1 - r < 0$.
2. Using (6.12),

$$\frac{dV}{d\omega} = -h(\hat{p}(\omega))(\hat{p}(\omega)y - 1 - r)\frac{d\hat{p}}{d\omega} > 0$$

since $\hat{p}(\omega)y - 1 - r < 0$, and $V(1) = V > C$ by assumption.

Now, for a firm with endowment $\omega$ such that $\omega < \omega_C$, the project will not be screened, and the expected return from implementing the project will be $E(p)y$, inferior to $1 + r$. Lending to the firm against a promised repayment $R(\omega)$ may be profitable only if

$$\frac{E(p)R(\omega)}{1-\omega} \geq 1+r.$$

But in this case the firm will obtain a return from the project inferior to $1+r$.

3. Using (6.11) and remembering that $p \to A(p)$ is increasing,

$$r(\omega) = \frac{1+r}{A(\hat{p}(\omega))} - 1.$$

To establish that $\hat{p}(\omega)$ is indeed increasing, apply the implicit function theorem to the system (6.10) and (6.11):

$$\begin{cases} [y - R(\omega)]\dfrac{d\hat{p}}{d\omega} - \dfrac{dR}{d\omega}\hat{p} = 1+r, \\ A'(\hat{p})R\dfrac{d\hat{p}}{d\omega} + A(\hat{p})\dfrac{dR}{d\omega} = -1-r. \end{cases}$$

Multiplying the first equation by $A(\hat{p})$ and the second by $\hat{p}$, and adding them, yields

$$\{A(\hat{p})\underbrace{[y - R(\omega)]}_{>0} + \underbrace{\hat{p}A'(\hat{p})R}_{>0}\}\frac{d\hat{p}}{d\omega} = (1+r)\underbrace{[A(\hat{p}) - \hat{p}]}_{>0}. \tag{6.13}$$

Therefore

$$\frac{d\hat{p}}{d\omega} > 0. \qquad \blacksquare$$

To obtain a better understanding of what underlies result 6.1, consider the benchmark case in which there is perfect information on $p$, so that the repayment, $R(\omega, p)$, is made contingent also on this variable. As expected, the first best allocation is obtained, and the cutoff probability is then $p^*$ for all $\omega$. Notice that a higher $p$ implies lower repayments, so that $R(\omega, 1) < R(\omega, p) < R(\omega, p^*)$ for $p > p^*$. When $p$ is not publicly observable, these different payments are replaced by their average $R(\omega)$, which results in taxing safer firms and subsidizing riskier ones. As a consequence of this subsidy (due to adverse selection), riskier firms are willing to implement the project even for low (inefficient) values of $p$, that is, in the range $(\hat{p}(\omega), p^*)$, thus increasing the cost of funds for all firms. Since the market imperfection is related to the entrepreneur's lack of endowment, the lower the level of endowment $\omega$, the larger the effect, which explains points 2 and 3 in result 6.1.

**Macroeconomic Implications**

This section discusses the macroeconomic consequences of result 6.1 by computing aggregate macroeconomic variables of this economy. Let $F(\omega)$ denote the distribution function of the firms' wealth, with associated density $f(\omega)$ and support $[\omega_0, \omega_1]$. The interesting case is when

$$\omega_0 < \omega_C < \omega_1,$$

that is, when there is a group of firms (those with $\omega$ between $\omega_0$ and $\omega_C$) that are actually credit-constrained.

The expected output (net of screening costs) of a firm of wealth $\omega$ is

$$q(\omega) = \begin{cases} (1+r)\omega & \text{if } \omega < \omega_C, \\ (1+r)\omega + V(\omega) - C & \text{if } \omega > \omega_C. \end{cases}$$

The aggregate economic variables (per capita) are

$$I = \mu \int_{\omega_C}^{\omega_1} (1 - H(\hat{p}(\omega)))\, dF(\omega) \qquad \text{(investment)},$$

$$S = \mu \int_{\omega_C}^{\omega_1} (1 - \omega)(1 - H(\hat{p}(\omega)))\, dF(\omega) \quad \text{(households' savings)},$$

$$q = 1 + r + \mu \int_{\omega_C}^{\omega_1} (V(\omega) - C)\, dF(\omega) \qquad \text{(output)},$$

($\mu$ denotes the proportion of entrepreneurs), to be compared with their first best levels:

$$I^* = \mu(1 - H(p^*)),$$

$$S^* = \mu(1 - \omega_e)(1 - H(p^*)),$$

$$q^* = 1 + r + \mu(V - C).$$

Therefore, it appears that the global performance (output $q$ and investment $I$) of this economy does not depend only on the fundamentals of investment ($p^*$, $\mu$, $V$, and $C$) but also on the financial situation of firms (captured here by the distribution of their initial wealth $\omega$). In particular, when many entrepreneurs have low wealth (when $F(\omega_C)$ is close to 1, or $\omega_e$ close to $\omega_C$), investment and output will be low even if the fundamentals are good. This situation is described by Bernanke and Gertler as financial fragility. Moreover, if there is a shock on the distribution of $\omega$ such that $\omega_e$ falls below $\omega_C$, a *collapse* of investment will occur as a result of the poor financial condition of firms.

Bernanke and Gertler discuss the policy implications of their results. For instance, if the types of agents are not observable (the case discussed here), a tax on successful investment projects (used to subsidize households) will be welfare-improving. This is because entrepreneurs tend to invest in too many projects ($\hat{p}(\omega) < p^*$). Therefore, by reducing the profitability of investments, a tax makes entrepreneurs more selective.

However, this positive effect is counterbalanced by a reduction of the screening activity because the value of the option is reduced by the tax. In the case in which types are observable, a more interesting welfare-improving policy consists of subsidizing entrepreneurs by taxing households. This could be interpreted as bailing out debtors (as in the less developed countries debt crisis) or as lending money to illiquid entrepreneurs. Bernanke and Gertler extend the interpretation of their results to justify the lender-of-last-resort policy of the Central Bank, aimed at protecting financial institutions from liquidity shocks.

## 6.4 Financial Development and Economic Growth

The idea that financial structure is a key factor of economic development can be traced back at least to Schumpeter (1934) and Gerschenkron (1962), when confronting cross-country experiences in economic development (see Da Rin and Hellmann 2002). This issue is critical because it may guide economic policy in developing countries. Consequently, it has been the object of many contributions and has taken different forms as progress in empirical research has allowed defining the issues more rigorously.

Examining this issue is complex, in the first place because the link between higher levels of economic development and more sophisticated financial systems might result from reverse causality. That is, the correlation between growth and the level of financial development might simply reflect that countries with higher income consume more financial services. Yet there is a clear theoretical argument for the opposite causality, asserting that financial development fosters economic growth because more sophisticated financial systems allow limiting financial market imperfections and channeling funds to the best investment opportunities. The models we have examined so far focus on this point: both Bernanke and Gertler (1990) and Holmström and Tirole (1997) make a clear link between the level of financial imperfections and the level of aggregate production. In a dynamic framework, a lower cost of screening (Bernanke and Gertler) or a higher level of monitored finance (Holmström and Tirole) would therefore lead to a higher rate of growth. More generally, Levine (2005) identifies five channels through which financial intermediaries may affect growth: providing ex ante information, monitoring investment, managing risk better, mobilizing savings, and facilitating the exchange of goods and services.

The first empirical contribution analyzing cross-country effects is due to King and Levine (1993), who established that the level of financial development was a good predictor of future economic growth. The results are confirmed when cross-region analysis is performed within a country (see Guiso, Sapienza, and Zingales 2004 for Italy). This was a way to circumvent the reverse causality issue. Yet there are a number of possible objections; financial development could be related to, or even caused by, other correlated variables. Still, the effect could be produced by a missing variable that is directly correlated to financial development. Rajan and Zingales (1998) take an alternative approach toward finding the real cause. They test whether the effect of financial development on growth is larger for firms that are more dependent on external finance. Their results confirm the effect of financial development on growth. Thus both King and Levine's and Rajan and Zingales's methodologies point at the same (robust) evidence: financial development is a key factor of economic growth.

The next step in the exploration of the financial development basis for growth was to study which specific aspects of financial development were crucial for economic development. This meant studying how different types of financial development fit different types of economic environments in order to foster growth. On this point, the key issue is the distinction between the respective roles of banks and financial markets. Securities markets provide informational feedback because they provide prices for the different types of assets. This may allow firms to make better decisions, as in Boot and Thakor (1997). On the other hand, as pointed out in chapters 2–4 of this book, banks perform specific functions: monitoring, providing liquidity insurance, smoothing intertemporal shocks, developing relationships with firms, allowing for a richer set of contracts. In addition, banks in a noncompetitive market structure may also have a role in coordinating investment in "big-push" type of models, where firms' productivity depends on the aggregate amount of investment, as suggested by Da Rin and Hellmann (2002).

Thus, a question arises as a refinement and extension of the initial evidence: Is a bank-based financial structure better than a market-based one to foster development? This may depend on the environment. It is quite possible that in environments with low contractual enforcement and limited legal protection, or in an economy with small firms and illiquid securities markets, bank monitoring performs better than arm's-length lending, whereas in the opposite type of environments market-based financial structures perform better. Alternatively, the financial services view states (Levine 2002) that the key issue is the level and quality of financial services, not the channel, banks, or markets through which these services are provided. Finally, it could be the case that the main determinant of financial development is the legal enforcement environment, irrespective of whether the financial system is bank-based or market-based. This is the view of La Porta et al. (2000).

The empirical evidence is somewhat controversial. On the one hand, the legal environment seems to be a clear factor of economic development. Thus, Beck, Levine, and Loayza (2000) establish two facts: development of the banking sector exerts a positive effect on growth, and legal and accounting systems (such as creditor rights, contract enforcement, and accounting standards) explain differences in the level of financial development.

Thus, according to Beck, Levine, and Loayza (2000), the causal link is from better contract enforcement and more rigorous accounting practices to better financial development, and from financial development to higher economic growth.

Also supporting the view of La Porta et al. (2000), Levine (2002) finds that "the legal system crucially determines financial development and that financial structure is not a particularly useful way to distinguish financial systems" (402). This is in line with the conclusion that Demirgüç-Kunt and Maksimovic (2002) reach when they analyze the determinants of firms' access to external finance.

Still, the role of financial structure appears relevant in Tadesse (2002). His results indicate that market-based countries outperform bank-based countries among those with developed financial systems, whereas the opposite is true for countries with less developed financial systems. Also, countries dominated by small firms grow faster with bank-based systems, a view quite consistent with the role of banks in monitoring and relationship lending.

Additional insight on the effect of financial development on growth can be achieved by investigating through what channels this effect operates. A number of contributions (see Papaioannou 2006) have succeeded in breaking the overall financial development–growth link into possible specific connections.

First, the rate of growth can be decomposed into two components: input growth and total-factor productivity (TFP) growth. Therefore, it is interesting to know whether the effect on growth comes from the release of additional savings into the productive process or from a better allocation of the same input to investment projects. The analysis of Jayaratne and Strahan (1996) on U.S. data shows that higher growth rates basically come from the quality of banking rather than from increased investment. In the same vein, Beck, Levine, and Loayza (2000) analyze the effect of financial development in a cross-country multiperiod setup and show that the main effect is on TFP; the long-run links between financial intermediary development and both physical capital growth and private saving rate are tenuous.

Second, if the source of the effect of financial development on growth is a better allocation of capital, this should have a number of testable implications. If we assume that the effect is brought about by competition, as initially argued by Schumpeter, more developed financial systems should be associated with a higher rate of entry and exit. Following this line of analysis, Bertrand, Schoar, and Thesmar (2007) showed that French banking deregulation led to increased firm return on assets in bank-dependent sectors and to increased entry and exit in finance-dependent industries. Similarly, Guiso, Sapienza, and Zingales (2004) showed that entry of new firms is much higher in financially developed regions of Italy.

A second testable implication is that financial development helps accommodate macroeconomic shocks. This has been shown to be the case both at the aggregate and the industry level. At the aggregate level, the key issue is to establish that the impact of volatility or exogenous macroeconomic shocks is mitigated by access to financial markets. This is tested by Aghion et al. (2005), who show that the interaction term between financial development and volatility is significant in explaining economic growth. In the same vein, Aghion et al. (2006) show that exchange rate volatility significantly reduces productivity growth in financially underdeveloped countries; this effect becomes positive above some level of financial development.

Building on Rajan and Zingales (1998), Cetorelli and Gambera (2001) show that both bank concentration and financial development promote the growth of finance-dependent sectors. Still, they also show that bank concentration has a direct negative impact on growth. At the industry level, Ciccone and Papaioannou (2006) show that financial development fosters the adjustment of capital investment to sectorial productivity shocks.

Finally, it is also interesting to consider the effect of financial development on the convergence or divergence of GDP per capita between countries. The idea here is that countries with technological backwardness have the opportunity to benefit from new technologies only if they have the necessary financial resources, whereas they will not be able to do so if they lack the adequate financial markets, in which case growth rates will diverge. The results show that the likelihood that a country will catch up with others increases with its financial development. Also, the effect of financial development for a country that catches up is positive but eventually diminishing, and the rate of growth of countries that fail to catch up increases with level of financial development.

So, overall, the case for a link between financial development and growth is quite strong, even if future research in this area needs to clarify some key issues.

## Notes

1. The cost of screening is assumed to be nonmonetary. That is, $C$ is the monetary equivalent of the cost of effort needed for screening projects.

2. Notice that when $\omega < 1$, $\hat{p}(\omega)$ is different from the best cutoff $p^* = (1+r)/y$. Indeed, if $\hat{p}(\omega)$ were equal to $p^*$, (6.10) would imply $p^* R(\omega) = (1+r)(1-\omega)$, which should be equal to $A(p^*)R(\omega)$, according to (6.11). Since $A(p^*) > p^*$, this is possible only when $R(\omega) = 1 - \omega = 0$, contradicting $\omega < 1$.

3. When $V(0) \geq C$, all projects are screened. This discussion focuses on the more interesting case in which $C > V(0)$.

## References

Aghion, P., G. M. Angeletos, A. Banerjee, and K. Manova. 2005. Volatility and growth: Credit constraints and productivity-enhancing investments. Mimeograph.

Aghion, P., P. Bachetta, K. Rogoff, and R. Ranciere. 2006. Exchange rate volatility and productivity growth: The role of financial development. NBER working paper 12117.

Beck, T., R. E. Levine, and N. V. Loayza. 2000. Finance and the sources of growth. *Journal of Financial Economics* 58 (1–2): 261–300.

Bencivenga, V. R., and B. D. Smith. 1991. Financial intermediation and endogenous growth. *Review of Economic Studies* 58 (2): 195–209.

Bernanke, B. 1983. Non-monetary effects of the financial crisis in propagation of the Great Depression. *American Economic Review* 73 (3): 257–276.

———. 1986. Alternative explanations of the money-income correlation. NBER working paper 1842.

Bernanke, B., and A. Blinder. 1987. Banking and macroeconomic equilibrium. In *New approaches to monetary economics*, ed. W. Barnett and K. Singleton. Cambridge: Cambridge University Press.

———. 1988. Credit, money and aggregate demand. *American Economic Review* 78 (2): 435–439.

Bernanke, B., and M. Gertler. 1987. Banking in general equilibrium. In *New approaches to monetary economics*, ed. W. Barnett and K. Singleton. Cambridge: Cambridge University Press.

———. 1989. Agency cost, net worth and business fluctuations. *American Economic Review* 79 (1): 14–31.

———. 1990. Financial fragility and economic performance. *Quarterly Journal of Economics* 105 (1): 87–114.

———. 1995. Inside the black box: The credit channel of monetary policy transmission. *Journal of Economic Perspectives* 9 (4): 27–48.

Bernanke, B., and H. James. 1991. The gold standard, deflation, and financial crisis in the Great Depression: An international comparison. In *Financial markets and financial crises*, ed. R. Hubbard. Chicago: University of Chicago Press.

Bertrand, M., A. Schoar, and D. Thesmar. 2007. Banking deregulation and industry structure: Evidence from the French banking reforms of 1985. *Journal of Finance* 62 (2): 597–628.

Bolton, P., and X. Freixas. 2000. Equity, bonds, and bank debt: Capital structure and financial market equilibrium under asymmetric information. *Journal of Political Economy* 108 (2): 324–351.

Boot, W. A., and A. V. Thakor. 1997. Financial system architecture. *Review of Financial Studies* 10 (3): 693–733.

Boyd, J., and E. Prescott. 1986. Financial intermediary coalitions. *Journal of Economic Theory* 38 (2): 211–232.

Cetorelli, N., and M. Gambera. 2001. Banking market structure, financial dependence and growth: International evidence from industry data. *Journal of Finance* 56: 617–648.

Ciccone, A., and E. Papaioannou. 2006. Financial development and efficient capital reallocation. European Central Bank and University of Pompeu Fabra. Mimeograph.

Cook, T., and T. Rowe. 1986. *Instruments of the money market*. 6th ed. Richmond, Va.: Federal Reserve Bank of Richmond.

Da Rin, M., and T. Hellmann. 2002. Banks as catalysts for industrialization. *Journal of Financial Intermediation* 11 (4): 366–397.

Demirgüç-Kunt, A., and V. Maksimovic. 2002. Funding growth in bank-based and market-based financial systems: Evidence from firm-level data. *Journal of Financial Economics* 65 (3): 337–363.

Diamond, D. W. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51 (3): 393–414.

Diamond, D. W., and P. H. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91 (3): 401–419.

Farmer, R. 1984. New theory of aggregate supply. *American Economic Review* 74 (5): 920–930.

Fisher, I. 1933. The debt deflation theory of great depressions. *Econometrica* 1 (4): 337–357.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton, N.J.: Princeton University Press.

Fuerst, T. 1992. Liquidity, loanable funds, and real activity. *Journal of Monetary Economics* 29 (1): 3–24.

Gerschenkron, A. 1962. *Economic backwardness in historical perspective: A book of essays*. Cambridge, Mass.: Harvard University Press.

———. 1992. Economic backwardness in historical perspective. In *The sociology of economic life*, ed. M. Granovetter and R. Swedberg. Boulder, Colo.: Westview Press.

Gertler, M. 1988. Financial structure and aggregate economic activity: An overview. *Journal of Money, Credit and Banking* 20 (3): 559–588.

Gertler, M., and S. Gilchrist. 1992. Monetary policy, business cycles and the behavior of small manufacturing firms. Economic Research Report 92-108, New York University, New York.

Goldsmith, R. 1969. *Financial structure and development.* New Haven: Yale University Press.

Gorton, G., and G. Pennacchi. 1990. Financial innovation and the provision of liquidity services. Wharton School, University of Pennsylvania, Philadelphia. Mimeograph.

Greenwood, J., and B. Jovanovic. 1990. Financial development, growth and the distribution of income. *Journal of Political Economy* 98 (5): 1076–1107.

Guiso, L., P. Sapienza, and L. Zingales. 2004. Does local financial development matter? *Quarterly Journal of Economics* 119 (3): 929–969.

Gurley, J., and E. Shaw. 1955. Financial aspects of economic development. *American Economic Review* 65: 515–538.

———. 1960. *Money in the theory of finance.* Washington: Brookings Institution.

Hart, O., and J. Moore. 1994. A theory of debt based on the inalienability of human capital. *Quarterly Journal of Economics* 109 (4): 841–879.

Holmström, B., and J. Tirole. 1997. Financial intermediation, loanable funds, and the real sector. *Quarterly Journal of Economics* 112 (3): 663–691.

Hoshi, T., A. Kashyap, and D. Scharfstein. 1991. Corporate structure, liquidity, and investment: Evidence from Japanese industrial groups. *Quarterly Journal of Economics* 106 (1): 33–60.

James, C. 1987. Some evidence on the uniqueness of bank loans. *Journal of Financial Economics* 19 (2): 217–235.

Jayaratne, J., and P. E. Strahan. 1996. The finance-growth nexus: Evidence from bank branching deregulation. *Quarterly Journal of Economics* 111 (3): 639–670.

Kashyap, A., and J. Stein. 1993. Monetary policy and bank lending. NBER working paper 4317.

Kashyap, A., and J. Stein. 2000. What do a million observations on banks say about the transmission of monetary policy? *American Economic Review* 90 (3): 407–428.

Kashyap, A., J. C. Stein, and D. W. Wilcox. 1993. Monetary policy and credit conditions: Evidence from the composition of external finance. *American Economic Review* 83 (1): 79–98.

King, R. G., and R. Levine. 1993. Financial intermediation and economic development. In *Capital markets and financial intermediation*, ed. C. Mayer and X. Vives. Cambridge: Cambridge University Press.

King, S. 1986. Monetary transmission: Through bank loans or bank liabilities? *Journal of Money, Credit and Banking* 18 (3): 290–303.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny. 2000. Investor protection and corporate governance. *Journal of Financial Economics* 58 (1–2): 3–27.

Levine, R. 2002. Bank-based or market-based financial systems: Which is better? *Journal of Financial Intermediation* 11 (4): 398–428.

———. 2005. Finance and growth: Theory and evidence. In *Handbook of Economic Growth*, ed. P. Aghion and S. Durlauf, vol. 1, ch. 12. Amsterdam: Elsevier.

Mihov, I. 2001. Monetary policy implementation and transmission in the European Monetary Union. *Economic Policy* 16 (33): 369–406.

Mishkin, F. S. 1978. The household balance sheet and the Great Depression. *Journal of Economic History* 38 (4): 918–937.

Modigliani, F., and M. Miller. 1958. The cost of capital, corporation finance and the theory of investment. *American Economic Review* 68 (3): 261–297.

Papaioannou, E. 2006. An overview and appraisal of the finance and growth literature. European Central Bank. Mimeograph.

Rajan, R. G. 1992. Insiders and outsiders: The choice between informed and arm's length debt. *Journal of Finance* 47 (4): 1367–1400.

Rajan, R. G., and L. Zingales. 1998. Financial dependence and growth. *American Economic Review* 88: 559–586.

Ramey, V. 1992. How important is the credit channel of monetary transmission? University of California, San Diego. Mimeograph.

Romer, C. D., and D. H. Romer. 1990. New evidence on the monetary transmission mechanism. In *Brookings papers on economic activity*, vol. 1, 149–198.

Roosa, R. V. 1951. Interest rates and the central bank. In *Money, trade and economic growth: Essays in honor of John H. Williams.* New York: Macmillan.

Scheinkman, J., and L. Weiss. 1986. Borrowing constraints and aggregate economic activity. *Econometrica* 54 (1): 23–45.

Schnadt, N. 1993. The domestic money markets of the UK, France, Germany and the US. Financial Markets Group, London School of Economics.

Schreft, S. L. 1990. Credit controls: 1980. Federal Reserve Bank of Richmond. *Economic Review* 76 (6): 25–55.

Schumpeter, J. 1934. *The theory of economic development*. Cambridge, Mass.: Harvard University Press. First published in German, 1912. Republished New York: Oxford University Press, 1961.

Sharpe, S. 1990. Asymmetric information, bank lending and implicit contracts: A stylized model of customer relationships. *Journal of Finance* 45 (4): 1069–1087.

Slovin, M., M. Sushka, and J. Polonchek. 1993. The value of bank durability: Borrowers as bank stakeholders. *Journal of Finance* 48 (1): 247–266.

Stein, J. C. 1998. An adverse-selection model of bank asset and liability management with implications for the transmission of monetary policy. *Rand Journal of Economics* 29 (3): 466–486.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71 (3): 393–410.

Sussman, O., and J. Zeira. 1995. Banking and development. Discussion paper 1127. Centre for Economic Policy Research, London School of Economics.

Tadesse, S. 2002. Financial architecture and economic performance: International evidence. *Journal of Financial Intermediation* 11 (4): 429–454.

Van den Heuvel, S. 2002. The bank capital channel of monetary policy. Wharton School, University of Pennsylvania, Philadelphia. Mimeograph.

Williamson, S. D. 1987. Financial intermediation, business failures, and real business cycles. *Journal of Political Economy* 95 (6): 1196–1216.

# 7 Individual Bank Runs and Systemic Risk

Bank panics were a recurrent phenomenon in the United States until 1934. According to Kemmerer (1910), the country experienced 21 bank panics between 1890 and 1908. Similarly, Friedman and Schwartz (1963) enumerate five bank panics between 1929 and 1933,[1] the most severe period in the financial history of the United States. Miron (1986) extensively documents this phenomenon, recalling its seasonal pattern prior to the founding of the Federal Reserve System (the Fed). Moreover, whereas the average yearly growth rate of real gross national product (GNP) was 3.75 during this period, Miron finds that if the years in which a bank panic occurred (or following a bank panic) are taken out of the sample,[2] the average growth rate becomes 6.82 percent. Similar phenomena affected England before the establishment of a Central Bank, as well as other European countries (Bordo 1990; Eichengreen and Portes 1987).

More recently, many countries have experienced banking crises, in part initiated by the general movement toward financial deregulation. For example, banking crises in East Asia, Mexico, and the Scandinavian countries started when the savings and loan crisis began to ebb away in the United States.

Therefore, it seems that without regulation, bank runs and bank panics are inherent to the nature of banking, and more specifically to the fractional reserve system. Indeed, bank deposit contracts usually allow depositors to withdraw the nominal value of their deposits on demand. As soon as a fraction of these deposits is used for financing illiquid and risky loans or investments, there is a possibility of a liquidity crisis. This chapter examines whether such deposit contracts are efficient and whether the fractional reserve system is justified despite the possibility of bank runs.

Most theoretical models have addressed this question in an aggregate framework, representing the whole banking industry by a unique entity. However, it is important to distinguish between *bank runs*, which affect an individual bank, and *bank panics*, which concern the whole banking industry,[3] the payment system, and the interbank market.

The conventional explanation for a bank run is that when depositors learn bad news about their bank, they fear bankruptcy and respond by withdrawing their own deposits. This bad news can be about the value of the bank's assets (fundamental bank run) or about large withdrawals (speculative bank run). Withdrawals in excess of the current expected demand for liquidity generate a negative externality for the bank experiencing the liquidity shortage because they imply an increase in the bank's probability of failure. But they can also generate an externality for the whole banking system if the agents view the failure as a symptom of difficulties occurring throughout the industry.

In such a case, a bank run may develop into a bank panic. Bagehot (1873) was one of the first to analyze how the Central Bank could prevent such contagion by playing the part of a lender of last resort (LLR). Section 7.7 is devoted to this question. The other sections are organized as follows. Section 7.1 recalls the model of liquidity insurance presented in section 2.2, and section 7.2 introduces a fractional reserve banking system and studies its stability. Sections 7.3 and 7.4 discuss bank runs. Section 7.5 is dedicated to interbank markets, and section 7.6 examines systemic risk and contagion.

## 7.1 Banking Deposits and Liquidity Insurance

This section recalls the simple model of liquidity insurance (Bryant 1980; 1981; Diamond and Dybvig 1983) introduced in chapter 2. It then discusses different institutional arrangements that can provide this liquidity insurance to individual economic agents.

### 7.1.1 A Model of Liquidity Insurance

Consider a one-good, three-dates economy in which a continuum of agents, each endowed with one unit of good at date $t = 0$, want to consume at dates $t = 1$ and $t = 2$. These agents are ex ante identical, but they are subject to independently identically distributed (i.i.d.) liquidity shocks in the following sense: with some probability $\pi_i$ ($i = 1, 2$, with $\pi_1 + \pi_2 = 1$), they need to consume at date $t = i$. The utility of agents of type $i = 1$ (impatient consumers) is $u(C_1)$, and that of agents of type $i = 2$ (patient consumers) is $u(C_2)$. Ex ante all agents have the same utility:[4]

$$U = \pi_1 u(C_1) + \pi_2 u(C_2). \tag{7.1}$$

Assume that $u$ is increasing and concave.

There is a storage technology that allows transfer of the good without cost from one date to the next. More important, there is also a long-term illiquid technology (with constant returns to scale): one unit invested at $t = 0$ gives a return $R > 1$ at

$t = 2$. The term *illiquid* reflects the fact that investments in this long-term technology give a low return $\ell \leq 1$ if they are liquidated prematurely at $t = 1$. This section determines the characteristics of the optimal (symmetric) allocation. It begins by studying two benchmarks: the autarkic situation and the allocation obtained when a financial market is opened.

### 7.1.2 Autarky

Autarky corresponds to the absence of trade between agents. Each of them independently chooses at $t = 0$ the level $I$ of his investment in the long-term technology and stores the rest $(1 - I)$. In the case of a liquidity shock at date $t = 1$, the investment is liquidated, yielding a consumption level

$$C_1 = \ell I + 1 - I. \tag{7.2}$$

If consumption occurs at date $t = 2$, the consumption level obtained is

$$C_2 = RI + 1 - I. \tag{7.3}$$

At date $t = 0$, consumers choose $I$ so as to maximize $U$ under constraints (7.2) and (7.3). Notice that since $\ell < 1 < R$, then $C_1 \leq 1$ and $C_2 \leq R$, with at least one strict inequality. This comes from the fact that the investment decision is always ex post inefficient with a positive probability: if $i = 1$, the efficient decision is $I = 0$, whereas it is $I = 1$ if $i = 2$. This inefficiency can be mitigated by opening a financial market.

### 7.1.3 The Allocation Obtained When a Financial Market Is Opened

Suppose that a bond market is opened at $t = 1$, whereby $p$ units of good at $t = 1$ are exchanged against the promise to receive one unit of good at $t = 2$. The consumption levels obtained by each consumer at dates 1 and 2 become

$$C_1 = pRI + 1 - I \tag{7.4}$$

and

$$C_2 = RI + \frac{1 - I}{p}. \tag{7.5}$$

In the first case, the impatient agent has sold $RI$ bonds (instead of liquidating his long-term investment), whereas in the second case the patient agent has bought $(1 - I)/p$ bonds at $t = 1$ (instead of storing the good for another period). Notice that $C_1 = pC_2$ and that the utility of the agent is increasing in $I$ if $pR > 1$, and decreasing if $pR < 1$. Since agents choose at date $t = 0$ the amount $I$ they invest in the long-run technology, an interior maximum exists only when $pR = 1$. Therefore, the only (interior) equilibrium price of bonds is $p = 1/R$, and the allocation obtained is $C_1 = 1, C_2 = R$, which Pareto-dominates the autarkic allocation. This is because the existence of a financial market ensures that the investment decisions are efficient. However, this market allocation is not Pareto-optimal in general, because liquidity risk is not properly allocated.

### 7.1.4 The Optimal (Symmetric) Allocation

Agents being ex ante identical, it is legitimate to focus on the (unique) symmetric optimal allocation obtained by

$$\mathscr{P}_1 \begin{cases} \max\limits_{C_1, C_2, I} \; U = \pi_1 u(C_1) + \pi_2 u(C_2) \\ \text{under the constraints} \\ \pi_1 C_1 = 1 - I, & (7.6) \\ \pi_2 C_2 = RI. & (7.7) \end{cases}$$

Replacing $C_1$ and $C_2$ by their values given by (7.6) and (7.7), $U$ becomes a function of the single variable $I$:

$$U(I) = \pi_1 u\left(\frac{1-I}{\pi_1}\right) + \pi_2 u\left(\frac{RI}{\pi_2}\right).$$

The solution $(C_1^*, C_2^*, I^*)$ of $\mathscr{P}_1$ is thus determined by the constraints (7.6) and (7.7) and the first-order condition:

$$-u'(C_1^*) + Ru'(C_2^*) = 0. \tag{7.8}$$

In general, the market allocation $(C_1 = 1, C_2 = R, I = \pi_2)$ does not satisfy (7.8) except in the peculiar case in which $u'(1) = Ru'(R)$. An interesting situation arises when $u'(1) > Ru'(R)$.[5] In this case, impatient consumers get more in the optimal allocation than in the market equilibrium ($C_1^* > 1$): they need to be insured against a liquidity shock at $t = 1$. The next section shows how a fractional reserve banking system can provide this liquidity insurance.

### 7.1.5 A Fractional Reserve Banking System

The optimal allocation characterized in the previous section can be implemented by a fractional reserve banking system in which banks collect the endowments of consumers (deposits) and invest a fraction of them in long-term investments while offering depositors the possibility of withdrawal on demand. A deposit contract $(C_1, C_2)$ specifies the amounts $C_1$ and $C_2$ that can be withdrawn, respectively, at dates $t = 1, 2$ for a unit deposit at $t = 0$. Competition between banks leads them to offer the optimal feasible deposit contract $(C_1^*, C_2^*)$ characterized earlier. A crucial question is whether this fractional reserve system is stable, that is, whether the banks will be able to fulfill their contractual obligations. This depends very much on the behavior of patient consumers, which in turn depends on their anticipations about the safety of their bank.

Consider the case of a patient consumer who anticipates that the bank will be able to fulfill its obligations. The consumer has the choice between withdrawing $C_2^*$ at date $t = 2$ or withdrawing $C_1^*$ at date $t = 1$ and storing it until $t = 2$. Since $R > 1$ and $u'$ decreases, equation (7.8) shows that $C_2^* > C_1^*$.

This means that if the patient consumer trusts her bank, she will always prefer to withdraw at $t = 2$. By the law of large numbers, the proportion of withdrawals at $t = 1$ will be exactly $\pi_1$. This determines the amount $\pi_1 C_1^*$ of liquid reserves that the bank has to make in order to avoid premature liquidation. With these reserves, the bank will be solvent with probability 1, and the consumers' expectations will be fulfilled. Thus there is an equilibrium of the banking sector that implements the optimal allocation. However, another equilibrium also exists, which leads to an inefficient allocation.

Suppose the patient consumer anticipates that all other patient consumers want to withdraw at $t = 1$. The bank will be forced to liquidate its long-term investments, yielding a total value of assets $\pi_1 C_1^* + (1 - \pi_1 C_1^*)\ell$, which is less than 1 and thus less than the total value of its liabilities ($C_1^*$). In the absence of other institutional arrangements, the bank will fail and nothing will be left at $t = 2$. Anticipating this, the optimal strategy for a patient consumer is to withdraw at $t = 1$. Thus the initial expectations of the consumer are self-fulfilling. In other words, there is a second Nash equilibrium[6] of the withdrawal game in which all consumers withdraw at $t = 1$ and the bank is liquidated: this is what is called an inefficient bank run.

**Result 7.1** In a fractional reserve banking system in which investment returns are high enough ($R > 1$), two possible situations may arise at equilibrium:

- An efficient allocation, when patient depositors trust the bank and withdraw only at $t = 2$
- An inefficient bank run, when all depositors withdraw at $t = 1$

Change in expectations was supposed to be the main channel of contagion in the nineteenth century and the beginning of the twentieth century. This channel seems also important today in countries that are vulnerable to currency crises. This type of crisis closely resembles bank runs because they occur as self-fulfilling prophecies. The bankruptcy of a first bank makes depositors update their beliefs concerning the other banks' solvency. Under the updated beliefs, depositors may prefer to run the second bank, which goes bankrupt. It does not matter whether the beliefs concern the second bank's solvency (its return on assets $R$, in the notation we have been using) or the proportion of agents that will run the bank. In both cases, patient depositors are perfectly rational in running the bank. As a consequence, contagion through change in expectations is easily modeled using the Bryant-Diamond-Dybvig framework and does not require extended development here.

## 7.2 The Stability of the Fractional Reserve System and Alternative Institutional Arrangements

### 7.2.1 The Causes of Instability

As was just shown, the fractional reserve banking system leads to an optimal allocation only if patient consumers do not withdraw early. There are two reasons they might want to withdraw.

If the relative return of date 2 deposits with respect to date 1 deposits $(C_2^*/C_1^* - 1)$ is less than what patient consumers can obtain elsewhere, either by storage (as in the present model) or more generally by reinvesting in financial markets, as in Von Thadden (1996), they will prefer to withdraw early. If consumers' types were observable, this could be avoided by forbidding patient consumers to withdraw early. In practice, however, liquidity needs are not publicly observable, and incentive compatibility constraints must be introduced. In a continuous time extension of the Bryant-Diamond-Dybvig model, Von Thadden (1996) shows that these incentive compatibility constraints are *always* binding somewhere, which severely limits the provision of liquidity insurance that can be obtained through a fractional reserve banking system.

The literature has paid more attention to a second cause of instability, arising from the fact that the game between depositors has two equilibria, one efficient and one inefficient. The inefficient equilibrium arises only when there is a coordination failure among depositors, coming from a lack of confidence in their bank. In any case, theoreticians dislike multiple equilibria, and they have tried to offer selection devices. For instance, Anderlini (1989) suggests a recourse to exogenous uncertainty ("sunspots") to determine which equilibrium will prevail. This might explain sudden confidence crises in real-world banking systems. On the other hand, Postlewaite and Vives (1987) suggest that some agents may observe signals that give them some information about the likelihood of a bank run (see problem 7.8.2); these are "information-based" bank runs. The following sections discuss several institutional arrangements that have been proposed to solve the instability problem of the fractional reserve system.

### 7.2.2 A First Remedy for Instability: Narrow Banking

A natural way to prevent the instability of the banking system is to require that under any possible circumstance all banks can fulfill their contractual obligations.

This idea given rise to the term *narrow banking*, which refers to a set of regulatory constraints on banks' investment opportunities that would make them safe in any possible event. But this is open to different interpretations, leading to three alternative views of narrow banking: (1) a bank with enough liquidity to guarantee repayment to all depositors even in case of bank run, (2) a bank that obtains enough liquidity after liquidation of its long-run technology to face a bank run, and (3) a bank that obtains enough liquidity after the securitization of its long-run technology to cope with a bank run. The discussion shows why, in terms of risk and resource allocation, narrow banking leads to an inefficient allocation for each of the three interpretations.

In the first version of narrow banking, the bank is required to have a reserve ratio of 100 percent: liquid reserves $(1 - I)$ at least equal to $C_1$, the maximum possible amount of withdrawals at date 1.[7] In practice, the maturity structure of banks' assets should be perfectly matched with that of their liabilities. In the present context, this means that the deposit contract $(C_1, C_2)$ offered by the bank must satisfy $C_1 \leq 1 - I$, and $C_2 \leq RI$.

The best deposit contract $(C_1, C_2)$ that can offer such a narrow bank is defined by

$$\mathscr{P}_2 \begin{cases} \max\limits_{I, C_1, C_2} \; U = \pi_1 u(C_1) + \pi_2 u(C_2) \\ \text{under the constraints} \\ C_1 \leq 1 - I, \qquad C_2 \leq RI. \end{cases}$$

As Wallace (1988; 1996) points out, the solution of $\mathscr{P}_2$ is dominated by that of $\mathscr{P}_1$. In fact, it is even dominated by the autarkic situation, which will be obtained if the second, milder version of the narrow banking proposal is adopted, in which the banks are allowed to liquidate some of their assets in order to satisfy unexpected withdrawals. If the bank has offered the deposit contract $(C_1, C_2)$, the amount $I$ invested in the long-term technology must now be such that

$$C_1 \leq \ell I + (1 - I)$$

(the liquidation value of the bank's assets at $t = 1$ covers the maximum possible amount of withdrawals). Similarly, at $t = 2$,

$$C_2 \leq RI + 1 - I,$$

which means a return to the autarkic situation.

Finally, the more modern, weaker version of narrow banking suggests replacing banks by money market funds that use the deposits they collect to buy (riskless) financial securities (Gorton and Pennacchi 1993). Alternatively, banks would be allowed to securitize their long-term assets in order to satisfy the withdrawals of depositors. It is easy to see that the best deposit contract that can be offered by such a money market fund is the market equilibrium characterized in section 7.1.3. Therefore, even with this last version involving money market funds (or monetary service companies), the narrow banking proposal is antagonistic to the efficient provision of liquidity insurance. It remains to be seen whether this efficient provision can be obtained under other institutional arrangements that would guarantee the stability of the banking system.

### 7.2.3 Regulatory Responses: Suspension of Convertibility or Deposit Insurance

If liquidity shocks are perfectly diversifiable, and if the proportion[8] $\pi_1$ of impatient consumers is known, one can get rid of the coordination problem that gives rise to inefficient bank runs. For instance, the bank could announce that it will not serve more than $\pi_1 C_1^*$ withdrawals at date $t = 1$. After this threshold, convertibility is suspended. Patient consumers therefore know that the bank will be able to satisfy its engagements at date 2, and thus they have no interest in withdrawing at date 1. The threat of a bank run disappears.[9]

An equivalent way to get rid of inefficient bank runs is to insure depositors. In this case, even if the bank is not able to fulfill its obligations, depositors receive the full value of their deposits. The difference is paid by a new institution, the deposit insurance system, financed by insurance premiums paid ex ante by the bank (or by taxes, if the system is publicly run). In the present simple framework, the existence of deposit insurance is enough to get rid of bank failures.[10]

The equivalence between these two systems breaks down as soon as one allows for a variability of the proportion $\pi_i$. In that case, the equilibrium without bank runs is characterized by a random amount of date 1 withdrawals. But since the level of investment has been already chosen, two situations may arise. First, if the realized value of $\pi_1$ is too high, the investment in the long-run technology will have to be liquidated at a loss, and the bank will not be able to meet its obligations at date 2. If $\pi_1$ is too low, the level of investment is also too low, and time 2 depositors will again not obtain the promised return. More generally, any type of regulation that is intended to cope with random withdrawals must take into account the fact that time 2 returns are contingent on time 1 withdrawals. In such a case, it is inefficient to set a critical level of liquidity demand (type 1 deposits) that triggers the suspension of convertibility. If this level is $\hat{f}$, a realization of $\pi_1$ with $\pi_1 > \hat{f}$ implies that type 1 agents will be rationed. Conversely, a realization with $\pi_1 < \hat{f}$ implies that a bank run may still develop because type 2 agents are actually too numerous in relation to the promised return, given the amount the bank has invested. Thus, even if it is true that the suspension of convertibility will eliminate bank runs, it will do so at a cost because the deposit contracts will then be less efficient in terms of risk sharing.

Deposit insurance, however, will allow for a contingent allocation. For instance, if the deposit insurance system is publicly run and financed by taxes, the government can levy a tax based on the realization of $\pi_1$. If the tax rate is the same across agents, it can be interpreted as resulting from the adjustment of the time 1 price of the good (inflationary tax). Of course, as pointed out by Wallace (1988) this is only possible if the potential taxpayers have not already consumed the good.

### 7.2.4 Jacklin's Proposal: Equity versus Deposits

Since a demand deposit economy achieves better risk sharing than a market economy but is vulnerable to bank runs, it is interesting to investigate whether other contractual arrangements achieve the same allocations without being prone to bank runs. Jacklin (1987) has shown that sometimes equity can do as well as deposit contracts.

Instead of the mutual bank of the previous section, consider a bank entirely financed by equity, which announces it will distribute a dividend $d$ at date 1. Accordingly, it keeps an amount of reserves equal to $d$ and invests $(1-d)$ in long-run technology. The shares of the bank are traded during period 1, once agents know their types and after the dividend is paid. Each share gives a right to $R(1-d)$ consumption units at time 2. The equilibrium price $p$ of the ex-dividend share in terms of time 1 consumption good depends on $d$. Therefore, changing the value of $d$ affects the period 1 utilities of the agents, so this mechanism will also allow for some improvement of the ex ante expected utility with respect to the market economy.

For a dividend $d$ and a price $p$ for the ex-dividend share, the behavior of each type of shareholder of the bank can be determined. Type 1 agents (impatient consumers) receive their dividends and sell their shares,

$$C_1 = d + p, \tag{7.9}$$

whereas type 2 agents (patient consumers) use their dividends $d$ to buy $d/p$ new shares, which gives them at date $t = 2$,

$$C_2 = \left(1 + \frac{d}{p}\right) R(1-d). \tag{7.10}$$

The price $p$ is determined by equality of supply and demand of shares:

$$\pi_1 = \pi_2 \frac{d}{p}, \tag{7.11}$$

which gives

$$p = \frac{\pi_2 d}{\pi_1}$$

and

$$C_1 = \frac{d}{\pi_1}, \qquad C_2 = \frac{R(1-d)}{\pi_2}. \tag{7.12}$$

Finally, the level of $d$ is determined ex ante (at $t = 0$) by stockholders,[11] who unanimously choose it to maximize $U$ under the constraints (7.12). Eliminating $d$ between these constraints yields the same budget constraint as for deposit contracts:

$$\pi_1 C_1 + \pi_2 \frac{C_2}{R} = 1.$$

Therefore, in the simple specification adopted here, the efficient allocation $(C_1^*, C_2^*)$ can also be obtained by participation contracts in which consumers are shareholders of the bank instead of depositors. The advantage of these participation contracts is that they are immune to bank runs. However, for more general specifications of agent utilities, Jacklin shows that equity contracts can be dominated by efficient deposit contracts, yielding a trade-off between stability and efficiency because deposit contracts can be destabilized by bank runs. Figure 7.1 illustrates this point. If the efficient allocation happens to be implementable in an equity economy, the three contractual frameworks are equivalent. This is the Diamond-Dybvig case. Still, if the efficient allocation is outside the banking economy area, the banking economy will typically perform better.

The reason for this domination is that equity contracts are necessarily *coalitionally incentive-compatible* in the sense that they are immune to early withdrawals (deviations) of coalitions of patient consumers, whereas deposit contracts are only *individually incentive-compatible*. If agents are allowed to trade their deposit contracts, these contracts become equivalent to equity contracts. This is the case because a time $t = 2$

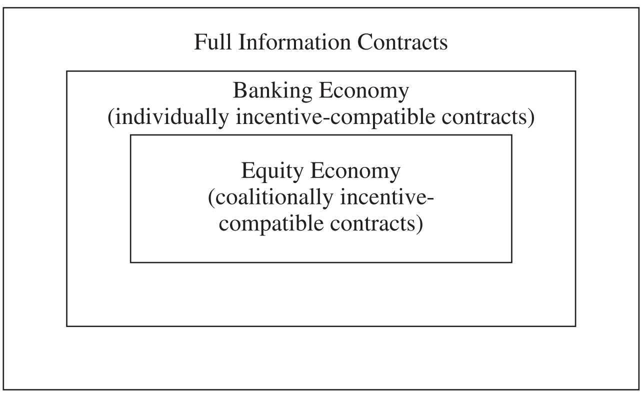



**Figure 7.1**
Different sets of contracts.

deposit has then a market price; in equilibrium, agents should be indifferent between selling their time $t = 2$ deposits in the market or cashing them at the bank.

## 7.3 Bank Runs and Renegotiation

Diamond and Rajan (2001) provide an alternative justification for banking based on the disciplinary role of bank runs. Their argument is based on the fact that contracting parties have the power to renegotiate the original agreed rules. The same asset may therefore be worth more to an entrepreneur developing a project than to its main financier, and the value of the asset to its main financier with experience in dealing with this kind of assets is higher than to other financiers. In this context, the mechanism of bank runs may limit the renegotiation power of a financier, since the threat to diminish the amount of repayment to depositors could trigger a bank run, whereas if the financier is dealing with a unique depositor the withdrawal of deposits is not a credible threat, and the financier can easily renegotiate a lower payment. In other words, the possibility of a bank run, which is usually characteristic of banks' fragility, here gives more bargaining power to depositors, and therefore may lead to a higher level of financing.

### 7.3.1 A Simple Model

Consider an economy with an excess of savings, so that the opportunity cost of funds is equal to 1. Entrepreneurs have projects but no cash, so they will borrow from agents that we call financiers. Still, as in Hart and Moore (1994), entrepreneurs cannot commit to the project in the future because they cannot alienate their human capital. As a consequence, the entrepreneurs may threaten to quit and bargain a better deal on their loan contract.

Formally, assume an entrepreneur invests $Iy (I < 1)$ in the project. This allows it to obtain a riskless cash flow $y$ at date $t = 1$. A financier is willing to provide funds in exchange for a repayment $R$. If the financier liquidates the project before time $t = 1$, it gets $V_1$.

A financier will be a specialized financial intermediary in the sense that it has experience in liquidating the type of firm it is lending to. If another external financier should replace the first one in the contract, it would only obtain a liquidation value for the assets equal to $\alpha V_1$, where $0 < \alpha < 1$. This means that the market is not ex post competitive, either because of a lack of specialized financiers, an interpretation that would make financiers akin to providers of venture capital, or because of the fact that there is a relationship with the borrower that is valuable to both parties (see section 3.6). Both interpretations, which are not incompatible, provide a rationale for financial intermediaries.

### 7.3.2 Pledgeable and Nonpledgeable Cash Flows

Assume that the borrower has all the bargaining power. As in Hart and Moore (1994), any contractual repayment higher than $V_1$ will be reduced to that level, so renegotiation-proof contracts will be those satisfying $R \leq V_1$.

The implication is that even if the entrepreneur has a high future expected cash flow $y$, it will not be able to borrow against this cash flow but only against the liquidation value of its assets, and the maximum amount of funds the entrepreneur is able to borrow is $V_1$.

Assume only the uninformed lender has funds. Two possible financial arrangements are then possible: (1) The entrepreneur borrows directly from the uninformed lender, and (2) the entrepreneur borrows from the financial intermediary, which in turn borrows from the uninformed lender. (This can be interpreted as a loan sale.)

Assume the entrepreneur has all the bargaining power when negotiating with another party and that the financial intermediary has all the bargaining power when negotiating with the uninformed lender. Then the entrepreneur can only fund its project up to the level of $\alpha V_1$. A larger amount would not be renegotiation-proof.[12]

In case 1, where the entrepreneur borrows directly from the uninformed lender, the Hart and Moore argument applies, and the maximum value of the repayment is $\alpha V_1$.

In case 2 the entrepreneur could, in principle, borrow $V_1$ because this is the amount it can extract from the entrepreneur. For this to be possible, the financial intermediary has to be able to raise $V_1$ from the uninformed lender. Still, the same renegotiation problem that limits the funds available to the entrepreneur will now limit the funds to the financial intermediary. Indeed, any repayment to the uninformed lender larger than $\alpha V_1$ will be renegotiated, and the uninformed lender's best option will be to accept the offer. In anticipation of this, the uninformed lender will only lend $\alpha V_1$, and therefore both cases lead to financing the project only to the level of $\alpha V_1$.

The existence of the financial intermediary, which could theoretically provide funds up to the amount $V_1$, is of no help because it is itself unable to raise (credibly) the required funds.

### 7.3.3 Bank Runs as a Discipline Device

Consider the case where the financial institution chooses to be funded not by a unique uninformed lender but by a demand deposit structure, thus committing to serve all deposits on a first-come, first-served basis. It is then possible for the bank to commit to a total repayment $V_1$.

Any attempt by the bank to threaten to withdraw its specific collection skills and to renegotiate down the total repayment, say, to $\alpha V_1$, will then trigger a bank run.

To see this, assume the financial intermediary raises an amount equal to $V_1$ by offering a deposit contract to two external financiers. This contract gives depositors the right to withdraw at any time an amount

$$\frac{d}{2} = \frac{V_1}{2},$$

and has the property of sequential service, first-come, first-served.

In this case, any attempt to renegotiate the payment on deposit by decreasing it by $\varepsilon$ will leave the two depositors facing the following game:

| | Withdraw | Wait |
|---|---|---|
| Withdraw | $\left(\frac{\alpha V_1}{2}, \frac{\alpha V_1}{2}\right)$ | $\left(\frac{d}{2}, \alpha V_1 - \frac{d}{2} - \varepsilon\right)$ |
| Wait | $\left(\alpha V_1 - \frac{d}{2} - \varepsilon, \frac{d}{2}\right)$ | $\left(\frac{d}{2} - \varepsilon, \frac{d}{2} - \varepsilon\right)$ |

The equilibrium outcome of the game is (withdraw, withdraw), that is, a bank run. The depositors will not gain anything by running the bank. Indeed, once they are in possession of the loan, the entrepreneur will make them a take-it-or-leave-it offer, and they will obtain $V_1/2$. The outcome is then the same as for the case of a loan sale: the entrepreneurs' rent increases. The loser will be the bank, which will go bankrupt.

Thus, a demand deposit structure allows the bank to commit to a larger nonrenegotiable repayment. This implies that the bank is able to act as a delegated monitor, committing in a credible way not to renegotiate the promised repayment to depositors. Thus the demand deposit structure of contracts gives the financial intermediary the ability to borrow additional funds, which in turn can be passed down to the entrepreneur. In other words, it is only because of its deposit structure that the services of the relationship lender can be used by the entrepreneur.

To summarize, in this model, the role of the bank is to "tie human capital to assets" (Diamond and Rajan 2000).

### 7.3.4 The Role of Capital

Under the previous assumptions, it is always optimal for a bank to hold zero capital. Still, when the liquidation value becomes random, the bank's structure of capital becomes a relevant issue. The randomness of $V_1$ could lead to runs even in the absence of opportunistic behavior by the banker. This justifies a role for bank capital, characterized by Diamond and Rajan (2000) as a softer claim that can be renegotiated in bad times.

In order to summarize the main result of Diamond and Rajan (2000), we assume that when the bank managers renegotiate with the equity holders, they have all the bargaining power. Also, we assume that in the event of a bank run, depositors cannot employ the services of the bank managers for a fee but simply get the value of the liquidated asset, $\alpha V$.

Assume $V$ can take two values $V_H$ and $V_L$ $(V_H > V_L)$ with probabilities $p$ and $1 - p$, respectively. Assume also $V_L < \alpha V_H$.[13]

In the case $d = V_L$, if $V_L$ occurs, depositors are paid, and neither the bank managers nor the equity holders obtain any payment. If $V_H$ occurs, then the bank managers threaten not to extract from the entrepreneurs all the value $V_H$ but only $\alpha V_H$. Since the bank managers have all the bargaining power, this leaves the equity holders with $\alpha V_H - V_L$ and the bank managers with $V_H(1 - \alpha)$. The maximum amount of funding the project will obtain is therefore $p(\alpha V_H - V_L) + V_L$, which is strictly lower than $E(V)$. The difference between the two is due to the cost of renegotiation when $H$ occurs. Equity holders will enter the project, but renegotiation implies that there is ex ante a cost of equity finance.

In the case $d = V_H$, with a probability $1 - p$ that the bank is bankrupt, there is a bank run and there is no role for equity finance because equity holders always receive a zero return. Nevertheless, the choice of $d = V_L$ or $d = V_H$ is endogenous, so a high level of deposits $d = V_H$ might be chosen whenever this allows raising a higher amount of funds. For $d = V_H$, the maximum amount raised is $pV_H + (1 - p)\alpha V_L$, and this could be larger than the amount obtained for $d = V_L$.

To summarize, as shown by comparing the cases for $d = V_L$ and $d = V_H$, capital allows avoiding bankruptcy, that is, having to pay $V_H$ when the value of assets are $V_L$. But this comes at a cost—the cost of bank managers extracting rents in case of success—a cost that investors will take into account ex ante.

## 7.4 Efficient Bank Runs

In the models examined up to now, the investment returns are certain. This implies that bank runs have a purely speculative origin. Yet it is reasonable to think that leakage of bad performance of bank loan portfolios should also trigger bank runs. Empirical evidence on bank runs seems to point in that direction. Thus bank runs can also have a fundamental origin, motivated by the expectation of poor performance of the banks.

There is no need to model explicitly the chain of events that trigger a fundamental bank run. When agents perceive a bad signal on the bank return during period 1, they may rationally decide to withdraw early. As Jacklin and Bhattacharya (1988) point out, the information on future returns modifies the relevant incentive-compatible constraint, and therefore agents of type 2 may prefer to withdraw at $t = 1$. Fundamental bank runs can thus provide an efficient mechanism for closing down inefficient banks. However, in practice it may be difficult to distinguish ex ante a fundamental bank run from a speculative one.

Gorton (1985) suggests a simple model in which agents obtain at $t = 1$ some information about the expected return on the bank's assets at $t = 2$. If the expected return on deposits is lower than the expected return on currency, there is a fundamental bank run. This simple structure enables Gorton to provide a rationale for the suspension of convertibility when there is asymmetric information on the date 2 return on deposits. To do so, it suffices to assume that the banks are able, by paying a verification cost, to transmit the true value of the expected return to depositors. If there is a bank run on a solvent bank, the bank is able to suspend convertibility and pay the verification cost, which stops the run (see problem 7.8.3 for a simplified version of Gorton's model). This explains "a curious aspect of suspension... that despite its explicit illegality, neither banks, depositors nor the courts opposed it at any time" (Gorton 1985, 177).

Chari and Jagannathan (1988) consider a model close to the Diamond-Dybvig (1983) model, in which they introduce a random return on investment that may be observed by some of the type 2 agents. If the signal that the agents receive indicates a poor performance, this makes them prefer the type 1 deposit, which is less sensitive to the variations in period 2 returns. The agents observe the total amount of withdrawals and use this information to decide their own action (withdraw or wait). Since the proportion of type 1 agents is not observable, it is impossible for an uninformed type 2 depositor to distinguish if the origin of the large withdrawal he observes comes from informed type 2 agents or simply from a large proportion of type 1 agents. The rational expectations equilibrium that is obtained therefore combines fundamental bank runs (the ones justified by the poor performance of the bank) and speculative bank runs, which develop as in the Diamond-Dybvig model but are here triggered by the fear of poor performance, anticipated by informed depositors.[14] Notice that this model assumes that the management keeps the bank open even if this implies a decrease in its net wealth. If management has incentives to maximize the bank's total value, then fundamental bank runs will never occur, and speculative bank runs will not occur either. However, under limited liability, the bank management may have an incentive to keep the bank running if this increases shareholder value, even if the total value of the bank is decreased.[15] Provided that the bank is worth more dead than alive, bank runs are efficient because they correct, at least partly, the incentives of management to forbear. Closing down the bank will be efficient whenever, given a signal $S$ on the future return for the long-run technology, the liquidation value is larger than the expected conditional return:

$$\ell > E(\tilde{R}|S).$$

The fact that deposits are demandable is therefore a key characteristic that would be desirable even if all consumers were of equal type ($\pi_1 = 1$). Thus in the Chari and Jaggannathan model we have a disciplining role for demandable debt, a point subsequently developed and emphasized by Calomiris and Kahn (1991), Qi (1998), and Diamond and Rajan (2001).

Three other contributions extend the Diamond-Dybvig framework to a context of random returns on the bank's assets. They do not focus directly on bank runs but rather on the risk sharing provided by demand deposit contracts in this framework.

Jacklin and Bhattacharya (1988) address the question of the relative performance of equity versus demand deposit in banks' financing, in a context where some type 2 agents are informed on the bank's future return. In an equity economy, equilibrium prices are fully revealing; in a demand deposit equilibrium, with suspension of convertibility, there is rationing on type 1 deposits, so these deposits are shared between type 1 and informed type 2 agents. Comparison of the relative performances for specific values of the model's parameters shows that for a lower dispersion of returns, demand deposits perform better, whereas for a large dispersion, equity financing is preferred. Still, the demand deposit contract can be improved if it is required to remain incentive-compatible after type 2 agents become informed (Alonso 1991).

Gorton and Pennacchi (1990) also consider that some of the type 2 agents are informed,[16] but the agents' behavior is not competitive, so the prices of an equity economy are not fully revealing. Informed traders benefit from trading in the equity market. Demand deposit contracts emerge then because they are riskless, so their value cannot be affected by informed trading. Gorton and Pennacchi establish that in equilibrium uninformed traders invest in deposits and informed traders invest in equity of the financial intermediary, so there is no other market for equity. In that way, the authors elaborate on Jacklin's (1987) contention that equity could perform as well as demand deposit contracts.

Allen and Gale (1998) consider the efficiency properties of bank runs in the case where the long-run technology yields a random return. They argue that if liquidation of the long-run technology is impossible ($\ell = 0$), then bank runs can be best information constrained efficient, thus allowing for optimal risk sharing between early and late withdrawing depositors. This is the case because, when returns on the long-run technology happen to be low, patient consumers will have an incentive to run the bank. By so doing, the remaining patient consumers withdrawing at time $t = 2$ will have a larger consumption, and this will go on until the equality $C_1 = C_2$ is reached and the incentives to run the bank disappear. Thus, as Allen and Gale argue, the problem is not bank runs per se but rather the cost of liquidating the long-run technology. When the cost of liquidation is taken into account, runs are not limited to a fraction of patient customers, and the best cannot be reached any longer.

## 7.5 Interbank Markets and the Management of Idiosyncratic Liquidity Shocks

Until now the discussion has maintained the convenient fiction that there is only one collective mutual bank. This section drops this assumption and focuses on the problems that arise precisely because of a multiplicity of banks.[17] These problems are again based on the coexistence of demand deposits on the liability side with nonmarketable loans on the asset side.

### 7.5.1 The Model of Bhattacharya and Gale

The Bhattacharya-Gale (1987) model is a variant of Diamond-Dybvig in which it is assumed that $\ell = 0$ (liquidation is impossible) and the consumption good cannot be stored, so there are no bank runs. The novelty is that there are now several banks confronted with i.i.d. liquidity shocks in the sense that their proportion of patient consumers (who withdraw at date 1) can be $\pi_L$ or $\pi_H$ (with $\pi_L < \pi_H$), with respective probabilities $p_L$ and $p_H$. Assume there are a large number of banks, so liquidity shocks of banks are completely diversifiable; the proportion of banks with few ($\pi_L$) early withdrawals is exactly $p_L$.

In an autarkic situation (absence of trade between the banks), each bank is completely restricted by its ex ante choice of investment $I$. The bank can offer only contingent deposit contracts

$$C_1(\pi) = \frac{1-I}{\pi}, \qquad C_2(\pi) = \frac{IR}{1-\pi},$$

where $\pi$ can be $\pi_L$ or $\pi_H$. Therefore, depositors bear the liquidity risk of their bank. This risk can be eliminated by opening an interbank market, which can decentralize the optimal allocation, obtained by solving

$$\begin{cases} \max\limits_{I, C_1^k, C_2^k} \sum\limits_{k=L,H} p_k[\pi_k u(C_1^k) + (1-\pi_k)u(C_2^k)] \\ \sum\limits_{k=L,H} p_k \pi_k C_1^k = 1 - I, \\ \sum\limits_{k=L,H} p_k(1-\pi_k)C_2^k = RI, \end{cases}$$

where $(C_1^k, C_2^k)$ is the deposit contract offered by a bank of type $k$, $k = L, H$.

The solution to this problem satisfies

$$C_1^k \equiv C_1^* = \frac{1-I^*}{\pi_a}, \qquad C_2^k \equiv C_2^* = \frac{RI^*}{1-\pi_a} \qquad (k = L, H), \tag{7.13}$$

where $\pi_a = p_L\pi_L + p_H\pi_H$ is the average proportion of early withdrawals across all banks. Equations (7.13) show that consumers are now completely insured against the liquidity risk faced by their bank: $C_1^*$ and $C_2^*$ are independent of $k$.

### 7.5.2 The Role of the Interbank Market

The implementation of this allocation by the interbank market is realized as follows. Banks of type $k = L$ face fewer early withdrawals than the average; therefore they have excess liquidity $M_L = 1 - I^* - \pi_L C_1^*$. On the contrary, banks of type $k = H$ have liquidity needs $M_H = \pi_H C_1^* - (1 - I^*)$. Conditions (7.13) imply that on aggregate, supply and demand of liquidity are perfectly matched:

$$p_L M_L = p_H M_H.$$

At date 2, banks of type $k = H$ will have excess liquidities, which they will use to repay the interbank loan they obtained at date 1. The interest rate $r$ on the interbank market will thus be determined by equaling this repayment with $(1 + r)M_H$, where $M_H$ is the amount of the loan obtained at date 1:

$$(1 + r)M_H = RI^* - (1 - \pi_H)C_2^*.$$

Computations yield the following:

$$1 + r = \left(\frac{\pi_a}{1 - \pi_a}\right)\left(\frac{I^*}{1 - I^*}\right)R. \tag{7.14}$$

### 7.5.3 The Case of Unobservable Liquidity Shocks

Bhattacharya and Gale (1987) also study the more difficult case in which the liquidity shock (the type $k$ of the bank) and the investment of the bank in the illiquid technology are not publicly observed. In that case, the best allocation derived earlier will typically not be implementable. Suppose, for instance, that the equilibrium interest rate $r$ on the interbank market (defined by (7.14)) is smaller than $R - 1$. Then all banks will have an interest in declaring that they are of type $H$ because this will entitle them to an interbank loan that they will use for investing in the illiquid technology, obtaining a positive excess return $R - (1 + r)$. To prevent this, the second-best solution will involve imperfect insurance of depositors (in the sense that $C_1^L < C_1^H$ and $C_2^L > C_2^H$) and overinvestment with respect to the best solution $I^*$. This case occurs when liquidity shocks are small, which can be shown to imply that the interest rate $r$ defined by (7.14) is smaller than $R - 1$.

Symmetrically, when liquidity shocks are large, then $1 + r > R$, and the second-best solution involves, on the contrary, underinvestment and reverse ordering of the consumption profiles ($C_1^L > C_1^H$ and $C_2^L < C_2^H$).[18]

Freixas and Holthausen (2005) extend the analysis of the interbank market to cross-country bank lending. They assume that the main barrier to building an integrated international interbank market is the presence of asymmetric information between different countries, which may prevail in spite of monetary integration or successful currency pegging. In order to address this issue, they consider a Bryant-Diamond-Dybvig model and find not only that an equilibrium with integrated markets need not always exist but also that when it does, the integrated equilibrium may coexist with one of segmentation of the interbank market.

## 7.6 Systemic Risk and Contagion

Systemic risk is usually defined as any risk that may affect the financial system as a whole (de Bandt and Hartmann 2002). It may originate either in the banking industry or in the financial markets. We focus here on the mechanisms that trigger a systemic banking crisis.

A systemic crisis may develop either as a result of a macroeconomic shock or as a result of contagion. The models we examined in section 7.2 allow us to understand why an aggregated liquidity shock, if sufficiently large, may trigger a systemic crisis. This will occur if all banks have to liquidate the long-run technology up to the point where residual depositors have incentives to join the run. In the same vein, efficient bank runs (section 7.4) may occur at the aggregated level when the returns on banks' assets are highly correlated. Patient depositors will prefer to withdraw at time $t = 1$, which may or may not be efficient depending on the information structure and the level of the liquidation cost.

Apart from liquidity and productivity shocks, a third macroeconomic shock, the one of the exchange rate, is particularly relevant to the study of systemic risk; the East Asian crises pointed out the close links between the currency crisis and banking crisis within a country.

Liquidity and productivity shocks are no longer considered the main sources of systemic risk because Central Banks have tools to accommodate such shocks.[19]

On the other hand, a systemic crisis may be the result of contagion. The failure of a bank may propagate to the whole banking industry. Clearly, the macroeconomic environment will be important in setting the conditions for this domino effect to occur, since a lower yield on loans, due to high loan losses, depletes banks capital and reduces the buffers each bank has to cope with risks.

Contagion may occur through four different (nonexclusive) channels:

- Change in expectations of investors
- Large-value payment systems

- Over-the-counter operations (mainly on derivatives)
- Interbank markets

### 7.6.1 Aggregate Liquidity and Banking Crises

Diamond and Rajan (2005) study the effect of bank runs on the aggregate demand of liquidity and on interest rates, and find that the very existence of a demand deposit structure allowing the banker to commit not to renegotiate the deposit repayment (see section 7.3), might aggravate banking crises. This will occur because bank runs might destroy liquidity rather than create it. As mentioned, runs are triggered by the bank's inability to repay depositors. So, in contrast to Diamond-Dybvig (1983), Diamond and Rajan (2001; 2005) find that bank runs are triggered by the asset side of the bank's balance sheet.

The basic mechanism of Diamond and Rajan (2005) is somewhat related to Fisher's (1933) debt deflation mechanism, but whereas Fisher emphasized the effect a decrease in the value of assets would have on the supply of credit (because of the lower value of collateral), Diamond-Rajan find a decrease in the value of assets will trigger more bank runs (fig. 7.2).

It is clear that liquidity shortages increase interest rates and that this in turn decreases the value of assets promising future cash flows, so the key issue is to understand why bank runs may exacerbate liquidity shortages, that is, why the excess demand for liquidity may be increased by bank runs.

In order to understand how bank runs may absorb liquidity, it is useful to briefly describe some of the aspects of the Diamond-Rajan (2005) model, which inherits the main characteristics of their previous work.

There are three dates, $t = 0, 1, 2$, and three types of agents: investors, bankers, and entrepreneurs, all of them risk-neutral. Investors consume at time $t = 1$, and entrepreneurs and bankers consume at time $t = 2$.

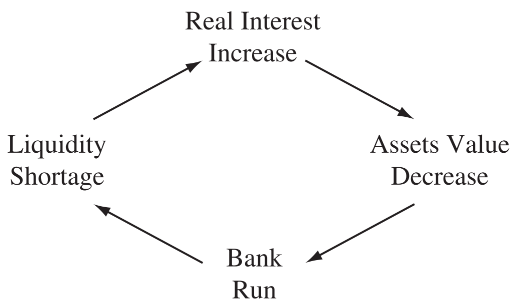



**Figure 7.2**
Debt deflation.

Entrepreneurs' projects are risky in the sense that the expected cash flow $y$ may be obtained either at time $t = 1$ (early project) or at time $t = 2$ (late project).

As mentioned in section 7.3, banks are partly financed by deposits and any shortfall on repayment triggers a bank run. This depends on the proportion of early versus late projects the bank faces. But once a bank run occurs, inefficient liquidation will take place, so that both early and late projects will be liquidated. Assume that liquidation of a project produces an amount $\alpha V_1$ at time $t = 1$ and $\alpha V_2$ at time $t = 2$.

On the other hand, early repayment as well as continuation of late projects leads to a repayment $R$ to the bank and a profit $y - R$ to the entrepreneur at the time when $y$ is realized. Since $\alpha(V_1 + V_2)$ is lower than $y$, project liquidation leads to a destruction of resources: the amount $y - R$ lost by the entrepreneur plus the amount lost by the bank. If the equilibrium interest rate is $\rho$, the bank's loss equals

$$R - \alpha\left(V_1 + \frac{V_2}{1+\rho}\right).$$

Absent a bank run, when the bank is confronted with a high proportion, $\mu$, of early repayments, its cash flow $\mu R$ is sufficient to pay all the depositors. To understand why bank runs may deplete liquidity and therefore increase the real interest rate, consider a bank on the verge of insolvency (the value of its assets just equals its depositors' claims) that is confronted with a proportion $\mu$ of early projects (and $1 - \mu$ of late projects) and therefore has to liquidate all its late projects.

If no bank run occurs, the liquidity generated by the bank $\mu R$ is absorbed by the depositors, but the bank will have to obtain additional liquidity from other banks in order to pay the remaining depositors $d - \mu R$. This will be possible by liquidating and selling the projects of the late entrepreneurs in proportion $1 - \mu$. Since the liquidated assets are assumed to be transparent, their price is the discount rate

$$\frac{1}{1+\rho}.$$

Hence, the bank obtains $\mu R$ from its projects and demands

$$(1-\mu)\frac{\alpha V_2}{1+\rho}$$

from other banks to satisfy its depositors. Consequently, the banks' depositors absorb

$$(1-\mu)\frac{\alpha V_2}{1+\rho}$$

liquid assets. Still, the bank generates liquid time $t = 1$ cash flows for $\mu(y - R)$ entrepreneurs.

By contrast if there is a bank run, depositors will repossess the bank's assets and thus destroy liquidity. Indeed, while the amount $\alpha V_2$ will still be sold, keeping the demand for liquidity unchanged, the bank's liquidity is lower because $\mu\alpha V_1 < \mu R$, and so is the external liquidity because the amount $\mu(y - R)$ that the early entrepreneurs received is destroyed.

Consequently, bank runs deplete liquidity. This leads to an increase in $\rho$, which in turn leads to additional (fundamental) bank runs.

Note that an increase in the interest rate $\rho$ implies a higher demand for liquidity. This is indeed the case because in order to obtain a unit of liquidity, a bank has to liquidate a fraction $\gamma$ of its late projects such that $\gamma \cdot \alpha V_2/(1 + \rho) = 1$. So a higher $\rho$ implies a higher value of $\gamma$.

Diamond and Rajan (2005) study the equilibrium and remark that multiple interest rates might be compatible with liquidity market clearing. So the coordination of beliefs will be the determinant for market allocation.

Finally, it is interesting to note that liquidity is defined here in terms of goods and that depositors' claims are also defined in real terms, so the introduction of money and prices, as in Allen and Gale (1998), may modify the model's implications.

### 7.6.2 Payment Systems and OTC Operations

In spite of their apparent differences, payment systems and over-the-counter (OTC) operations share some similarity because they both give rise to credit risk. We develop the analysis focusing on the payment system, as Freixas and Parigi (1998) do, and then we reinterpret the results in the context of OTC operations.

The motivation for the analysis of payment systems stems from the change that has taken place in this field. Banking authorities in developed countries have progressively introduced Real Time Gross System (RTGS), where each individual payment order becomes irrevocable in real time provided the sending bank has a sufficient balance on its account at the Central Bank. RTGS has completely replaced traditional net systems, where payment orders of each institution are netted against all the others, and a net credit or debit position is obtained for each of them. The issue that arises is therefore under what conditions this substitution of RTGS for net payment systems is efficient. Contagion is a key issue because in a net payment system banks are net creditors (debtors) of one another as a result of the operations their clients realize. The bankruptcy of a bank means that it will default on its payments to other banks, thus affecting their solvency.

Freixas and Parigi (1998) consider a Diamond-Dybvig economy with two locations where patient agents are uncertain about the location where they want to consume. As a consequence, there is a demand for a payment system, and the properties of net systems and RTGS could be compared in terms of expected future consumption.

In a Diamond-Dybvig framework the disadvantages of RTGS become apparent. Banks have to keep more reserves and are unable to invest as much as they would like in the more profitable long-run technology. As a consequence, when returns are certain, a net system always dominates. Still, when returns are uncertain, a net payment system causes the bankruptcy of one institution to propagate to others, whereas RTGS provides a perfect fire wall. The comparison of net system versus RTGS boils down to a trade-off between efficiency and safety.

The choice of a payment system, net or gross, depends on characteristics of the environment, in particular, how large the opportunity cost of holding reserves and how large the probability of a bank failure. RTGS is shown to be preferred when the probability of bank failures increases, when transaction volume increases, and when the opportunity cost of liquid reserves increases. These features may help to explain the recent move from net to gross systems.

An analysis of OTC versus market-based operations on derivatives could be pursued simply as a reinterpretation of the preceding results. Organized markets require traders to post collateral. In a Diamond-Dybvig framework this implies that banks invest less in the long-run technology in order to hold more liquidity. Hence OTC operations are equivalent to the net payment system, whereas organized market operations are the equivalent of RTGS.

A consequence of this analysis is that banks prefer net payment systems and OTC operations, whereas regulators prefer RTGS and organized (collateralized) markets. Holthausen and Rønde (2000) are concerned precisely with this issue.

### 7.6.3 Contagion through Interbank Claims

Contagion and systemic risk are explored by Allen and Gale (2000) and Freixas, Parigi, and Rochet (2000). Both contributions aim at explaining how contagion works, and they share some conclusions:

- The level of buffers each bank has, broadly defined as capital, subordinated debt, or the claim of patient depositors is a key determinant of contagion.
- The way in which the failure of a bank is resolved has an impact on the propagation of crises.
- The system of cross-holdings of assets and liabilities among banks, including those implicit in payment system arrangements (bilateral or multilateral credit lines), is essential in triggering systemic crises.
- The specific architecture of this system of cross-holdings matters. A system where each bank borrows only from one bank is more fragile than a system where the sources of funds are more diversified.

The focus of the two articles is different. In Allen-Gale the crisis is basically a liquidity crisis; in Freixas-Parigi-Rochet the crisis stems from a coordination problem. Although both articles conclude that a Central Bank is needed, they differ in the policy prescription. Allen-Gale advise preventing contagion by injecting liquidity globally (through repos or open market operations, as suggested by Goodfriend and King (1988)). Freixas-Parigi-Rochet require that liquidity be provided to a specific financial intermediary.

Both groups draw on the Diamond-Dybvig framework with a short-run storage technology and a long-run technology that is costly to liquidate. This framework in extended in order to consider $N$ commercial banks $(i = 1, \dots, N)$.

**Contagion in the Allen-Gale Model**

The model of Bhattacharya and Gale (1987) is the starting point of the Allen-Gale analysis. Absent aggregate liquidity shocks, (and abstracting from the issue of asymmetric information regarding liquidity shocks), efficient allocation corresponds to a consumption profile $(C_1^*, C_2^*)$ that is independent of the banks' idiosyncratic liquidity shocks. To achieve this allocation, the banks with liquidity needs borrow from the ones with excess liquidity. The interbank market, credit lines, or cross-holdings of deposits among banks are possible mechanisms that implement efficient allocation by transferring liquidity among banks. It is therefore necessary that at least one of these mechanisms be in place for efficient allocation to be implemented.

Starting from this setup of aggregate certainty, Allen and Gale introduce the following perturbation: with probability zero, each bank faces the average demand for liquidity except one bank (say, bank 1), which faces the average demand plus $\varepsilon$. In this event, the aggregate liquidity shock is larger, and this implies that the allocation $(C_1^*, C_2^*)$ is not feasible anymore. The shock, initially affecting bank 1, will propagate and affect all the other banks through the very same interbank links that were designed to channel liquidity efficiently and implement $(C_1^*, C_2^*)$.

The propagation mechanism is quite simple. Assume that $N = 3$, that all depositors are treated equally, and that each bank borrows from its neighbor, as in the credit chain of figure 7.3. An $\varepsilon$ liquidity deficit forces bank 1 to borrow this amount from bank 2. Then bank 2 becomes illiquid and has to borrow from bank 3, which in turn will borrow from bank 1. But since bank 1 is already illiquid, it will be forced to liquidate its long-run assets at a cost, and this may trigger its bankruptcy. Then the value of bank 3's assets may become insufficient to cover its liquidity needs. Bank 3 is therefore forced to liquidate all its assets. The mechanism continues until all banks fail.

For bankruptcy and propagation to occur, banks have to have a high level of debt and a low level of buffers. Clearly, if the losses derived from the liquidation of assets at time $t = 1$ could be borne by equity holders or by patient depositors (which would

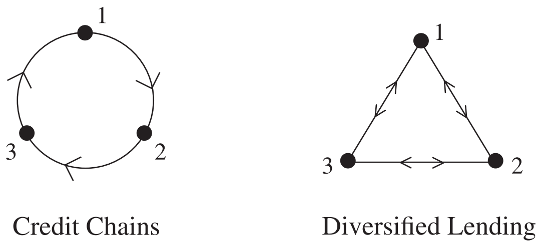



**Figure 7.3**
Two examples of interbank borrowing architecture.

still reach a higher utility level by waiting than by forcing liquidation), the process would stop.

The extent of the contagion will also depend on the architecture of interbank borrowing. To see why, notice that although bank 1 does not have a sufficient buffer to cope with the liquidity shock, it could be the case that bank 1 and 3 together do. But the importance of the liquidity demand from one bank to another depends on the architecture of interbank borrowing claims. Comparing the two extreme cases of a credit chain and diversified lending, where each bank has deposited the same amount in every other bank (see fig. 7.3), an $\varepsilon$ shock for bank 1 translates into the same shock for bank 3 in the first case and only to a shock of $\varepsilon/2$ in the diversified lending case. As a consequence, the same buffer level is much more effective in avoiding a domino effect in the complete claims or diversified lending case than in the credit chain case.

**Contagion in the Freixas-Parigi-Rochet Model**

Freixas, Parigi, and Rochet extend the Freixas and Parigi (1998) framework in order to explore contagion through interbank markets. The question they raise is whether the interbank market is a sufficient guarantee against a liquidity shock affecting one bank. Consider an economy where depositors travel and are uncertain about their consumption location. Depositors have two ways to transfer money at the right location: either they cash their deposit at time $t = 1$ and travel to their consumption location with cash in their pockets, or they travel with a check payable at the location of destination (representing the use of the interbank market). In a Diamond-Dybvig framework, the first solution implies that too much cash is used. Therefore, as in Allen and Gale (2000), access to interbank credit reduces the cost of holding liquid assets.

Still, depositors have to make a strategic choice, whether to travel with cash or with a check. This depends on their assessment of the probability that the check will be repaid by the bank at the location where they travel. If they do not trust the bank at their destination, they will withdraw cash at their own bank, which generates externalities across banks. Consequently, two equilibria coexist: the efficient one, where depositors travel with a check to their destination and long-run investment is preserved, and an inefficient one, called the gridlock equilibrium, without any interbank lending. In this equilibrium the expectations of each bank, that there will be no liquidity available, are self-fulfilled. Freixas-Parigi-Rochet analyze the possibilities of contagion when one of the banks is closed by supervisory authorities. As in Allen-Gale, they show that the likelihood of contagion depends on the architecture of interbank payments. A diversified lending situation is less fragile than a credit chain situation. They also show that some banks may be more likely than others to provoke contagion.

## 7.7 Lender of Last Resort: A Historical Perspective

Following the ideas of Bagehot (1873), the Central Banks of most countries have adopted a position of lender of last resort (LLR) in the sense that under certain conditions, commercial banks facing liquidity problems can turn to them for short-term loans. Although this definition is functionally correct, it amalgamates completely different interventions, which reflect different strategies on behalf of the Central Banks.

To begin with, the provision of liquidity to the market on a regular basis, through auctions or a liquidity facility is the responsibility of the LLR. It corresponds to lending (anonymously) to the market against good collateral and is one of the main tasks of a Central Bank.

Second, LLR lending could be directed at a specific group of solvent illiquid institutions. This type of facility could be arranged through a discount window or as an exceptional short-term loan. The term *emergency lending assistance* has been used to refer to this situation.

Third, LLR lending could be aimed at an insolvent institution, in which case the term *lending* barely conceals the fact that the LLR is injecting capital.

Notice that an institution could be ex ante solvent and ex post insolvent, or the other way around, so the decision of the LLR has to be taken with incomplete information. Also, since the rescue of a bank involves, directly or indirectly, taxpayer funds, a number of rescue operations are done under the pretense of granting a loan to an ex ante solvent institution so as to avoid political pressure.

Finally, the LLR could play a role as a crisis manager, involving no lending at all, but necessary so as to coordinate agents that would otherwise liquidate an otherwise solvent institution.

The discussion here examines various arguments justifying the role of the LLR. We start with a quick review of Bagehot's doctrine and of the criticisms that have addressed it. Then we examine the practice of LLR intervention as well as the historical evidence about the efficiency of the LLR system in preventing systemic risk.

### 7.7.1 Views on the LLR Role

The idea that market mechanisms cannot insure against liquidity shocks has to be based on arguments that sustain the existence of a market failure. The classical argument was forcefully put forward by Bagehot (1873), emphasizing the difficulty a bank will face if it must transmit credible information to the market during a crisis. In his words, "Every banker knows that if he has to prove that he is worthy of credit, however good may be his argument, in fact his credit is gone" (68). The classical price argument, which implies that an increase in interest rates would compensate lenders for the increased risk they take when lending to a bank facing a crisis, may in fact act as a signal of an unsound position and therefore discourage potential lenders. Market failure can thus be traced to asymmetric information on the banks' solvency.

The idea of the Central Bank's acting as the LLR is associated with the work of Bagehot. He argues that

- the LLR has a role in lending to illiquid, solvent financial institutions;
- these loans must be at a penalty rate, so that financial institutions cannot use the loans to fund their current lending operations;
- the lending must be open to solvent financial institutions provided they have good collateral (valued at prepanic prices);
- the LLR must make clear in advance its readiness to lend any amount to an institution that fulfills the conditions on solvency and collateral (credibility).

There have been some criticisms of Bagehot's view:

- Goodhart (1987; 1995) writes that the clear-cut distinction between illiquidity and insolvency is a myth because the banks that require the assistance of the LLR are already under suspicion of being insolvent. The existence of contagion is the additional argument that may induce the systematic rescue of *any* bank.
- Goodfriend and King (1988) argue that LLR functions must be restricted to the use of open market operations. Humphrey (1986) claims that this would have been Bagehot's viewpoint had he known of open market operations.
- Proponents of free banking do not challenge the existence of market failure but suggest that the market would still lead to a better allocation than a public LLR would.
- Repullo (2005) shows that when banks' risk-taking decisions are explicitly taken into account, the existence of the LLR does not increase the incentives to take risk, whereas penalty rates do.

A second issue that evokes disagreement is the classical view that the rules governing LLR behavior should be clearly stated. Most of the time, this is opposed by Central Banks. In the United States, for instance, the Fed has always stressed that discounting is a privilege, not a right. The supporters of this view say that ambiguity in the policy will help to bring some market discipline (in contradiction to the view, also held by Central Banks, that disclosure could have a destabilizing effect on the payment system). In fact, the effect of ambiguity is a transfer of wealth from small to large banks, because there is no ambiguity that large institutions are "too big to fail." Thus, ambiguity is to some extent illusory and is equivalent to repaying all large banks' liabilities and rescuing only the solvent ones among the small banks (if they are able to prove that they are solvent).

It is clear that these positions should result from social welfare maximization, taking into account asymmetric (or costly) information and all the externalities that the behavior of the LLR may have: contagion, panics, and effects on securities markets, as well as the moral hazard problem. Therefore, at least theoretically, the differences among the views of the LLR's role could be traced to differences in the appreciation of, say, the social cost of individual bank failure, bank panics, and contagion effects.[20]

### 7.7.2 Liquidity and Solvency: A Coordination Game

Rochet and Vives (2004) revisit Bagehot's assertion that the LLR should provide liquidity assistance to illiquid but solvent banks. This view had been criticized, in particular by Goodfriend and King (1988), on the grounds that in Bagehot's time, interbank and money markets were underdeveloped and lacked the efficiency that now exists most developed countries, where a solvent bank will be able to find liquidity assistance. So Central Bank lending is not needed anymore and should even be prohibited in the case where Central Banks are prone to forbearance, especially under political pressure.

Rochet and Vives (2004) are able to rejuvenate Bagehot's doctrine by showing that even modern, sophisticated, interbank markets will not necessarily provide liquidity assistance to a solvent bank; in other words, a solvent bank can indeed be illiquid. The reason is a potential coordination problem between investors (typically other banks) who may have different opinions about the solvency of the bank requiring liquidity assistance. In this context, the decision of each individual investor to renew, say, a large certificate of deposit will be based not only on his own opinion about the bank's solvency (fundamental risk) but also on his assessment about the decisions of other investors (strategic risk). The reason is that a large withdrawal by other investors (the modern form of a bank run, illustrated by what happened to Continental Illinois in 1985) might force the bank to liquidate some of its assets at a loss ("fire sales") or simply borrow from other investors at a penalty rate or for an amount strictly lower than the value of the assets that the bank can offer as collateral ("haircut"). As a result, liquidity problems might provoke the insolvency of an initially solvent bank. So, the optimal decision of each individual investor (renew the CD or not) depends on his expectation of what others will do. The higher the proportion of other investors that are ready to lend, the more likely each individual investor is to lend. Thus there is strategic complementarity between investors' decisions. In the limit case of perfect information, where all investors know exactly the value of the bank's assets, this strategic complementarity leads to the possibility of multiple equilibria. In fact, when the value of the bank's assets is either very large or very low, there is still a unique equilibrium, characterized by lending in the former case and withdrawing in the latter, but in the intermediate region (where the bank is solvent but not "supersolvent") there are two equilibria resembling bank run models à la Diamond-Dybvig.

By using global game techniques as in Morris and Shin (2000), Rochet and Vives (2004) show that in the imperfect information case, where investors have different opinions about the bank's solvency, there is a unique equilibrium. In this equilibrium, the fraction of investors who withdraw continuously decreases (as the value of the bank's assets increases) from 1 (when the bank is insolvent) to 0 (when the bank is "supersolvent"). There is a critical value of bank's assets (between the solvency and supersolvency thresholds) such that whenever the value of the bank's assets falls below this threshold (which is strictly higher than the solvency threshold) the proportion of investors who withdraw becomes so large that the bank will not find enough liquidity support from the interbank markets. Thus there is a possibility that a solvent bank might be illiquid.

Rochet and Vives (2004) then use their model to show how a Bagehot LLR might increase social welfare by avoiding inefficient closures of solvent banks. This is particularly clear if the Central Bank has access to regulatory information that allows it to assess properly the value of the bank's assets.[21] Interestingly, it would not help in this case to instruct the supervisor (or the Central Bank) to disclose this information publicly, because this would lead to a multiplicity of equilibria (as in the perfect information case) and thus to some form of instability. Thus disclosure of supervisory information is not a good idea, implying a caveat on the promotion of a transparency policy.

Finally, Rochet and Vives (2004) discuss how other regulatory instruments like solvency and liquidity requirements can complement the LLR activities of the Central Bank so as to provide appropriate incentives for banks' shareholders while avoiding too many bank closures.

An alternative link relating liquidity and bank crises is put forward by Acharya and Yorulmazer (2006). They consider a framework similar to that of Diamond and Rajan (2005) and analyze the effect of bank failures on the supply of assets available for acquisition rather than on the interest rate. Since bank failures decrease available liquidity, a bank bankruptcy leads to a decrease in asset prices. So, bank failures are a cumulative phenomenon: they depress the price of banks' assets and this in turn increases the number of bank failures.

Confronted with this, the regulator has the option to intervene in an number of ways. Acharya and Yorulmazer show that when the number of bank failures is low, the optimal ex post policy is not to intervene, but when this number is sufficiently large, the regulator will optimally adopt a mixed strategy and choose randomly which banks to assist. This mixed strategy is justified by the fact that the regulator sets a liquidity target that may be lower than the amount required to bail out all banks. This liquidity target limits banks' assets sales and the fall of asset prices that could severely distort asset allocation (as inefficient users of the asset who are liquidity-long would then end up owing it all) and thus prevents additional bank bankruptcies.

Finally, Acharya and Yorulmazer remark that the policy of liquidity assistance to surviving banks in the purchase of failed banks is equivalent to an ex post bailout policy but provides better incentives ex ante.

### 7.7.3 The Practice of LLR Assistance

The practice of Central Banks in injecting money through regular repo auctions or other procedures falls outside the scope of this discussion. This section is concerned with the lending provided by the LLR (usually Central Banks) to individual institutions.

Casual empirical evidence seems to indicate that in general LLR assistance is directed to insolvent banks, or at least to banks that ex post are insolvent. In particular, "too big to fail" banks are always rescued, as illustrated by the Continental Illinois, Crédit Lyonnais, or Banesto events, even if the last two were insolvent as a result of fraudulent operations.

Although it is always difficult to distinguish between illiquid and insolvent banks, in the United States the discount window has been open to thrift institutions with the worst CAMELS ratings.[22] Thus, as Kaufman (1991) states, it has provided a disguised way to bail out insolvent banks.

The empirical analysis of bank resolutions confirms the idea that LLR lending is often directed to bail out banks. Goodhart and Schoenmaker (1995) support that view with evidence on the effective bailout policies of Central Banks all over the world. Out of a sample of 104 failing banks, 73 ended up being rescued and 31 liquidated. Since the Central Bank is in charge of an orderly liquidation, it is no surprise that absent institutional structures that would allow for an orderly closure of financial institutions, the Central Bank prefers to rescue them rather than risking a contagion.

This view is confirmed by Central Bank complementary measures when a bank is liquidated. Case studies of bank failures, for instance, the Herstatt bank in 1974 or Barings in 1995, show that the Central Bank was ready to lend to any bank that would have been hit by the bankruptcy, in order to limit a contagion effect.

The fact that LLR lending tends to bail out a bank or help it forbear is an important difference between nineteenth-century and modern practices, which may clarify the debates about the LLR. The existence of a closure/bailout policy is an important part of the overall regulation of the banking industry. The fact that it may be implemented by the LLR is, to say the least, confusing.

### 7.7.4 The Effect of LLR and Other Partial Arrangements

The evidence on the LLR mechanism points unambiguously to the conclusion that it has helped to avoid bank panics. Miron (1986), Bordo (1990), and Eichengreen and Portes (1987), among others, support this view.[23] Their results were obtained either by examining the effects of creating the LLR in a given country, thus assuming the ceteris paribus clause for changes in the banking system, or by comparing different countries and assuming that the ceteris paribus clause is satisfied for other factors affecting the frequency of financial panics. By monitoring the banks' solvency and payment system, the LLR mitigates the risk of contagion, the importance of which has been emphasized by Aharony and Swary (1983), Humphrey (1986), Guttentag and Herring (1987), Herring and Vankudre (1987), and Saunders (1987).

The evidence obtained by Miron (1986) on the effects of creating the Federal Reserve Board in the United States shows the importance it has had on limiting bank runs. Prior to the Fed's founding, autumn and spring were the stringent money quarters, during which panics tended to occur. The founding of the Fed provided the U.S. economy with an LLR, and the frequency of bank panics immediately decreased. The change in seasonal patterns for both interest rates and the loan-reserve ratio confirms the importance of the founding of the Fed as a way out of seasonal liquidity–triggered bank panics. Between 1915 and 1928 the banking system experienced no financial panics, although several recessions occurred during the periods 1918–1919, 1920–1921, 1923–1924, and 1926–1927. The 1920–1921 recession was quite severe.[24]

On the other hand, the panics observed during the 1929–1933 period may be considered as providing an argument against the effectiveness of the LLR policy. Still, it is clear that during that period the Fed did not conduct the open market operations necessary to provide banks with adequate reserves. According to Friedman and Schwartz (1963), the series of bank failures that produced an unprecedented decline in the money stock could have been prevented. Meltzer (1986) makes the same point: "The worst cases of financial panics arose because the Central Bank did not follow Bagehotian principles" (83).

Bordo (1990) examines the changes that occurred in the United States and the United Kingdom before and after the creation of an LLR system. Before 1866 the Bank of England tended to react by protecting its own gold reserves, which could even worsen panics. After that date, the Bank of England adopted Bagehot's policy and thus "prevented incipient crises in 1878, 1890, and 1914 from developing into full-blown panics, by timely announcements and action" (23). Bordo compares the two countries during the 1870–1913 periods and sees striking similarities in their business cycles: similar declines in output, price reversals, and declines in money growth. Still, the United States had four panics during this period while the United Kingdom had none. Evidence on Germany, Sweden, and Canada supports analogous views (Bordo 1986; Humphrey and Keleher 1984).

## 7.8 Problems

### 7.8.1 Bank Runs and Moral Hazard

Consider a Diamond-Dybvig economy with a unique good and three dates, where banks managers have a choice of the technology they implement. This choice is unobservable and consists in investing one unit in either project G or B, where project G yields $G$ with probability $p_G$ and zero otherwise, and project B yields $B$ with probability $p_B$ and zero otherwise, where $G < B$, and $p_G G > p_B B$.

A continuum of agents is endowed with one unit at time $t = 0$. Of these agents, a nonrandom proportion $\pi_1$ will prefer to consume at time $t = 1$, and the complementary proportion $\pi_2$ will prefer to consume at time $t = 2$.

The agents' utility function is

$$\begin{cases} U(C_1) & \text{for impatient consumers,} \\ \rho U(C_2) & \text{for patient consumers,} \end{cases}$$

so that the ex ante expected utility is $\pi_1 U(C_1) + \pi_2 \rho U(C_2)$.

If there are bank runs, they coincide with sunspots that occur with probability $\alpha$.

1. Assuming that the risk-neutral bank manager brings in equity, and the other agents have deposit contracts, compute under what conditions the $G$ allocation is obtained. Interpret this condition in terms of regulation.

2. In what follows, we restrict our attention only to the particular case of risk-neutral depositors, $U(C) = C$. What is the optimal contract? What are the manager's incentives to implement $G$? Do they depend upon $\alpha$? Could we propose a better contract by defining an equity economy?

### 7.8.2 Bank Runs

Consider an economy with a unique good and three dates, with a storage technology that yields a zero net interest and a standard long-run technology that yields $R$ units with certainty at time $t = 2$ but yields only $L$ $(L < 1)$ if prematurely liquidated at time $t = 1$. Both technologies are available to any agent.

A continuum of agents is endowed with one unit at time $t = 0$. Of these agents, a nonrandom proportion $\pi_1$ will prefer to consume at time $t = 1$, and the complementary proportion $\pi_2$ will prefer to consume at time $t = 2$.

The agents' utility function is

$$\begin{cases} \sqrt{C_1} & \text{for impatient consumers,} \\ \rho\sqrt{C_2} & \text{for patient consumers,} \end{cases}$$

so that the ex ante expected utility is $\pi_1\sqrt{C_1} + \pi_2\rho\sqrt{C_2}$.

Assume first that $\rho R > 1$.

1. Compute the first-order condition that fully characterizes the optimal allocation. Compare it with the market allocation that is characterized by $C_1 = 1$ and $C_2 = R$.
2. Consider a banking contract where a depositor's type is private information. Are bank runs possible? If so, for what parameter values?
3. Is the optimal contract implementable within an equity economy, where each agent has a share of a firm that distributes dividends, and a market for ex-dividend shares opens at time $t = 1$, as suggested by Jacklin?

Assume now that $\rho R < 1$.

4. What would be the optimal banking contract? Are bank runs possible? if so, for what parameter values?
5. Is the optimal contract implementable within an equity economy à la Jacklin?

### 7.8.3 Information-Based Bank Runs

This problem is adapted from Postlewaite and Vives (1987). Consider a one-good, three-dates, two-agent economy in which the gross return is $r_1$ $(<1)$ for an investment during the first year $(t = 0$ to $t = 1)$, $r_2$ for an investment during the second year, and $r_3$ for an investment during the third year. Assume $2r_1 - 1 > 0$, and $2r_1r_2 - 1 > 0$. The preferences can be of three types. If an agent is of type 1, her utility is $U(x_1)$; of type 2, $U(x_1 + x_2)$; and of type 3, $U(x_1 + x_2 + x_3)$. The probability that agent 1 is of type $i$ and agent 2 is of type $j$ is $p_{ij}$.

The (exogenous) banking contract allows each agent to withdraw the amount initially deposited without penalty at dates 1 and 2, but interest can be collected only if the agent waits until date 3.

1. Define $a_t^i$ as the strategy that consists in withdrawing everything at time $t$. Write the matrix of payments when both agents initially deposit one unit.

2. Consider the restriction of the game to strategies $a_1^i$ and $a_2^i$. What is the equilibrium if $r_1 > (2r_1 - 1)r_2$, and $1 > r_1 r_2$? Is this an efficient allocation?

3. Returning to the initial matrix, assume that $(2r_1 - 1)r_2 r_3 > 1$. Describe the equilibrium by establishing the optimal strategy for each type. Will there be any bank runs?

### 7.8.4 Banks' Suspension of Convertibility

This problem is adapted from Gorton (1985). Consider a three-dates economy $(t = 0, 1, 2)$ with a unique consumption good that cannot be stored but can be invested. There is a continuum of agents of total measure 1, each having one unit of the good as an initial endowment. Agents have identical risk-neutral preferences, represented by

$$U(C_1, C_2) = C_1 + \frac{1}{1+\rho} C_2,$$

where $C_t$ denotes consumption at time $t$.

The only available technology yields returns $r_1$ at time 1, and $r_2$ at time 2. A signal $s$, belonging to the interval $[\underline{s}, \bar{s}]$, characterizes the distribution of $r_2$. Assume that if $s_1 > s_2$, the distribution of $r_2$ conditionally on $s_1$ first-order dominates the distribution of $r_2$ conditionally on $s_2$.

1. Show that the optimal consumption decision is given by

$$\begin{cases} C_1 = 1 + r_1 \quad \text{and} \quad C_2 = 0 & \text{if } 1 + \rho > E[1 + \tilde{r}_2 | s], \\ C_1 = 0 \quad \text{and} \quad C_2 = 1 + \tilde{r}_2 & \text{if } 1 + \rho < E[1 + \tilde{r}_2 | s], \\ \text{undetermined} & \text{if } 1 + \rho = E[1 + \tilde{r}_2 | s]. \end{cases}$$

2. A mutual fund contract is defined as one in which an investment of $I_0$ gives a right to $I_0(1 + r_1)d_1$ at time 1, and an investment of $I_1$ at time 1 gives a right to $I_1(1 + \tilde{r}_2)(1 - d_1)$ at time 2, where $d_1$ is the fraction that is withdrawn at period 1. Assume agents invest $Q$ in the mutual fund equity, and that the fund is liquidated if and only if it has repurchased all the investor shares. Show that the optimal allocation is obtained.

3. A deposit contract is defined as the right to withdraw amounts $d_1$ and $d_2$ such that $d_1 \leq D(1 + r_D)$ and $d_2 = (D(1 + r_D) - d_1)(1 + r_D)$, where $D$ is the initial deposit and $r_D$ is the promised rate on deposits, $(r_D > \rho)$. When the bank fails to pay the amount

due, its assets are distributed in proportion to the depositors' rights, so that if the bank fails during period 1,

$$d_1 = 1 + r_1, \qquad 1 + r_1 < D(1 + r_D),$$

and if it fails during period 2,

$$d_2 = (1 + r_1 - d_1)(1 + r_2), \qquad d_2 < D(1 + r_D - d_1)(1 + r_D).$$

Equity holders are period 2 residual claimants. The bank will choose to close only if this increases its expected net present value. Speculative bank runs are defined as ones that happen independently of $s$, and fundamental bank runs as ones that arise for low values of $s$. Also, $\delta(\hat{d}_1, s)$ is defined as the expected period 2 return on deposits when the other agents withdraw $\hat{d}_1$.

3a. Show that $\delta(\hat{d}_1, s)$ is increasing (resp. decreasing) in $\hat{d}_1$ if $1 + r_1 > D(1 + r_D)$ (resp. $1 + r_1 < D(1 + r_D)$).

3b. Characterize the different Nash equilibria that obtain depending on the values of $\rho(0, s), \rho(D(1 + r_D), s)$, and $1 + \rho$, and show that for some of these values a speculative bank run obtains, while others result in a fundamental bank run.

3c. Show that this contract does not lead to the optimal allocation.

4. Assume now that $r_2$ is observable by the bank's management, whereas depositors observe only $s$. Show that if the bank's equity holders find it profitable to pay an auditing cost $c$ in order to make $r_2$ publicly observable while suspending convertibility, it is Pareto-superior to do so.

### 7.8.5 Aggregated Liquidity Shocks

This problem, adapted from Hellwig (1994), studies the allocation of interest rate risk in an extension of the Bryant-Diamond-Dybvig model (see section 7.1). There are three possible technologies (all with constant returns to scale):

- A short-term investment at date 0 that yields a return $r_1 = 1$ at date 1 (storage technology).
- A long-term investment at date 0 that yields a return $R > 1$ as of date 2 but can also be liquidated at date 1 for a return $L < 1$.
- A short-term investment at date 1 that yields a random return $\tilde{r}_2$ at date 2. $\tilde{r}_2$ is observed only at date 1. It is assumed that $1 \leq \tilde{r}_2 \leq R/L$.

The consumption profile $(C_1, C_2)$ may now depend on $\tilde{r}_2$ (depositors may bear some of the interest rate risk). This is because when $\tilde{r}_2$ is large, some quantity $x(\tilde{r}_2)$ of the available consumption good can be invested in the short-term technology rather than used for immediate consumption by impatient consumers.

The optimal allocation is obtained by solving

$$\mathscr{P}_3 \begin{cases} \max E[\pi_1 u(C_1(\tilde{r}_2)) + \pi_2 u(C_2(\tilde{r}_2))] \\ \pi_1 C_1(\tilde{r}_2) + x(\tilde{r}_2) = 1 - I, \\ \pi_2 C_2(\tilde{r}_2) = RI + \tilde{r}_2 x(\tilde{r}_2), \\ x(\tilde{r}_2) \geq 0. \end{cases}$$

Let $(C_1^*(r_2, M),\ C_2^*(r_2, M))$ denote the solution of

$$\begin{cases} \max \pi_1 u(C_1) + \pi_2 u(C_2), \\ \pi_1 C_1 + \dfrac{\pi_2 C_2}{r_2} = M. \end{cases}$$

1. Show that $C_1^*$ is increasing in $M$ and decreasing in $r_2$.

2a. Show that the solution of $\mathscr{P}_3$ satisfies when some investment takes place at $t = 1 (x(\tilde{r}_2) > 0)$:

$$\begin{cases} C_1(\tilde{r}_2) = C_1^*\left(\tilde{r}_2, 1 + \left(\dfrac{R}{\tilde{r}_2} - 1\right) I\right), \\ C_2(\tilde{r}_2) = C_2^*\left(\tilde{r}_2, 1 + \left(\dfrac{R}{\tilde{r}_2} - 1\right) I\right). \end{cases} \tag{7.15}$$

2b. Show that the solution of $\mathscr{P}_3$ satisfies when no investment takes place at $t = 1$, and depositors bear no interest rate risk:

$$\begin{cases} C_1(\tilde{r}_2) = \dfrac{1 - I}{\pi_1}, \\ C_2(\tilde{r}_2) = \dfrac{RI}{\pi_2}. \end{cases} \tag{7.16}$$

3. Show that case 2a occurs when $\tilde{r}_2$ is larger than some threshold $r_2^*$. In other words, when the investment opportunities are good enough at $t = 1$ ($\tilde{r}_2 \geq r_2^*$), it is optimal to let depositors bear some risk, even though a complete immunization would be possible, since the allocation defined by (7.16) is always feasible.

### 7.8.6 Charter Value

This problem is based on Calomiris and Kahn (1991). Consider a one-good, three-period economy with risk-neutral agents and zero (normalized) interest rates where banks invest one unit of good in order to obtain random cash flows $\tilde{y}$, $\tilde{y} \in \{\underline{y}, \bar{y}\}$ that are nonobservable, and denote by $p$ the probability of success ($\tilde{y} = \bar{y}$). The bank owner-manager brings in some capital $K$. Depositors invest $D = 1 - K$ and are promised a return $R$ at time $t = 2$ in a time deposit. The bank manager is able to "take the money and run," but at a cost, so that he obtains only $\alpha\tilde{y}$, $0 < \alpha < 1$, and the bank's assets are completely depleted.

1. For which values of $K$ will financial intermediation exist depending on the probability $p$? Compute the level of $R$ for which the depositors break even as a function of $K$. Is the strong form of the Modigliani-Miller theorem (the value of a firm, sum of the market value of debt and equity is constant) satisfied? Why?

2. Assume the bank has a charter value, defined as the net present value of future profits. How will this change the incentive problem?

## 7.9 Solutions

### 7.9.1 Banks Runs and Moral Hazard

1. A contract gives the depositor the right to withdraw $C_1$ at time 1 or wait and consume $\overline{C}_2$ if the project is successful and $\underline{C}_2$ if it fails. If $I$ is invested in the long-run technology, the bank manager will implement the $G$ technology provided that

$$p_G(IG - \pi_2\overline{C}_2) \geq p_B(IB - \pi_2\overline{C}_2).$$

In terms of regulation this implies that payments to the depositors cannot be too generous in case of success. The condition is satisfied if a sufficient amount of capital is brought in by the managers–equity holders.

2. Optimal contract: $C_1 = \underline{C}_2 = 0$, $I = 1$ provided $p_G G > 1$. The manager's incentives are

$$p_G(G - \pi_2\overline{C}_2) \geq p_B(B - \pi_2\overline{C}_2).$$

There are no bank runs, so $\alpha$ is irrelevant. An equity contract could replicate the deposit contract but cannot do better.

### 7.9.2 Bank Runs

1. The first-order condition is

$$C_2 = \rho^2 R^2 C_1,$$

and the efficient allocation is

$$C_1 = \frac{1}{\pi_1 + \rho^2 R\pi_2}, \qquad C_2 = \frac{\rho^2 R^2}{\pi_1 + \rho^2 R\pi_2}.$$

2. Since $C_2 > C_1$, the only type of run we have to consider is the Nash equilibrium, where every agent is better off withdrawing when all other agents are doing the same. This will occur if $C_1 > \pi_1 C_1 + (1 - \pi_1)L$, that is, if

$$\frac{1}{\pi_1 + \rho^2 R \pi_2} > \dot{L}.$$

Thus, for instance, if $\rho^2 R > 1$, and $L = 1$, bank runs will never occur.

3. In a dividend economy with $d$ being the cash dividend and $p$ the price at time $t = 1$ of the ex-dividend share,

$$C_1 = d + p \quad \text{and} \quad C_2 = (d + p)\frac{R(1 - d)}{p},$$

so

$$\frac{C_2}{C_1} = \rho^2 R^2 = \frac{R(1 - d)}{p}.$$

Equality of supply and demand of ex-dividend shares at $t = 1$ implies

$$\pi_1 = \pi_2 \frac{d}{p}.$$

Then

$$\rho^2 R + \frac{\pi_1}{\pi_2} = \frac{1}{p},$$

that is,

$$p = \frac{\pi_2}{\pi_1 + \rho^2 R \pi_2} \quad \text{and} \quad d = \frac{\pi_1}{\pi_1 + \rho^2 R \pi_2}.$$

An equity economy will implement the best.

4. If $\rho R < 1$, the optimal allocation would imply $C_2 < C_1$, so type 2 agents prefer to withdraw and store. The optimal deposit allocation would imply $C_2 = C_1$, so it will not reach the best. Bank runs are always possible because $C_2 = C_1 > 1$ and $L < 1$.

5. An equity economy will not reach the best because the demand for ex-dividend shares by impatient types is zero for

$$\frac{R(1 - d)}{p} < 1.$$

### 7.9.3 Information-Based Bank Runs

1. The matrix of payments is

| | | Agent 1 $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|---|
| Agent 2 | $a_1$ | $(r_1, r_1)$ | $(1, (2r_1 - 1)r_2)$ | $(1, (2r_1 - 1)r_2r_3)$ |
| | $a_2$ | $((2r_1 - 1)r_2, 1)$ | $(r_1r_2, r_1r_2)$ | $(1, (2r_1r_2 - 1)r_3)$ |
| | $a_3$ | $((2r_1 - 1)r_2r_3, 1)$ | $((2r_1r_2 - 1)r_3, 1)$ | $(r_1r_2r_3, r_1r_2r_3)$ |

2. The restriction to $(a_1, a_2)$ shows a game with the "prisoner's dilemma" structure. Strategy $a_1$ dominates strategy $a_2$. If $r_2 > 1$, the allocation is inefficient.

3. The optimal strategy will be

$$\begin{cases} a_1^i & \text{if } i\text{'s type is 1 or 2,} \\ a_3^i & \text{if } i\text{'s type is 3.} \end{cases}$$

Therefore bank runs occur when one of the agents is of type 2.

### 7.9.4 Banks' Suspension of Convertibility

1. For a given realization of $s$, solve

$$\max_{0 \leq C_1 \leq 1+r_1} C_1 + \frac{1}{1+\rho} E[(1 + r_1 - C_1)(1 + \tilde{r}_2)|s],$$

which gives the desired result.

2. Let $d_1$ be the fraction of the mutual fund that is withdrawn at time 1. For a given realization of $s$, solve

$$\max_{0 \leq d_1 \leq 1} (1 - Q)\left\{ d_1(1 + r_1) + \frac{1}{1+\rho} E[(1 + r_1)(1 + \tilde{r}_2)(1 - d_1)] \right\},$$

so

$$d_1 = \begin{cases} 0 & \text{if } 1 + \rho > E[1 + \tilde{r}_2|s], \\ 1 & \text{if } 1 + \rho < E[1 + \tilde{r}_2|s], \\ \text{undetermined} & \text{if } 1 + \rho = E[1 + \tilde{r}_2|s]. \end{cases}$$

The rules on the closing of the mutual fund imply that agents withdrawing $(1 - Q)(1 + r_1)$ also obtain their capital $Q(1 + r_1)$ so that their consumption is optimal. If fees are introduced, the solution is unchanged if these fees are proportional to withdrawals or if they are redistributed to equity holders in the form of dividends.

3a. The equity holders have a call option on the time 2 value of the bank's assets. Therefore, if closing the bank generates a zero profit, which happens when $1 + r_1 < D(1 + r_D)$, the bank will never liquidate its investment, independently of the value of the signal $s$.

Consider the cases for which the bank does not close down during period 1.

Let

$$\delta(\hat{d}_1, s) = E\left[\min\left[\frac{(1 + r_1 - \hat{d}_1)(1 + \tilde{r}_2)}{D(1 + r_D) - \hat{d}_1}, 1 + r_D\right]\middle| s\right].$$

The depositor will choose $d_1$ so as to solve

$$\begin{cases} \max\limits_{d_1} \; d_1 + \dfrac{1}{1+\rho}\delta(\hat{d}_1, s)[D(1 + r_D) - d_1] \\ 0 \leq d_1 \leq D(1 + r_D). \end{cases}$$

Since all agents are identical, a Nash equilibrium obtains for $d_1 = \hat{d}_1$ when there is a unique solution, so it is a symmetrical equilibrium. Examine the symmetrical equilibria:

$$\delta(\hat{d}_1, s) = \frac{1 + r_1 - \hat{d}_1}{D(1 + r_D) - \hat{d}_1}\int_{-1}^{\hat{r}_2(\hat{d}_1)} (1 + r_2)\varphi(r_2)\, dr_2 + (1 + r_D)\int_{\hat{r}_2-(\hat{d}_1)}^{\infty} \varphi(r_2)\, dr_2,$$

where $\varphi(r_2)$ is the generalized density function of $\tilde{r}_2$, and $\hat{r}_2(\hat{d}_1)$ is the value of $r_2$ for which

$$\frac{(1 + r_1 - \hat{d}_1)(1 + r_2)}{D(1 + r_D) - \hat{d}_1} = 1 + r_D.$$

Canceling out the terms in $\hat{r}_2(\hat{d}_1)$ yields

$$\frac{d\delta}{d\hat{d}_1} = \frac{1 + r_1 - D(1 + r_D)}{D(1 + r_D) - \hat{d}_1}\int_{-1}^{\hat{r}_2(\hat{d}_1)} (1 + r_2)\varphi(r_2)\, dr_2.$$

3b. If $1 + r_1 > D(1 + r_D)$, $\delta(D(1 + r_D), s) > \delta(0, s)$, and

(1) $d_1 = \hat{d}_1 = 0$ for $1 + \rho < \delta(0, s)$,

(2) $d_1 = \hat{d}_1 = d_1^*$ for $1 + \rho < \delta(d_1^*, s)$,

(3) $d_1 = (1 + r_D)D$ for $1 + \rho < \delta(D(1 + r_D), s)$,

case (3) corresponds to a fundamental bank run. In case (2) the solution is undetermined, so (infinitely many) asymmetrical equilibria are obtained, provided that

$$\delta\left(\int_0^1 d_1(t)\, d\mu(t), s\right) = 1 + \rho.$$

If $1 + r_1 < D(1 + r_D)$, $\delta(0, s) > \delta(1 + r_1, s)$, and

(4) $1 + \rho > \delta(0, s)$, then $d_1 = \hat{d}_1 = 1 + r_1$, since $1 + \rho > \delta(1 + r_1, s)$;

(5) $\delta(0, s) > 1 + \rho > \delta(1 + r_1, s)$, there are three solutions, $d_1 = \hat{d}_1 = 0$; $\hat{d}_1$ with $\delta(d_1^*, s) = 1 + \rho$; and $d_1 = \hat{d}_1 = 1 + r_1$;

(6) $1 + \rho < \delta(1 + r_1, s)$, then $d_1 = \hat{d}_1 = 0$, since $1 + \rho < \delta(0, s)$.

Thus, in case (5) speculative bank runs occur.

3c. The optimal allocation clearly does not obtain in equilibrium.

4. The bank's management will decide to suspend convertibility only if time 2 profit is greater than the auditing cost. This implies that the suspension of convertibility will take place only when time 2 profits are strictly positive for the $r_2$ that is observed. But this in turn implies that depositors obtain $r_D$ with certainty, so they are better off under the suspension of convertibility.

### 7.9.5 Aggregated Liquidity Shocks

1. $(C_1^*, C_2^*)$ is characterized by the first-order condition $u'(C_1^*) = r_2 u'(C_2^*) = \lambda(r_2, M)$, where $\lambda(r_2, M)$, the Lagrange multiplier associated to the budget constraint, is such that this budget constraint is satisfied. Namely,

$$\pi_1(u' - 1)(\lambda(r_2, M)) + \pi_2(u' - 1)\left(\frac{\lambda(r_2, M)}{r_2}\right) = M.$$

Denoting by $\varphi(\lambda, r_2, M)$ the mapping

$$\varphi(\lambda, r_2, M) = \pi_1(u' - 1)(\lambda) + \pi_2(u' - 1)\left(\frac{\lambda}{r_2}\right) - M,$$

we see that $\lambda(r_2, M)$ is defined implicitly by

$$\varphi(\lambda(r_2, M), r_2, M) = 0.$$

$u'$ being decreasing (since $u$ is strictly concave), we see that $\varphi$ is decreasing in $\lambda$ and $M$, and increasing in $r_2$. Therefore, $\lambda(r_2, M)$ is decreasing in $r_2$ and increasing in $M$, as was to be established.

2a. If the third constraint $(x \geq 0)$ does not bind, $x$ can be eliminated between the first and the second constraint, leading to a unique budget constraint:

$$\pi_1 C_1 + \pi_2 \frac{C_2}{r_2} = 1 + \left(\frac{R}{r_2} - 1\right) I.$$

Thus by definition of $(C_1^*, C_2^*)$ the solution of $\mathscr{P}_3$ satisfies

$$\begin{cases} C_1(r_2) = C_1^*\left(r_2, 1 + \left(\dfrac{R}{r_2} - 1\right)I\right), \\ C_2(r_2) = C_2^*\left(r_2, 1 + \left(\dfrac{R}{r_2} - 1\right)I\right). \end{cases}$$

2b. If $x = 0$, then the first and second constraints give directly

$$\begin{cases} C_1 = \dfrac{1 - I}{\pi_1}, \\ C_2 = \dfrac{RI}{\pi_2}. \end{cases}$$

3. In case 2a, we have $x = 1 - I - \pi_1 C_1(r_2)$. According to part 1, $C_1$ is decreasing in $r_2$. Thus $x(r_2) \geq 0 \Leftrightarrow r_2 \leq r_2^*$, where $r_2^*$ is defined by $x(r_2^*) = 0$.

### 7.9.6 Charter Value

1. The bank manager has incentives to abstain from absconding with the money, when $y$ if:

$$y - R \geq \alpha y, \tag{7.17}$$

we consider two cases.

In case 1 (riskless case), expression (7.17) is satisfied for both $\underline{y}$ and $\bar{y}$. (More precisely, it is satisfied for $\underline{y}$ and this implies it will be satisfied for $\bar{y}$, because $(1 - \alpha)\underline{y} \geq R$ implies $(1 - \alpha)\bar{y} \geq R$. This implies that there is no risk for depositors and thus $R = 1 - K$, which implies

$$(1 - \alpha)\underline{y} \geq 1 - K, \tag{7.18}$$

or equivalently,

$$K \geq 1 - (1 - \alpha)\underline{y}. \tag{7.19}$$

In case 2 (risky case), expression (7.17) is satisfied only for $\underline{y}$, and investors are only repaid with probability $p$, so $R = (1 - K)/p$:

$$(1 - \alpha)p\bar{y} \geq 1 - K > (1 - \alpha)p\underline{y} \tag{7.20}$$

Thus, if $1 - K > (1 - \alpha)p\bar{y}$, only case 1 occurs. If $1 - K < (1 - \alpha)p\bar{y}$, then either $(1 - \alpha)\underline{y} < 1 - K$ and only case 2 occurs, or else $(1 - \alpha)\underline{y} \geq 1 - K$ and the two equilibria for both cases are possible. This corresponds to a self-fulfilling prophesy. If depositors believe the equilibrium is risky, the repayment has to be set accordingly, and the manager has an incentive to abscond in case of low return. But if depositors believe the equilibrium is riskless, the manager has the right incentives and does not take the money and run.

The Modigliani-Miller theorem will hold locally within regions. Still, it is clear that when we go from case 1 to case 2, an expected amount $p(1-\alpha)\underline{y}$ of the bank's assets are destroyed, so the structure of liabilities does affect the value of the firm.

2. The existence of a charter value is a substitute for $K$. It would decrease the value of $K$ required for every type of equilibrium.

## Notes

1. The precise definition of a bank panic varies from one author to another. Kemmerer himself enumerates six major bank panics and fifteen minor ones during the period 1890–1908. Benston and Kaufman (1986), using another definition, found bank panics on three occasions: 1874, 1893, and 1908. Gorton (1985) also found panics in 1884, 1890, and 1896.

2. Miron uses this indicator to evaluate the real (as opposed to financial) effects of bank panics. However, the inverse causality cannot be dismissed; it could be argued that decreases in GNP tend to shrink the value of banks' assets, thus triggering bank panics.

3. Of course, bank runs can develop into bank panics; this is the contagion phenomenon.

4. For simplicity, we do not discount the utility of consumption at date 2.

5. This is true, for instance, if $R \to Ru'(R)$ is decreasing, which corresponds to assuming that the elasticity of substitution between periods is smaller than 1.

6. There is also a mixed-strategy equilibrium that is not considered here.

7. See Kareken (1986) and Mussa (1986).

8. By the law of large numbers, the realized proportion equals the theoretical frequency. This changes for aggregate liquidity risk (see section 7.6).

9. Engineer (1989) has shown that suspension of convertibility may fail to prevent a bank run if the Diamond-Dybvig (1983) model is extended to a framework with four dates and three types of agents.

10. The situation is more complex if the return on banks' assets is uncertain, and if moral hazard can occur (see, for instance, Freeman 1988).

11. This is the key difference with respect to market equilibrium. In a market equilibrium agents are free to choose the level of investment $I$. Here this level is set so as to maximize expected utility.

12. This is only true insofar as the uniformed lender cannot make the entrepreneur and the financial intermediation compete. This is an important assumption in Diamond and Rajan (2001) that is embodied in the timing of offers and counteroffers.

13. The alternative assumption, $V_L > \alpha V_H$, is easier to deal with and allows for equity financing, but the cost of equity financing does not appear. In this case, if $V$ takes the value $V_H$, the banker cannot renegotiate its payment to the equity holders without triggering a bank run. Thus there is no cost of using equity, and the maximum amount $E(V)$ can be raised from the two types of claim holders: depositors and equity holders.

14. Temzelides (1997) studies a repeated version of the Diamond-Dybvig model and models equilibrium selection (between the efficient and the panic equilibria) by an evolutionary process. He shows that the probability of panic decreases with the size of banks, and he studies the possibility of contagion effects.

15. As reported in Benston et al. (1986), this "gambling for resurrection" behavior is frequently observed when a bank faces a crisis. Therefore, although it is unattractive from a theoretical viewpoint because the manager contract is not the optimal one, this assumption is consistent with casual empiricism.

16. See chapter 2.

17. Adao and Temzelides (1995) introduce Bertrand competition between banks in the Diamond-Dybvig model. They show that surprisingly Bertrand equilibria may imply positive profits.

18. Other interesting approaches to the role of the interbank market are provided by Aghion, Bolton, and Dewatripont (1988) and Bhattacharya and Fulghieri (1994).

19. Schwartz (1986) argues that the severe consequences of the Great Depression in the United States could have been considerably limited had the Fed properly conducted lender-of-last-resort operations.

20. See also Smith (1984) for a model of the role of the LLR in the presence of adverse selection.

21. In the case where the Central Bank has only imperfect information about the bank's asset value, there is a trade-off between lending to insolvent banks and refusing to lend to solvent banks.

22. CAMELS is a confidential supervisory rating given to each regulated financial institution (bank) as part of an examination process undertaken by federal and state banking agencies. This rating is based on financial statements of the bank and on-site examination. The components of a bank's condition that are assessed are Capital, Assets, Management, Earnings, Liquidity, and (since 1997) Sensitivity to market risk. The scale is from 1 (strongest) to 5 (weakest). These ratings are not released to the public but only to the top management of the banking company. This is to prevent a bank run on a bank with a bad CAMELS rating.

23. See also the references in these articles.

24. Miron (1986) makes a simple test using a Bernoulli distribution. He estimates that prior to the founding of the Fed the probability of having a panic during a given year was 0.316. This implies that the probability of having no bank panic during the fourteen years 1914–1928 was only 0.005. Miron rejects the hypothesis of no change in the frequency of panics at the 99 percent level of confidence.

## References

Acharya, V., and T. Yorulmazer. 2006. Cash-in-the-market pricing and optimal resolution of bank failures. London Business School. Mimeograph.

Adao, B., and T. Temzelides. 1995. Beliefs, competition and bank runs. Discussion paper, Research Department, Federal Reserve Bank of Philadelphia.

Aghion, P., P. Bolton, and M. Dewatripont. 1988. Interbank lending and contagious bank runs. DELTA, Ecole Normale Supérieure, Paris. Mimeograph.

Aharony, J., and I. Swary. 1983. Contagion effects of bank failures: Evidence from capital markets. *Journal of Business* 56 (3): 305–322.

Allen, F., and D. Gale. 1998. Optimal financial crises. *Journal of Finance* 53 (4): 1245–1284.

———. 2000. Financial contagion. *Journal of Political Economy* 108: 1–33.

Alonso, I. 1991. On informationally based bank runs: The Jacklin and Bhattacharya paper revisited. University of Minnesota, Minneapolis. Mimeograph.

Anderlini, L. 1989. Theoretical modelling of banks and bank runs. In *The economics of missing markets, information and games*, ed. F. Hahn. Oxford: Oxford University Press.

Bagehot, W. 1873. Lombard Street: A description of the money market. London: H. S. King.

Benston, G., R. Eisenbeis, P. Horvitz, E. Kane, and G. Kaufman. 1986. *Perspectives on safe and sound banking: Past, present and future.* Cambridge, Mass.: MIT Press.

Benston, G., and G. Kaufman. 1986. Risks and failures in banking: Overview, history, and evaluation. In *Deregulating financial services: Public policy in flux*, ed. G. Kaufman and R. Kormendia. New York: Ballinger.

Bhattacharya, S., and P. Fulghieri. 1994. Uncertain liquidity and interbank contracting. *Economics Letters* 44: 287–294.

Bhattacharya, S., and D. Gale. 1987. Preference shocks, liquidity and central bank policy. In *New approaches to monetary economics*, ed. W. Barnett and K. Singleton. Cambridge: Cambridge University Press.

Bordo, M. D. 1986. Financial crises, banking crises, stock market crashes and the money supply: Some international evidence, 1887–1933. In *Financial crises and the world banking system*, ed. F. Capie and G. E. Wood. London: Macmillan.

———. 1990. The lender of last resort: Alternative views and historical experience. Federal Reserve Bank of Richmond. *Economic Review* 76 (1): 18–29.

Bryant, J. 1980. A model of reserves, bank runs, and deposit insurance. *Journal of Banking and Finance* 4: 335–344.

———. 1981. Bank collapse and depression. *Journal of Money, Credit and Banking* 13 (4): 454–464.

Calomiris, C., and C. Kahn. 1991. The role of demandable debt in structuring optimal banking arrangements. *The American Economic Review* 81: 497–513

Chari, V. V., and R. Jagannathan. 1988. Banking panics, information and rational expectations equilibrium. *Journal of Finance* 43 (3): 749–761.

De Bandt, O., and P. Hartmann. 2002. Systemic risk in banking: A survey. In *Financial crises, contagion and the lender of last resort*, ed. C. Goodhart and G. Illing. London: Oxford University Press.

Diamond, D. W., and P. H. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91 (3): 401–419.

Diamond, D. W., and R. G. Rajan. 2000. A theory of bank capital. *Journal of Finance* 55 (6): 2431–2465.

———. 2001. Liquidity risk, liquidity creation and financial fragility: A theory of banking. *Journal of Political Economy* 109 (2): 287–327.

———. 2005. Liquidity shortage and banking crises. *Journal of Finance* 60 (2): 615–647.

Eichengreen, B., and R. Portes. 1987. The anatomy of financial crises. In *Threats to international financial stability*, ed. R. Portes and A. Swoboda. Cambridge: Cambridge University Press.

Engineer, M. 1989. Bank runs and the suspension of deposit convertibility. *Journal of Monetary Economics* 24 (3): 443–454.

Fisher, I. 1933. The debt-deflation theory of great depressions. *Econometrica* 1 (4): 337–357.

Freeman, S. 1988. Banking as the provision of liquidity. *Journal of Business* 61 (1): 45–64.

Freixas, X. 1999. Optimal bail out policy, conditionality and constructive ambiguity. Economic and business working paper 400, Universitat Pompeu Fabra, Barcelona.

Freixas, X., and C. Holthausen. 2005. Interbank market integration under asymmetric information. *Review of Financial Studies* 18 (2): 459–490.

Freixas, X., and B. Parigi. 1998. Contagion and efficiency in gross and net interbank payment systems. *Journal of Financial Intermediation* 7 (1): 3–31.

Freixas, X., B. Parigi, and J. C. Rochet. 2000. Systemic risk, interbank relations, and liquidity provision by the central bank. *Journal of Money, Credit and Banking* 32 (2): 611–638.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960.* Princeton, N.J.: Princeton University Press.

Goodfriend, M., and R. King. 1988. Financial deregulation, monetary policy, and central banking. In *Restructuring banking and financial services in America*, ed. W. Haraf and R. M. Kushmeider. Washington: American Enterprise Institute.

Goodhart, C. 1987. Why do banks need a central bank? *Oxford Economic Papers* 39 (1): 75–89.

———. 1995. The central bank and the financial system. Cambridge, Mass.: MIT Press.

Goodhart, C., and H. Huang. 1999. A model of the lender of last resort. Working paper 99-39, International Monetary Fund, Washington, D.C.

Goodhart, C., and D. Shoenmaker. 1995. Should the functions of monetary policy and banking supervision be separated? *Oxford Economic Papers* 47: 539–560.

Gorton, G. 1985. Banks' suspension of convertibility. *Journal of Monetary Economics* 15: 177–193.

Gorton, G., and G. Pennacchi. 1990. Financial intermediaries and liquidity creation. *Journal of Finance* 45 (1): 49–71.

———. 1993. Money market funds and finance companies: Are they the banks of the future? In *Structural change in banking*, ed. M. Klausner and L. White, 173–227. New York: Irwin.

Guttentag, J., and R. Herring. 1987. Emergency liquidity assistance for international banks. In *Threats to international financial stability*, ed. R. Portes and A. Swoboda. Cambridge: Cambridge University Press.

Hart, O., and J. Moore. 1994. A theory of debt based on the inalienability of human capital. *Quarterly Journal of Economics* 109: 841–879.

Hellwig, M. 1994. Liquidity provision, banking, and the allocation of interest rate risk. *European Economic Review* 38 (7): 1363–1389.

Herring, R., and P. Vankudre. 1987. Growth opportunities and risk-taking by financial intermediaries. *Journal of Finance* 42 (3): 583–599.

Holmström, B., and J. Tirole. 1998. Private and public supply of liquidity. *Journal of Political Economy* 106 (1): 1–40.

Holthausen, C., and T. Rønde. 2002. Regulating access to international large-value payment systems. *Review of Financial Studies* 15: 1561–1586.

Humphrey, D. B. 1986. Payments finality and the risk of settlement failure. In *Technology and the regulation of financial markets*, ed. A. Saunders and L. J. White, ch. 8. Lexington, Mass.: Lexington Books.

Humphrey, T. M. 1986. The classical concept of the lender of last resort. In *Essays on inflation*, ed. T. Humphrey. 5th ed. Richmond, Va.: Federal Reserve Bank of Richmond.

Humphrey, T. M., and R. E. Keleher. 1984. The lender of last resort: A historical perspective. *Cato Journal* 4 (1): 275–318.

International interbank market: A descriptive study. 1983. Economic paper 8, Bank for International Settlements.

Jacklin, C. J. 1987. Demand deposits, trading restrictions and risk sharing. In *Contractual arrangements for intertemporal trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

———. 1993. Market rate versus fixed rate demand deposits. *Journal of Monetary Economics* 32: 237–258.

Jacklin, C. J., and S. Bhattacharya. 1988. Distinguishing panics and information-based bank runs: Welfare and policy implications. *Journal of Political Economy* 96 (3): 568–592.

Kareken, J. 1986. Federal bank regulatory policy: A description and some observations. *Journal of Business* 59 (1): 3–48.

Kaufman, G. 1991. Lender of last resort, too large to fail, and deposit-insurance reform. In *The reform of Federal Deposit Insurance: Disciplining the government and protecting the taxpayer*, ed. J. R. Barth and R. D. Brumbaugh Jr., 246–258. New York: Harper Business.

Kemmerer, E. W. 1910. *Seasonal variations in the relative demand for money and capital in the United States.* Washington: Government Printing Office.

Meltzer, A. 1986. Financial failures and financial policies. In *Deregulating financial services: Public policy in flux*, ed. G. Kaufman and R. Kormendia. Cambridge, Mass.: Ballinger.

Miron, J. A. 1986. Financial panics, the seasonality of the nominal interest rate, and the founding of the Fed. *American Economic Review* 76 (1): 125–140.

Morris, S., and H. S. Shin. 2000. Rethinking multiple equilibria in macroeconomic modelling. NBER Macroeconomics Annual. Cambridge, Mass.: MIT Press.

Mussa, M. 1986. Safety and soundness as an objective of regulation of depository institutions: Comments on Kareken. *Journal of Business* 59 (1): 97–117.

Postlewaite, A., and X. Vives. 1987. Bank runs as an equilibrium phenomenon. *Journal of Political Economy* 95 (3): 485–491.

Qi, J. 1998. Deposit liquidity and bank monitoring. *Journal of Financial Intermediation* 7 (2): 198–218.

Repullo, R. 2005. Liquidity, risk-taking, and the lender of last resort. *International Journal of Central Banking* 1: 47–80.

Rochet, J. C., and X. Vives. 2004. Coordination failures and the lender of last resort: Was Bagehot right after all? *Journal of the European Economic Association* 2 (6): 1116–1147.

Saunders, A. 1987. The inter-bank market, contagion effects and international financial crises. In *Threats to international financial stability*, ed. R. Portes and A. Swoboda. Cambridge: Cambridge University Press.

Smith, B. 1984. Private information, deposit interest rates, and the "stability" of the banking system. *Journal of Monetary Economics* 14 (3): 294–317.

Temzelides, T. 1997. Evolution, coordination and banking panics. *Journal of Monetary Economics* 40 (1): 163–183.

Von Thadden, E. L. 1996. Optimal liquidity provision and dynamic incentive compatibility. Cahiers de Recherches Economiques 9604. Université de Lausanne, Switzerland.

Wallace, N. 1988. Another attempt to explain an illiquid banking system: The Diamond-Dybvig model with sequential service taken seriously. Federal Reserve Bank of Minneapolis. *Quarterly Review* 12 (4): 3–16.

———. 1996. Narrow banking meets the Diamond-Dybvig model. Federal Reserve Bank of Minneapolis. *Quarterly Review* 20 (1): 3–13.

# 8 Managing Risks in the Banking Firm

The management of risks can be seen as the major activity of banks as well as other financial intermediaries such as insurance companies. Commercial banks, investment banks, and mutual funds have to control and select the risks inherent in the management of deposits, loans portfolios of securities, and off-balance-sheet contracts.

Since the risks that a bank has to manage are diverse, several classifications have been proposed, some of which are worth mentioning here. Thus economists have put forward the fundamental distinction between *microeconomic risks*, or *idiosyncratic risks*, which can be diversified away through the law of large numbers, and *macroeconomic risks*, or *systematic risks*, which cannot. Unlike property and casualty insurance companies, which essentially deal with microeconomic risks, banks and life insurance companies generally have to deal with both types of risks.

Another fundamental distinction, valid for any type of firm, is between *liquidity risk*, which appears when a firm is not certain to repay its creditors on time, and *solvency risk*, which appears when the total value of a firm's assets falls below the total value of its liabilities. Like any limited liability firm, banks are subject to both types of risks, but the consequences of these risks are much more dramatic for banks than for the other sectors of the economy. This has justified the implementation, in most countries, of complex regulation systems, studied in chapter 9.

The classification that this chapter uses stems from the classification of banking activities offered in chapter 1. The credit activity of banks is affected by *default risks*, or *credit risks*, which occur when a borrower is not able to repay a debt (principal or interest).[1] *Liquidity risks* occur when a bank must make unexpected cash payments. This type of risk essentially comes from the specificity of the demand deposit contract. Unlike the creditors of other kinds of firms, depositors are allowed to demand their money at any time. Consequently, the deposit activity is affected by the risk of an unexpected massive withdrawal by depositors. Similarly, *interest rate risks* are generated by the activity of maturity transformation of short-term deposits into long-term loans. Finally, *market risks* affect the portfolios of marketable assets (and liabilities) held by banks. Credit, liquidity, interest rate, and market risks are examined successively in this chapter.

Confronted with severe banking crises in a large number of countries in the last two decades of the twentieth century, banking authorities have gradually imposed capital requirements that depend on the composition of banks' portfolios. A piecemeal approach has been followed by these authorities. As explained in more detail in chapter 9, they have started to deal with credit risk, then proceeded to interest rate risk, and finally to market risk. Before studying in full detail the economic implications of these regulations, we briefly consider these different capital regulations.

## 8.1 Credit Risk

### 8.1.1 Institutional Context

Defining and measuring credit risk is equivalent to determining how the market evaluates the probability of default by a particular borrower, taking into account all the possibilities of diversification and hedging provided by financial markets. In part, the level of risk depends on the institutional arrangements to which the banks are subject, either through the interbank money market or through specialized institutions created for this purpose. This connection between the institutional framework and the different elements that determine the pricing of credit risk is particularly important in applied work.[2] Since this section is mainly concerned with the theoretical foundations of default risk, this discussion simply lists the points that are relevant to measuring credit risk (see, e.g., Hempel and Simonson (1991) for a more detailed description of the institutional context and its relation to credit risk).

Clearly the riskiness of a loan will be affected by the existence of

- collateral;
- compensating balances;
- endorsement.

But other characteristics of the credit market will also be relevant: Do banks share information on their creditors? How is the bankruptcy process settled? The reader should keep in mind that the (random) return on a loan will depend on all these features.

Notice that in the process of international competition, as well as in a process of market integration like the one undergone by Europe, differences between these institutions or regulations across countries are fundamental, not only because they may represent a barrier to entry but also because they may tend to concentrate some banking activities in countries that provide more efficient institutions and regulations.

### 8.1.2 Evaluating the Cost of Credit Risk

To explain how credit risk affects the competitive pricing of loans, we begin with a simple approach that justifies the use of the risk spread (the difference between the interest rate on a risky loan and the riskless rate for the same maturity) as a measure of the credit risk of an asset. We show how the risk spread is determined by the borrower's probability of default and then examine a more complete approach based on option pricing (Merton 1974).

**A Simple Interpretation of Risk Spread**

Assuming that default risk is diversifiable and that the bank can diversify this risk away through a large population of borrowers, the only thing that matters is the probability of default.[3] Credit scoring methods, the analogue of actuarial techniques used by insurers, allow banks to estimate a priori this probability of default based on the observable characteristics of the loan applicant.

From a financial viewpoint, the value of such a loan (subject to a diversifiable credit risk) is nothing but the expected present value of the borrower's repayments. Leaving aside interest rate risk for the moment, assume that the refinancing rate $r$ is constant, and take $e^{-r}$ as the one-period discount factor.[4] Consider now a risky loan, characterized by a series of promised repayments $(C_1, C_2, \dots, C_n)$ at future dates $(t_1, t_2, \dots, t_n)$. Assume that if the firm defaults, the bank receives nothing (zero recovery rate). The expected cost of default risk for this loan can be measured by the difference between

$$P_0 = \sum_{k=1}^{n} C_k e^{-rt_k}, \tag{8.1}$$

the value of the loan if there were no default risk, and

$$P = \sum_{k=1}^{n} C_k p_k e^{-rt_k}, \tag{8.2}$$

the value of the risky loan, where $p_k$ denotes the probability that the $k$th repayment will not be defaulted, assuming that there are no partial repayments.

In practice, however, the most commonly used instrument for evaluating the cost of default risk is the difference (spread) between the yield to maturity $R$ of the risky loan and the refinancing rate $r$. $R$ is defined implicitly by the equation

$$P = \sum_{k=1}^{n} C_k e^{-Rt_k}, \tag{8.3}$$

and this determines the value of the spread $s = R - r$.

Bierman and Hass (1975) and Yawitz (1977) have proved the following simple result.

**Result 8.1** If the firm's default follows a Poisson process of intensity $\lambda$, the spread $s$ is independent of the characteristics of the loan. It is equal to the intensity of the Poisson process:

$$s = \lambda. \tag{8.4}$$

***Proof*** In a Poisson process of intensity $\lambda$, the probability of survival at date $t_k$ is by definition

$$p_k = e^{-\lambda t_k}. \tag{8.5}$$

Combining (8.2), (8.3), and (8.5) yields

$$\sum_{k=1}^{n} C_k e^{-Rt_k} = \sum_{k=1}^{n} C_k e^{-\lambda t_k} e^{-rt_k}.$$

This equation in $R$ has a unique solution, which turns out to be independent of $C_1, \ldots, C_n$:

$$R = r + \lambda.$$

Therefore,

$$s = R - r = \lambda. \qquad \blacksquare$$

Consequently, if one considers corporate debt of a certain type, the spread can be considered (in a first approximation) as the instantaneous probability of failure $\lambda$ that the market assesses implicitly to the particular class of borrowers under consideration. For instance, a spread of fifty basis points roughly indicates a failure probability of $1 - e^{-\lambda} \sim \lambda = 0.5$ percent per year.

**The Option Approach to Pricing Default Risk**

The simple approach just explained relies on three assumptions that are not very satisfactory: (1) the instantaneous probability of failure is constant and exogenous; (2) credit risk is completely diversifiable; and (3) in case of failure, the residual value of the firm (or the recovery rate) is zero. Consider now what happens when these assumptions are relaxed.

When credit risk is not completely diversifiable, a risk premium must be introduced, and the analysis becomes more involved. However, financial markets provide insurance possibilities for banks. Therefore, the risk premium quoted by banks must be in line with the ones prevailing in the securities market. This will allow using the model for pricing risky debts, proposed by Merton (1974).

Consider a firm that plans to borrow a certain amount $D_0$ at date $t = 0$ and repay $D$ at date $t = T$. The yield to maturity $r_L$ is defined by

$$D = D_0 e^{r_L T}. \tag{8.6}$$

Let $V(t)$ denote the value at date $t$ of the firm's total assets, assumed to be marketable at no cost, and assume that the firm has no further debt outstanding. Two things can happen at date $T$:

- If $D \leq V(T)$, the firm is solvent, and the bank gets $D$ as promised.
- If $D > V(T)$, the firm is bankrupt, its assets are liquidated, and the bank gets only $V(T)$.

The terminal payoff to the bank is thus

$$\min(D, V(T)), \tag{8.7}$$

and the market value of the firm's equity at date $T$ is

$$\max(0, V(T) - D). \tag{8.8}$$

This last formula is exactly the payoff of a call option[5] on the firm's assets with a strike price equal to $D$. Consequently, from a purely financial viewpoint, granting a risky loan to a limited liability firm is similar to buying the firm's assets and selling a call option to its stockholders. Of course, this approach is somewhat simplistic because it neglects intermediate payments and liquidation costs. Also, in most cases, several loans of different maturities and seniorities coexist. But this funda4-mental insight will allow explicit evaluation of the cost of credit risk in some simple cases.

The only further restriction to be imposed is an assumption on the probability distribution of $V(t)$. Following Merton (1974), assume that $V(t)$ follows a geometric random walk, which is equivalent to saying that instantaneous returns on $V$ are Gaussian, independent, and identically distributed:

$$\frac{dV}{V} = \mu\, dt + \sigma\, dZ,$$

where $\mu$, $\sigma$ are constant ($\sigma > 0$), and $Z$ is a standard Wiener process.[6] Under this assumption, the market value $C$ of a call option on $V$ can be computed by the Black-Scholes formula (1973).[7] Since this discussion focuses on the (market) value $D_0$ of the loan, directly compute

$$D_0 = V - C,$$

and the following is obtained:

$$D_0 = VN(h_1) + De^{-rT}N(h_2), \tag{8.9}$$

where

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \left( exp - \frac{1}{2} t^2 \right) dt$$

is the cumulative of the standard Gaussian distribution, and

$$\begin{cases} h_1 = \dfrac{1}{\sigma\sqrt{T}} \log \dfrac{De^{-rT}}{V} - \dfrac{1}{2}\sigma\sqrt{T}, \\ h_2 = -\dfrac{1}{\sigma\sqrt{T}} \log \dfrac{De^{-rT}}{V} - \dfrac{1}{2}\sigma\sqrt{T}. \end{cases}$$

Again, the interest spread $s$ is defined as the difference between the yield to maturity $r_L$ of the risky loan and the riskless rate $r$. Using formula (8.6) yields

$$s = r_L - r = -\frac{1}{T} \log \frac{D_0}{De^{-rT}}.$$

Now, (8.9) yields

$$s = -\frac{1}{T} \log \left[ \frac{N(h_1)}{d} + N(h_2) \right],$$

where

$$d = \frac{De^{-rT}}{V}$$

is the debt-to-asset ratio.[8] By a comparison with result 8.1, one can show that $N(h_2)$ can be interpreted as the (risk-adjusted) probability of survival at $T$. In the simpler model, the probability of survival was $p_T = e^{-\lambda T}$, and the spread was $s = \lambda = -1/T \log p_T$.

Here the model is richer because

- the probability of failure is not exogenous; it depends in particular on the indebtedness of the firm;
- the market pricing of risk is taken into account;
- the recovery rate on the debt is not zero.

Merton then studies the influence of parameters $d$, $\sigma$, and $T$ on $s$, obtaining the following properties.

**Result 8.2**

1. The interest spread $s$ increases with the debt-to-asset ratio $d$. Thus the more indebted firms will pay higher interest rates.

2. The interest spread $s$ increases with the volatility $\sigma$ of the firm's assets. Thus firms having riskier activities will pay higher interest rates.

3. The global risk premium $sT$ increases with the maturity of the loan.[9] Thus longer loans will be more costly.

All these results, proved in problem 8.6.2, confirm casual empiricism. Note, however, that $s$ is not necessarily increasing in $T$, although the global risk premium $sT$ is.

Several restrictive assumptions limit the validity of the Merton model: deterministic refinancing rate $r$, no interest payments, no other outstanding loans. When stochastic interest rates are introduced (Décamps 1996), parts (1) and (3) of result 8.2 remain valid but not (2). When there is a negative correlation between $r$ and the value $V$ of the firm's asset, $s$ is minimized for a positive value of $\sigma$. Intuitively, the firm's assets provide some insurance against fluctuations of $r$. More generally, Décamps shows that when interest rates are stochastic, the volatility $\sigma$ ceases to be a good measure of risk for the loan. When interest payments are introduced, compound option techniques have to be used (Geske 1977), and no explicit formula is available. Finally, when several loans are simultaneously outstanding, seniority rules are fundamental. This topic has been the subject of interesting theoretical literature (e.g., Bizer and DeMarzo 1992), but the implications on the pricing of risky loans have not yet been explored.

### 8.1.3 Regulatory Response to Credit Risk

The capital requirements imposed by regulators for dealing with credit risk all derive from the capital regulations (Basel Accord) designed in 1988 by the Basel Committee on Banking Supervision. These regulations were designed for internationally active banks but later extended to all banks (with some adaptations) by national regulators. In the United States, for example, they were implemented in 1990 and then incorporated into the Federal Deposit Insurance Corporation Improvement Act. In the European Union they gave rise to the Capital Adequacy Requirement, implemented in January 1993. These regulations require that the capital of banks be equal to at least 8 percent of a weighted sum of the volumes of risky assets held by the banks (plus an additional term related to off-balance-sheet activities, like derivatives).[10] The weights are supposed to reflect different credit risks associated with different categories of assets. This weight system was subsequently revised (Basel II). The analytical modeling of this issue sheds some light on the conditions of the Basel I and Basel II agreements.

Starting from scratch, the regulator's objective function can be approached by assuming the regulator sets a probability of solvency for banks equal to $q$ (e.g., $q$ could be set at 99.9 percent). This is a simplification; it disregards the extent of the losses that occur conditional on default.

A bank's default occurs whenever the value of its losses is larger than its capital. If $x_i$ is the exposure to obligor $i$, and $U_i$ is the random loss per dollar exposure ($U_i = 0$ in case of survival), then

$$\sum_{i=1}^{n} x_i \tilde{U}_i = \tilde{L}_n$$

is the total portfolio loss for a portfolio of $n$ different assets (obligors). For a given capital $K$, bankruptcy occurs when $L_n > K$. This is equivalent to setting a loss ratio

$$\tilde{\ell}_n = \frac{\tilde{L}_n}{\sum x_i}$$

larger than

$$k = \frac{K}{\sum x_i},$$

where $k$ is the capital ratio.

Prudential regulation will impose a level of capital $K$ such that

$$\text{Proba}(\tilde{\ell}_n \geq k) = 1 - q. \tag{8.10}$$

A preliminary remark is order. In a perfectly diversified portfolio with uncorrelated credit risk, the law of large numbers implies that the required capital level is zero. Consequently, the issue of capital requirements is concerned with correlated risks and imperfect diversification. Correlated risks are a natural consequence of the existence of common (e.g., macroeconomic) factors that drive the random losses $U_i$. Imperfect diversification is the natural result of a bank's strategy because a bank may specialize in a particular industry.

Capital regulation is usually designed as a rule based on a vector of coefficients $\alpha_i$, one for each type of risk, so that

$$K \geq \sum_{i=1}^{n} \alpha_i x_i. \tag{8.11}$$

Gordy (2003) examines this issue rigorously by studying under what conditions a minimum capital requirement like (8.11) can implement a prudential regulation based on a probability of default, as in (8.10). Having weights $\alpha_i$ implies that each exposure contribution to total risk is portfolio-invariant.

Denote by $f$ the realization of the common factors $F$. Gordy shows that when the portfolio is sufficiently diversified, $L_n - E[L_n|f]$ converges to zero almost surely. When this is the case, the probability of bankruptcy can be analyzed as the probability of $E(L_n|f) > k$.

Define $E[U_i|f] = \mu_i(x)$; then

$$E[\ell_n|f] = E\left[\frac{1}{\sum x_i}\sum_{i=1}^{n} x_i\tilde{U}_i|f\right] = \sum_{i=1}^{n}\frac{x_i}{\sum x_i}\mu_i(f).$$

If there is a single common factor, it is possible to obtain a critical value $\hat{f}$ for which (8.10) obtains. This $\hat{f}(q)$ is the value of the common factor that triggers bankruptcy:

$$\text{Proba}\left(\sum_{i=1}^{n}\frac{x_i}{\sum x_i}\mu_i(\hat{f}) \geq k\right) = 1 - q.$$

It is therefore possible to obtain portfolio-invariant weights $\alpha_i$ by setting $\alpha_i = \mu_i(\hat{f})$.

Consequently, the regulator has to set the coefficients (e.g., as in the standardized approach of Basel II) equal to the expected loss per dollar exposure conditional on $\hat{f}$. So, if the target probability $q$ is 99.9 percent, the function $\mu_i(f)$ and the distribution of portfolio will determine the value, for instance, the rate of growth, such that the probability of a shortfall of capital is 0.1 percent. This in turn determines the coefficients $\alpha_i$.

Gordy (2003) emphasizes that the unique factor is not just a sufficient condition but also a necessary one. If two factors $f_1$ and $f_2$ drive the distribution of losses, then a bank with a higher proportion of $f_1$ risks and a bank with a higher proportion of an $f_2$ risk should have different $\alpha_i(f_1, f_2)$ weights in order to reach the same probability of solvency $q$.

## 8.2 Liquidity Risk

At the individual level, liquidity management is not fundamentally different for banks than for other firms. It can even be seen as a particular case of the general problem of managing inventories of any sort. However, for reasons that were discussed in chapter 7, the situation is different at the macroeconomic level because the liquidity problems of a single bank can propagate very quickly, affect other banks (externality), and give rise to systemic risk. This has justified the creation of three mechanisms designed to limit the possible extension of these liquidity problems: the lender of last resort, deposit insurance, and reserve requirements. These mechanisms are studied in detail in chapter 9.

This chapter focuses on the microeconomic level. The management of reserves is studied in section 8.2.1. Section 8.2.2 shows how the introduction of liquidity risk modifies the conclusions of the Monti-Klein model. Section 8.2.3 presents another paradigm for banks' behavior based on inventory management. Following Ho and Saunders (1981), a bank is assimilated to a security dealer, who determines an interest margin (bid-ask spread) between loans and deposits as a function of inventory risks.

### 8.2.1 Reserve Management

Consider the problem of a bank that wants to determine the quantity $R$ of liquidities (reserves) to be held, out of a total amount $D$ of deposits. The remaining $(D - R)$ is assumed to be invested in riskless (but illiquid) loans. Using for simplicity a static framework, we suppose that the net amount of withdrawals at the end of the period is a random variable $\tilde{x}$. If the realization $x$ of $\tilde{x}$ is greater than $R$, the bank has a liquidity shortage and has to pay a penalty $r_p(x - R)$, proportional to the shortage. A more reasonable assumption would be that the bank can borrow from the Central Bank or possibly from other banks at rate $r_p + r_D$ (where $p$ stands for penalty, and $r_D$ is the rate on deposits), but this would imply more than one period, so this example adopts the first approach. The rate $r_p$ is, of course, higher than the rate of return $r_L$ on loans, which itself is higher than the interest rate $r$ on reserves. Suppose that deposits are costless for the bank (no interest paid, no management cost) and that the bank is risk-neutral. The bank's expected profit is

$$\Pi(R) = r_L(D - R) + rR - r_p E[\max(0, \tilde{x} - R)] - r_D D.$$

The last term in this formula (the expected cost of liquidity shortages) is a convex function of $R$, which is differentiable under the assumption that the random variable $\tilde{x}$ has a continuous density $f(x)$. Let $C(R)$ denote this cost:

$$C(R) = r_p \int_R^D (x - R) f(x)\, dx,$$

$$C'(R) = -r_p \int_R^D f(x)\, dx = -r_p \Pr[\tilde{x} \geq R],$$

$$C''(R) = r_p f(R) \geq 0.$$

This implies that $\Pi(R)$ is a concave differentiable function of R. It is maximum when

$$\Pi'(R) = -(r_L - r) + r_p \Pr[\tilde{x} \geq R] = 0.$$

Thus the optimal amount of reserves $R^*$ is determined by the relation

$$\Pr[\tilde{x} \geq R^*] = \frac{r_L - r}{r_p}. \tag{8.12}$$

**Result 8.3** The optimal amount of reserves is the amount for which the marginal opportunity cost of holding reserves equals the expected cost of liquidity shortage. Alternatively, the optimal probability of liquidity shortage is just equal to the ratio of the liquidity premium $(r_L - r)$ to the penalty interest rate $r_p$.

For instance, if $r_L - r$ equals 3 percent, and $r_p$ is as large as 15 percent, the probability of liquidity shortage is 20 percent, which seems relatively important. Thus, formula (8.10) seems to overestimate a little the actual probability of liquidity shortage. Several modifications of this basic model have been suggested to increase its predictive power. For instance, if the cost of refinancing $r_p$ increases with the amount borrowed by the bank, or if the bank is risk-averse, the optimal probability of liquidity shortage will clearly decrease. Another improvement is to introduce adjustment costs and information acquisition on depositors' behavior (Baltensperger and Milde 1976; Stanhouse 1986). To increase the realism of the model, one can also introduce different kinds of deposits, and associate compulsory reserves with different coefficients, as well as penalty systems when these requirements are not met. The analysis becomes more complex but is not substantially modified.

The next section shows how introducing reserves management into the Monti-Klein model provides a solution to the asset-liability separation puzzle mentioned in chapter 3. With uncertainty and liquidity requirements, asset and liability decisions become interdependent.

### 8.2.2 Introducing Liquidity Risk into the Monti-Klein Model

One of the conclusions of the Monti-Klein model (see chapter 3) was not entirely satisfactory: when there exists an infinitely elastic source of funds (money market), the optimal policy of a (monopolistic) bank will be characterized by a separation between the pricing of assets (loans) and liabilities (deposits). This seems to contradict the stylized facts of (modern) bankers' behavior; bankers insist on the necessity of global asset-liability management. Moreover, this separation result would imply that any regulation on deposits has no effect on the credit market.

Prisman, Slovin, and Sushka (1986) show how introducing liquidity risk into the Monti-Klein model may alter this result. The simplest way to bring in liquidity risk is to introduce some randomness into the volume of funds collected or distributed by the bank. Thus the demand for loans can be stochastic, as Prisman, Slovin, and Sushka assume, or it can be the volume of deposits that is subject to random shocks, as assumed here. The (monopolistic) bank is assumed to choose the rates $r_L$ and $r_D$ of loans and deposits, taking into account the (downward-sloping) demand function for loans,

$$L = L(r_L),$$

and the (upward-sloping) supply function of deposits,

$$D = D(r_D),$$

from which a random amount $\tilde{x}$ of withdrawals will be subtracted at the end of the period. Assuming no other source of funds is available to the bank, the amount of reserves is simply

$$R = D(r_D) - L(r_L). \tag{8.13}$$

As before, the reserves are assumed to yield a return $r$, but in addition the bank must pay a proportional penalty $r_p$ in case of liquidity shortage at the end of the period. The expected profit of the bank is thus

$$\Pi = r_L L(r_L) - r_D D(r_D) + rR - r_p E[\max(0, \tilde{x} - R)],$$

or, using (8.13),

$$\Pi = (r_L - r)L(r_L) + (r - r_D)D(r_D) - r_p E[\max(0, \tilde{x} - D(r_D) + L(r_L))].$$

Make the usual assumptions on $L$ and $D$ to ensure that $\Pi$ is quasi-concave in $r_L$ and $r_D$: $DD'' - 2D'^2 < 0$ and $LL'' - 2L'^2 < 0$. Under these assumptions, the maximum is characterized by the first-order conditions

$$\begin{cases} \dfrac{\partial \Pi}{\partial r_L} = (r_L - r)L'(r_L) + L(r_L) - r_p \Pr[\tilde{x} \geq R]L'(r_L) = 0, \\ \dfrac{\partial \Pi}{\partial r_D} = (r - r_D)D'(r_D) - D(r_D) + r_p \Pr[\tilde{x} \geq R]D'(r_D) = 0. \end{cases}$$

Introducing the elasticities of the demand for loans and the supply of deposits,

$$\varepsilon_L = -\frac{r_L L'(r_L)}{L(r_L)}, \qquad \varepsilon_D = \frac{r_D D'(r_D)}{D(r_D)},$$

the optimum value of $r_L$ and $r_D$ is

$$r_L^* = \frac{r + r_p \Pr[\tilde{x} \geq R]}{1 - (1/\varepsilon_L)}, \tag{8.14}$$

$$r_D^* = \frac{r + r_p \Pr[\tilde{x} \geq R]}{1 - (1/\varepsilon_D)}. \tag{8.15}$$

Thus the only difference between these formulas and those obtained in the Monti-Klein model is that in this example the cost of the bank's resource is higher than $r$ because it now includes the expected cost of a liquidity shortage. Since the probability of such a shortage depends on $R$, the difference between $D$ and $L$, this introduces the desired dependence between assets and liabilities.

Prisman, Slovin, and Sushka perform an interesting comparative static analysis, the results of which are as follows.

**Result 8.4**

1. If the penalty rate $r_p$ increases, the rates $r_L^*$ and $r_D^*$ also increase. Consequently the volume of credit $L$ decreases, and the volume of deposit $D$ increases.

2. If the variance of $\tilde{x}$ increases (withdrawals become more uncertain), the impact on $L$ depends on the sign of $R$. In the most plausible case $(R > 0)$ this impact is negative; the volume of credit decreases.

These results are proved in problem 8.5.1.

### 8.2.3 The Bank as a Market Maker

Securities traders, such as brokers and dealers in the London Stock Exchange, specialists in the New York Stock Exchange, or market makers in the Paris Bourse, play a fundamental role in the provision of liquidity in modern financial markets. Financial economists, such as Ho and Stoll (1980) studied the determination of the bid-ask prices as a function of the characteristics of a security as well as the inventory policy of the trader. Ho and Saunders (1981) had the interesting idea of adapting this modeling to banking activity, thus providing a new paradigm for banking behavior. Indeed, like the market maker, a bank can be seen as providing liquidity to the market. Like the market maker, it will hold *illiquid* assets and therefore consider the risk of an unbalanced portfolio with extreme positions either long (because it has granted more loans than desired) or short (because it has taken too many deposits). It is worth emphasizing that this approach explains the illiquidity of banks' assets and liabilities and therefore views a bank as different from a mutual fund.

In the Ho-Saunders approach, a bank is considered an intermediary (market maker) on the market for funds, which sets a deposit rate $r_D$ and a loan rate $r_L$ (the equivalent of ask and bid prices) as a function of its inventory level and of the volatility of interest rates.

To maintain the analogy with the market maker, assume there is no credit risk and no difference in maturities between deposits and loans. Suppose that the bank is confronted with stochastic arrivals of depositors and borrowers, modeled by Poisson processes of respective intensities $\lambda_D$ and $\lambda_L$ (as usual, $L$ stands for loans and $D$ for deposits). For simplicity, loans and deposits have the same size $Q$ (standardization) and the same duration (no transformation). So the only thing that matters for the bank is the difference $(L-D)$, that is, its net inventory $I$ resulting from its commercial activity. The bank also has a (fixed) portfolio $\gamma$ of marketable assets and a money market (positive or negative) position $M$, both of which are inherited from the past and result from the need to fund loans or to invest excess liquidities. The total wealth of the bank at the end of the period is

$$\tilde{W} = \gamma(1+\tilde{r}_\gamma) + M(1+r) + I(1+\tilde{r}_I), \tag{8.16}$$

where $\tilde{r}_\gamma$ (resp. $\tilde{r}_I$) is the random return on the bank portfolio (resp. on the credit activity), and $r$ is the (deterministic) money market return. The objective of the bank is assumed to be of the mean-variance type:[11]

$$U = E(\tilde{W}) - \frac{1}{2}\rho \operatorname{var}(\tilde{W}), \tag{8.17}$$

where $\rho$ is a risk aversion coefficient. Using (8.16) and (8.17), compute $U$ as a function of $I$ and $M$:

$$U = U(I, M) = \gamma(1+r_\gamma) + M(1+r) + I(1+r_I) - \frac{1}{2}\rho[\sigma_\gamma^2\gamma^2 + 2\sigma_{\gamma I}\gamma I + \sigma_I^2 I^2], \tag{8.18}$$

where

$$r_\gamma = E(\tilde{r}_\gamma), \qquad r_I = E(\tilde{r}_I), \qquad \sigma_\gamma^2 = \operatorname{var}(\tilde{r}_\gamma),$$

$$\sigma_{\gamma I} = \operatorname{cov}(\tilde{r}_\gamma, \tilde{r}_I), \qquad \sigma_I^2 = \operatorname{var}(\tilde{r}_I).$$

Consider now the increase in the bank's utility consecutive to the market-making activity. The mechanism is as follows. The bank sets margins $a$ and $b$ for deposits and loans, which means that the bank sells securities (attracts deposits) at a bid price $Q(1+a)$ and buys them (grants loans) at an ask price $Q(1-b)$. This means that by paying $Q(1+a)$, the depositor will obtain at the end of the period $Q(1+\tilde{r}_I)$. Therefore, the rate of return for depositors will be

$$\left(\frac{1+\tilde{r}_I}{1+a} - 1\right).$$

Similarly, the rate paid by borrowers is

$$\left(\frac{1+r_{\tilde{I}}}{1-b} - 1\right).$$

In particular, if $\tilde{r}_I$ were deterministic, these rates would be, respectively,

$$\begin{cases} r_D = \dfrac{r_I - a}{1 + a} & \text{for depositors,} \qquad (8.19) \\ r_L = \dfrac{r_I + b}{1 - b} & \text{for borrowers.} \qquad (8.20) \end{cases}$$

When attracting an additional deposit, the bank obtains an increase of utility equal to

$$\begin{aligned}(\Delta U|\text{deposit}) &= U(I - Q, M + Q(1 + a)) - U(I, M) \\ &= Q\{(1 + a)(1 + r) - (1 + r_I)\} \\ &\quad - \frac{1}{2}\rho\{\sigma_I^2(Q^2 - 2QI) - 2\sigma_{\gamma I}\gamma Q\}. \end{aligned} \tag{8.21}$$

Similarly, when granting an additional loan, the bank gets

$$\begin{aligned}(\Delta U|\text{loan}) &= U(I + Q, M - Q(1 - b)) - U(I, M) \\ &= Q\{(1 + r_I) - (1 - b)(1 + r)\} \\ &\quad - \frac{1}{2}\rho\{\sigma_I^2(Q^2 + 2QI) + 2\sigma_{\gamma I}\gamma Q\}. \end{aligned} \tag{8.22}$$

When setting margins $a$ and $b$, the (monopolistic) bank takes into account not only the direct effect on the quantities but also the effect on supply and demand. More specifically, depositors and borrowers are assumed to arrive randomly, according to Poisson processes, the intensities $\lambda_D$ and $\lambda_L$ of which are decreasing functions, respectively, of $a$ and $b$. Ho and Saunders adopt a linear symmetric specification:

$$\lambda_D = \alpha - \beta a, \qquad \lambda_L = \alpha - \beta b. \tag{8.23}$$

The optimal margins $a$ and $b$ are the ones that maximize the (expected) increase in utility:

$$\Delta U = \lambda_D(\Delta U|\text{deposit}) + \lambda_L(\Delta U|\text{loan}).$$

The first-order conditions give

$$\begin{cases} \dfrac{d\lambda_D}{da}(\Delta U|\text{deposit}) + \lambda_D Q(1 + r) = 0, & (8.24) \\ \dfrac{d\lambda_L}{db}(\Delta U|\text{loan}) + \lambda_L Q(1 + r) = 0. & (8.25) \end{cases}$$

Using (8.23) and adding (8.24) and (8.25) yields

$$0 = -\beta[(\Delta U|\text{deposit}) + (\Delta U|\text{loan})] + (\lambda_L + \lambda_D)Q(1 + r),$$

and using $\lambda_D + \lambda_L = 2\alpha - \beta(a+b)$ yields

$$(\Delta U|\text{deposit}) + (\Delta U|\text{loan}) = Q(1+r)\left\{2\frac{\alpha}{\beta} - s\right\},$$

where $s = a + b$ is the bid-ask spread (total margin). The optimal values of $a$ and $b$ are complicated expressions, involving in particular $\gamma$ and $I$. But the optimal spread $s$ has a simple expression, independent of $\gamma$ and $I$. Replacing $(\Delta U|\text{deposit})$ and $(\Delta U|\text{loan})$ by their values given by equations 8.21 and 8.22 yields

$$Q(1+r)s - \rho\sigma_I^2 Q^2 = Q(1+r)\left\{2\frac{\alpha}{\beta} - s\right\},$$

or finally,

$$s = \frac{\alpha}{\beta} + \frac{1}{2}\frac{\rho\sigma_I^2 Q}{(1+r)}. \tag{8.26}$$

**Result 8.5** The total margin between loans and deposits is the sum of two terms:

- $\alpha/\beta$ is the risk-neutral spread that would be chosen by a risk-neutral monopoly. It depends on the elasticities of supply and demand.
- The other term is a risk premium, proportional to the risk aversion coefficient $\rho$ to the variance of the return on the credit activity $\sigma_I^2$ (which is itself related to the volatility of interest rates) and to the size of transaction $Q$.
- The volume of inventories $I$ does not affect $s$ (but it affects $a$ and $b$).

## 8.3 Interest Rate Risk

Interest rate risk can be defined as the risk that fluctuations in interest rates adversely affect the market value of a bank's assets and liabilities (or its interest income). It is a direct consequence of banks' traditional activity of maturity transformation (see chapters 1 and 2) and has become a crucial preoccupation of bank managers, shareholders, and supervisors since the 1980s, when interest rate volatility started to increase. The dramatic effects of an interest rate hike were highlighted, for example, by the crisis of the U.S. savings and loan industry in the 1980s. The core activity of this industry was to finance long-term, fixed-rate mortgages with short-term deposits. As long as long-term interest rates were above short-term interest rates (in the 1970s rates on deposits were even capped at a very low level under Regulation Q), this transformation activity was profitable. But in the early 1980s interest rates on deposits rose significantly because of two factors: fierce competition for deposits between S&Ls, banks, and money market funds; and inversion of the yields curve (short-term rates rose above long-term rates). Since the S&Ls could not revise the mortgage rates on their portfolio of loans, they could not avoid making huge losses. With an inverted yield curve, the transformation activity had become structurally money-losing. Fortunately, such an inversion is not a frequent event, so that maturity transformation is on average a profitable activity. However, the U.S. savings and loan industry was not prepared for this event, which triggered the big S&L crisis of the late 1980s. This convinced the banking industry that something had to be done to seriously manage interest rate risk. This section gives a short introduction to the management of interest rate risk by banks while keeping the general focus of this book, namely, microeconomic and strategic aspects. For a more detailed (and more management-oriented) presentation, the reader is referred to Fabozzi and Konishi (1991).

### 8.3.1 The Term Structure of Interest Rates

Denote (for all dates $t$ and $t+\tau$) by $B(t, t+\tau)$ the price at date $t$ of a default-free bond (zero-coupon bond) paying 1 at date $(t+\tau)$. By definition, the term structure of interest rates (or yield curve) is defined at each date $t$ as the curve $\tau \to Y(t, t+\tau)$, where $Y(t, t+\tau)$ is the yield to maturity of the bond, defined by[12]

$$B(t, t+\tau) = \exp - \tau Y(t, t+\tau). \tag{8.27}$$

The function $\tau \to B(t, t+\tau)$ is sometimes called the actualization function. Knowledge of this function (or equivalently, of the yield curve $\tau \to Y(t, t+\tau)$) is very valuable for investors because it allows computing immediately the (theoretical) market value of any risk-free, fixed-income security. Indeed, such a security is characterized by a sequence $(C_1, \ldots, C_n)$ of promised future payments, at respective dates $(t+\tau_1, \ldots, t+\tau_n)$. In the absence of arbitrage opportunities, the market value of such a security is necessarily equal to

$$V(t) = \sum_{i=1}^{n} C_i B(t, t+\tau_i) = \sum_{i=1}^{n} C_i \exp - \tau_i Y(t, t+\tau_i),$$

the sum of the actualizations (using the actualization rates given by the yield curve) of all future payments.

For simplicity, we restrict the discussion to default-free, fixed-income securities. As mentioned in section 8.1, additional risk premia would have to be incorporated if we dealt with defaultable bonds. Note also that in practice there are not enough zero-coupon bonds traded on the market to be able to observe directly $B(t, t+\tau)$ for all maturities. Therefore, the yield curve is typically constructed by a combination of interpolation and statistical methods that use the market prices of all fixed-income securities, which often involve intermediate coupons.

Thus, the term structure of interest rates reveals, at each date $t$, the prices that the market implicitly assigns to a transfer of funds for any duration $\tau$. As for all market prices, it results from the equilibrium of supply and demand in the markets for funds of different maturities. In particular, long-term rates reflect the expectations of the market about the future financing needs of the government (issuance of Treasury bonds), whereas short-term interest rates are strongly influenced by the monetary policy of the Central Bank. An early contribution of Modigliani and Sutch (1966) put forward the idea that debt markets tend to be segmented and that investors and borrowers have "preferred habitats" (like to specialize in instruments with a given maturity), so that short-term and long-term interest rates are determined relatively independently. On the contrary, modern theory (witnessing the rapid development and sophistication of primary, secondary, and derivative markets for fixed-income securities) insists on the strong connections between interest rates of different maturities that are induced by the presence of arbitrageurs. In the ideal world of complete markets and absence of arbitrage opportunities, one can establish a unifying formula that gives the price of any zero-coupon bond as the expectation of the actualization factor (for the corresponding period) under a probability distribution $Q$ (called the risk-adjusted probability) that incorporates risk premia:

$$B(t, t+\tau) = E_t Q\left[\exp - \int_0^{\tau} \tau(t+s)\, ds\right], \tag{8.28}$$

where $r(t) = \lim_{\tau \to 0} Y(t, t+\tau)$ is the "instantaneous" interest rate (it can be identified in practice as the overnight repo rate in the interbank market). The notation $E_t^Q$ means that the expectation is taken under the risk-adjusted probability $Q$ and conditionally on the information available at date $t$. Using (8.27), the definition of the yield to maturity, we obtain

$$Y(t, t+\tau) = -\frac{1}{\tau} \ln E_t Q\left[\exp - \int_0^{\tau} r(t+s)\, ds\right]. \tag{8.29}$$

In the absence of uncertainty, the expectation operator disappears, the "ln" and "exp" cancel out, and we obtain a very simple formula:

$$Y(t, t+\tau) = \frac{1}{\tau} \int_0^{\tau} r(t+s)\, ds,$$

according to which the interest rate between date $t$ and $t+\tau$ equals the average of the (future) instantaneous rate over the same period. Equation (8.29) is more complicated because it takes uncertainty into account and contains nonlinearities,[13] but as a first approximation one can say that $Y(t, t+\tau)$ reflects the expectations formed by the market about the average of future values of the short rate $r$ on the corresponding period $[t, t+\tau]$.

For practical purposes, several economists have developed Markovian models where all relevant information is summarized by a small number (typically one or two) of state variables or factors (typically the short rate and possibly the long-term rate $\ell$) that capture most of the uncertainty about future rates and allow computing the yield curve only as a function of these state variables. For instance, Vasicek (1977)[14] uses continuous-time arbitrage theory to build a one-factor model that gives an explicit formula for the term structure:

$$Y(t, t+\tau) = \ell + (r(t) - \ell)\alpha(\tau) + \sigma^2\beta(\tau), \tag{8.30}$$

where $r(t)$ is the short rate at date $t$ (the only state variable), $\ell$ and $\sigma$ are two parameters (long rate and volatility, assumed to be constant), and $\alpha(\tau)$, $\beta(\tau)$ are weight functions (such that $\alpha(0) = 1$, $\beta(0) = 0$, $\alpha(+\infty) = 0$). More complex models were developed subsequently that allow for nonconstant $\ell$ and $\sigma$, but no explicit formulas are available in general.

### 8.3.2 Measuring Interest Rate Risk Exposure

Consider a simple model of a bank that finances long-term assets $A$ through short-term deposits $D$ and equity $E$ (with $A = E + D$, all values nominal). Suppose that both deposits and assets are repaid at fixed dates $t + \tau_D$ and $t + \tau_A$, respectively. Assuming that both deposits and assets can be traded on the market (and neglecting default risk), we can compute the market value at date $t$ of the bank's equity:

$$V_t = AB(t, t+\tau_A) - DB(t, t+\tau_D). \tag{8.31}$$

Notice that $V_t$ differs from the book value $E = A - D$, which does not account for market rate or for maturities. Under the classical simplifying assumption that the term structure is flat ($Y(t, t+\tau) \equiv r(t)$ for all $\tau$), $V_t$ is a deterministic function of $r(t)$, and formula (8.31) becomes

$$V_t = V(r(t)) = A(\exp -\tau_A r(t)) - D(\exp -\tau_D r(t)). \tag{8.32}$$

Suppose now that at date 0, a bank manager tries to predict the effect of changes in interest rates on $V_t$, the market value of the bank at the future date $t$. Under the assumptions behind (8.32), $V_t$ only depends on the future value of the (unique) interest rate $r(t)$. A natural way to measure the exposure of the bank to interest rate risk is to compute the sensitivity of $V$ to variations of $r(t)$:

$$S \stackrel{\text{def}}{=} -\frac{1}{V}\frac{dV}{dr} = \frac{\tau_A A(\exp -\tau_A r(t)) - \tau_D D(\exp -\tau_D r(t))}{V_t},$$

which can be rewritten as

$$S = \lambda\tau_A + (1-\lambda)\tau_D, \tag{8.33}$$

where

$$\lambda = \frac{A \exp -\tau_A r(t)}{V_t}$$

is a measure of leverage ($\lambda > 1$), and $\tau_A$ and $\tau_D$ are, respectively, the durations of assets and deposits. Notice that since $\lambda > 1$ and $\tau_A > \tau_D$, $S$ is positive, which means that the bank is exposed to a rise in interest rates, $dV/dr < 0$. Notice also that $S$ increases when $\lambda$ increases (leverage) and when $\tau_A - \tau_D$ increases (maturity transformation).

This methodology can be extended to the case where assets and liabilities involve several dates of payment.[15] More elaborate models of the term structure can easily be used, like factor models à la Vasicek (1977) or Cox, Ingersoll, and Ross (1981), leading to formulas where $V_t$ depends not only on the short rate $r(t)$ but also on the long rate and on a volatility index. Such methods provide estimates of the sensitivity of the bank's net worth, not only to a parallel movement of interest rates like (8.33) but also to deformations of the shape (slope and curvature) of the yield curve.

### 8.3.3 Applications to Asset Liability Management

Being aware of the dangers of a too large exposure to interest rate risk, large banks have progressively developed different methods, regrouped under the general term of asset liability management (ALM) methods, for coping with this risk. We give here a brief outline of the general principles behind these methods and refer the reader to Fabozzi and Konishi (1991) for a detailed presentation.

A first way to decrease interest rate exposure is to look for a better matching of maturities on the asset and liability sides of the balance sheet. This can be obtained both by decreasing the maturity on the asset side (e.g., by securitizing loans) and by increasing the maturity on the liability side (e.g., by substituting long-term debt to demandable deposits). Of course this logic cannot be pursued to the limit because, in our view, asset transformation belongs to the core of commercial banks' activities. Imposing a perfect matching of asset and liability maturities would in fact be equivalent to a narrow banking system.

Given that transformation cannot be completely eliminated, banks have tried to implement a second approach for decreasing their interest rate exposure, namely, using derivative instruments like futures, options, and swaps. For example, consider the bank of section 8.3.2 with volumes of assets $A$ and deposits $D$ and respective repayment dates $t + \tau_A$ and $t + \tau_D$. Suppose that at date 0 the bank's management wants to hedge the market value of the bank's equity at date $t$ (say, the date on which the board of administrators meets) against fluctuations of the interest rate $r(t)$. A natural hedging instrument is a futures contract delivering at date $t$ a fixed-income instrument (say, a Treasury bill of nominal $B$ and maturity $t+\tau$) at a prespecified price $F$, to be paid at date $t$. The market value $B_t$ of the T-bill at date $t$ can be determined by using the formula developed previously:

$$B_t = B \exp -\tau Y(t, t+\tau).$$

Under the simplifying assumption of a flat yield curve $(Y(t, t+\tau) \equiv r(t))$ this becomes

$$B_t = B(r(t)) = B \exp -\tau r(t).$$

Consider now the hedged market value $V_t^H$ of the bank at date $t$ if it sells $x$ futures contracts. It consists of the commercial market value $V_t$ plus the gains or losses on the futures markets,

$$V_t^H = V_t + x(F - B_t).$$

Under the simplifying assumptions, this hedged market value only depends on $r(t)$:

$$V_t^H = V^H(r(t)) = V(r(t)) + x[F - B(r(t))].$$

Suppose now that the bank's managers anticipate a value $r^*$ for the interest rate at date $t$ but want to protect themselves against small fluctuations around this anticipated value. Using a first-order Taylor expansion, one can write

$$V_t^H \sim V^H(r^*) + [V'(r^*) - xB'(r^*)](r(t) - r^*).$$

To the first order, hedging will be perfect if the derivative of $V^H$ is zero, which is obtained for a unique value of the futures position $x$:

$$x^* = \frac{V'(r^*)}{B'(r^*)}.$$

$x^*$ can also be expressed as a function of the sensitivities

$$S = -\frac{V'(r^*)}{V(r^*)} = \lambda\tau_A - (1-\lambda)\tau_D$$

for the bank's value and

$$S_F = -\frac{B'(r^*)}{B(r^*)} = r$$

for the futures contract. We finally get

$$x^* = \frac{S}{S_F} \times \frac{V(r^*)}{B(r^*)}.$$

Notice that, with this futures position, hedging will be perfect only to the first order. For a better approximation, it is important to take also into account second-order derivatives, related to the convexity of functions $V(r)$ and $B(r)$.

## 8.4 Market Risk

The modern theory of portfolio management was developed by Sharpe (1964), Lintner (1965), and Markowitz (1952). It is of course interesting for banks, which often hold large portfolios of marketable assets. More important, this portfolio theory has led to another paradigm for banking behavior, essentially developed by Pyle (1971) and Hart and Jaffee (1974). The idea is to assimilate all assets and liabilities of the bank into securities of a particular sort, and to consider the whole bank itself as a portfolio manager who controls an enormous portfolio of these securities. In this approach, the only specificity of the bank's liabilities is that they correspond to short positions in the bank's portfolio. Portfolio theory is discussed briefly in section 8.4.1. The Pyle-Hart-Jaffee approach is presented in section 8.4.2, and an application to the analysis of capital requirements is presented in section 8.4.3.

### 8.4.1 Portfolio Theory: The Capital Asset Pricing Model

This presentation of portfolio theory is very brief, since there are excellent references on that topic (e.g., Ingersoll 1987). The brilliant idea of Sharpe, Lintner, and Markowitz was to simplify the general problem of optimal portfolio selection by assuming that investors' preferences $U$ depend only on the first two moments $\mu$ and $\sigma^2$ (mean and variance) of the random liquidation value of their portfolio. This can be justified by assuming that investors have quadratic Von Neumann–Morgenstern preferences, or that stochastic distributions of asset returns belong to a particular parameterized family (normal, or more generally, elliptical random variables).

Let $W$ denote the initial wealth of the investor and $x_i$ $(i = 1, \dots, N)$ be the amount invested in the $i$th risky asset. The vector $x = (x_1, \dots, x_N)$ thus represents the risky portfolio held by the investor. The rest of her wealth $(W - \sum_{i=1}^{N} x_i)$ is invested in a riskless asset of return $R_0$. The random returns $(\tilde{R}_i)_i$ of risky assets have first and second moments denoted

$$E(\tilde{R}_i) = R_0 + \rho_i \qquad (i = 1, \dots, N),$$

$$\operatorname{cov}(\tilde{R}_i, \tilde{R}_j) = v_{ij} \qquad (i, j = 1, \dots, N).$$

At the end of the period, the investor's wealth is

$$\tilde{W} = \left[\left(W - \sum_{i=1}^{N} x_i\right) R_0 + \sum_{i=1}^{N} x_i \tilde{R}_i\right].$$

The first two moments of this random variable are

$$\mu = E[\tilde{W}] = W\left(R_0 + \sum_{i=1}^{N} x_i \rho_i\right), \tag{8.34}$$

$$\sigma^2 = \mathrm{var}(\tilde{W}) = \left(\sum_{i=1}^{N}\sum_{j=1}^{N} v_{ij} x_i x_j\right) W^2. \tag{8.35}$$

Under the mean variance assumption, the investor will choose $x$ so as to maximize her utility function $U(\mu, \sigma^2)$ (where $\partial U/\partial \mu > 0$, $\partial U/\partial \sigma^2 < 0$) under constraints (8.34) and (8.35). The first-order conditions for a maximum are

$$\frac{\partial U}{\partial \mu}\frac{\partial \mu}{\partial x_i} + \frac{\partial U}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial x_i} = 0,$$

or

$$\frac{\partial U}{\partial \mu} \cdot \rho_i + 2\frac{\partial U}{\partial \sigma^2}\sum_j v_{ij} x_j = 0 \qquad (i = 1, \ldots, N). \tag{8.36}$$

Let $\rho = (\rho_1, \ldots, \rho_N)$ denote the vector of expected excess returns, and $V = (v_{ij})_{i,j=1,\ldots,N}$ the variance-covariance matrix of risky assets, assumed to be invertible. The first-order conditions can be written in a more compact form:

$$-\lambda\rho + Vx = 0,$$

where[16]

$$\lambda = \frac{-(\partial U/\partial \mu)}{2(\partial U/\partial \sigma^2)},$$

or

$$x = \lambda V^{-1}\rho. \tag{8.37}$$

Since $V$ and $\rho$ are independent of the investor, this relation implies that all investors will choose colinear risky portfolios. A more financially appealing way of expressing this result is that all investors obtain their preferred portfolio by a combination of the riskless asset and a fixed portfolio $V^{-1}\rho$, interpreted as a mutual fund. The only difference in behavior among investors is captured by the coefficient $\lambda$: a more risk-averse agent will buy more of the riskless asset and less of the risky mutual fund.

An interesting aspect of the Capital Asset Pricing Model (CAPM) (which is not farther discussed in this book) consists in writing a general equilibrium formulation of equation (8.37). Consider the two main implications of this classical model. If market portfolio $x_M$ is defined as the aggregation of all individual (risky) portfolios, (8.37) then has two important consequences:

- Since $x_M$ is the sum of individual risky portfolios, which are all colinear, these individual risky portfolios may conversely be considered as all colinear to $x_M$. Therefore, the market portfolio can be used as the mutual fund described previously.
- The expected excess return of any asset $i$ at equilibrium is proportional to the regression coefficient $\beta_i$ of $\tilde{R}_i$ on the return $\tilde{R}_M$ of the market portfolio. Indeed, $\beta_i$ is itself proportional to the $i$th component of the vector $Vx_M$, which according to (8.37) (and the previous observation) is proportional to $\rho_i$.

The following section discusses the application of the CAPM to the modeling of banks' behavior: the Pyle-Hart-Jaffee approach.

### 8.4.2 The Bank as a Portfolio Manager: The Pyle-Hart-Jaffee Approach

This section shows how mean-variance analysis can provide an adequate tool for modeling the management of market risk by commercial banks. Pyle (1971) and Hart and Jaffee (1974) studied a new paradigm for describing the behavior of financial intermediaries. When this paradigm is compared with the one developed by Klein (1971), the main differences are that in the Pyle-Hart-Jaffee approach the markets for assets (and liabilities) are assumed to be *competitive* and that *risk* is explicitly taken into account.[17]

As a first illustration of this new paradigm, consider the simple case of only two risky financial products $L$ and $D$, to be interpreted later as loans and deposits. The bank is assimilated to a portfolio manager, who has to decide the amounts $x_L$ and $x_D$ to be invested in these two risky activities, the rest of his wealth being invested in reserves (riskless asset). Make no assumption a priori on the sign of $x_L$ and $x_D$. The competitive behavior means that the bank takes the returns $\tilde{r}_L$, $\tilde{r}_D$, and $r$ of these activities as given. Therefore, the (random) profit of the bank is

$$\tilde{\pi} = [\tilde{r}_L x_L + \tilde{r}_D x_D + r(W - x_L - x_D)],$$

or

$$\tilde{\pi} = Wr + (\tilde{r}_L - r)x_L + (\tilde{r}_D - r)x_D.$$

Using the same notation as before, the objective function of the bank can be expressed as

$$\Phi(x) = U(E(\tilde{\pi}), \mathrm{var}(\tilde{\pi})).$$

If $x^*$ maximizes $\Phi$, the first-order condition implies, as before,

$$x^* = \lambda V^{-1}\rho, \tag{8.38}$$

where

$$V = \begin{pmatrix} \mathrm{var}(\tilde{r}_L) & \mathrm{cov}(\tilde{r}_L, \tilde{r}_D) \\ \mathrm{cov}(\tilde{r}_L, \tilde{r}_D) & \mathrm{var}(\tilde{r}_D) \end{pmatrix},$$

$$\lambda = \frac{-(\partial U/\partial \mu)}{2(\partial U/\partial \sigma^2)}, \quad \text{and} \quad \rho = \begin{pmatrix} \bar{r}_L - r \\ \bar{r}_D - r \end{pmatrix}.$$

**Result 8.6** If $\bar{r}_D < r < \bar{r}_L$ and $\mathrm{cov}(\tilde{r}_L, \tilde{r}_D) > 0$, then $x_L^* > 0$ and $x_D^* < 0$.

This result can be seen as an endogenous explanation for the intermediation activity of banks. If the expected excess returns on the deposit and loan activities are respectively negative and positive, and if the covariance between these returns is positive, then a competitive portfolio manager will invest a negative amount on deposits (he will issue such instruments) and a positive amount on loans. In other words, he would have loans on the asset side of the balance sheet ($L = x_L^*$) and deposits on the liability side ($D = -x_D^* > 0$). If the conclusion does not hold, it means that either $x_D^* > 0$, and the bank borrows at the riskless rate to invest in two types of loans, or $x_L^* < 0$, in which case the bank offers two types of deposits and invests the proceeds at the riskless rate.

***Proof*** The proof is derived from equation (8.38):

$$x^* = \begin{pmatrix} x_L^* \\ x_D^* \end{pmatrix} = \lambda V^{-1}\rho = \frac{\lambda}{\Delta}\begin{pmatrix} \mathrm{var}(\tilde{r}_D) & -\mathrm{cov}(\tilde{r}_L, \tilde{r}_D) \\ -\mathrm{cov}(\tilde{r}_L, \tilde{r}_D) & \mathrm{var}(\tilde{r}_L) \end{pmatrix}\begin{pmatrix} \bar{r}_L - r \\ \bar{r}_D - r \end{pmatrix},$$

where the following formula is used for inverting a $2 \times 2$ matrix:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -c \\ -b & a \end{pmatrix},$$

and $\Delta = ad - bc$ is the determinant of $V$.

Now $\lambda$ is positive because of risk aversion

$$\left(\frac{\partial U}{\partial \mu} > 0, \frac{\partial U}{\partial \sigma^2} < 0\right),$$

and $\Delta$ is positive because $V$ is a positive definite matrix. Therefore,

$$x_L^* = \frac{\lambda}{\Delta}\left[\underbrace{\mathrm{var}(\tilde{r}_D)}_{>0}\underbrace{(\bar{r}_L - r)}_{>0} - \underbrace{\mathrm{cov}(\tilde{r}_L, \tilde{r}_D)}_{>0}\underbrace{(\bar{r}_D - r)}_{<0}\right],$$

and $x_L^*$ is positive.

Similarly,

$$x_D^* = \frac{\lambda}{\Delta}\left[\underbrace{-\mathrm{cov}(\tilde{r}_L, \tilde{r}_D)}_{<0}\underbrace{(\bar{r}_L - r)}_{>0} + \underbrace{\mathrm{var}(\tilde{r}_L)}_{>0}\underbrace{(\bar{r}_D - r)}_{<0}\right],$$

and $x_D^*$ is negative. ■

Notice that result 8.6 gives only a sufficient condition. The necessary condition for $x_L^* > 0$ is $\mathrm{var}(\tilde{r}_D)(\bar{r}_L - r) > \mathrm{cov}(\tilde{r}_L, \tilde{r}_D)(\bar{r}_D - r)$, and for $x_D^* < 0$ is $\mathrm{cov}(\tilde{r}_L, r_D) \cdot (\bar{r}_L - r) > \mathrm{var}(\tilde{r}_L)(\bar{r}_D - r)$. This allows financial intermediaries to exist even if $\bar{r}_D > r$ or $\bar{r}_L < r$, provided that $\mathrm{cov}(\tilde{r}_L, \tilde{r}_D) > 0$.

Another interesting outcome of the mean-variance approach is a comparative statics analysis of the bank's behavior. How are the volumes of deposits attracted and loans issued affected by changes in the expectations or variances of returns? The answer is given by the following result.

**Result 8.7**

1. $x_L^*$ is an increasing function of $(\bar{r}_L - r)$ and a decreasing function of $(\bar{r}_D - r)$ and $\mathrm{var}(\tilde{r}_L)$.
2. $|x_D^*|$ is an increasing function of $(\bar{r}_L - r)$ and a decreasing function of $(\bar{r}_D - r)$ and $\mathrm{var}(\tilde{r}_D)$.

***Proof*** It is a direct consequence of the formulas for $x_L^*$ and $x_D^*$ obtained in the proof of result 8.6 (remembering that $x_D^* < 0$). The only properties that are not obvious are

$$\frac{\partial x_L^*}{\partial\, \mathrm{var}(\tilde{r}_L)} = -\frac{x_L^*}{\Delta}\frac{\partial \Delta}{\partial\, \mathrm{var}(\tilde{r}_L)} < 0$$

and

$$\frac{\partial |x_D^*|}{\partial\, \mathrm{var}(\tilde{r}_D)} = -\frac{|x_D^*|}{\Delta}\frac{\partial \Delta}{\partial\, \mathrm{var}(\tilde{r}_D)} < 0.$$ ■

Hart and Jaffee (1974) have extended the analysis of Pyle to the case of an arbitrary number of assets and liabilities, also introducing additional constraints. For instance, a no-short-sales requirement can be introduced by constraining $x_i$ to be positive if $i$ belongs to the asset side of the balance sheet, and negative if $i$ belongs to the liability side (instead of endogenizing it as in Pyle 1971). Similarly, reserve requirements, liquidity ratios, and solvency ratios can be introduced as linear constraints on the different entries of the bank's balance sheet. Thus a competitive theory of financial intermediaries is obtained in which all the posts of the balance sheet are determined in the same fashion as the portfolio of an individual investor. This approach has several problematic aspects.

As in the CAPM, the model predicts that all banks should hold colinear (risky) portfolios. This is not consistent with the diversity of banks' balance sheets that is observed in practice.

If the bank's capital is considered as just another liability, the wealth $W$ of the bank becomes endogenous. No utility function can be assumed, since the identity of the bank's owners becomes irrelevant. The only restriction on the whole balance sheet of the bank (including equity) is that it be a mean-variance-efficient portfolio. Then there is a fundamental indetermination on the size of banks at equilibrium. If a given balance sheet is mean-variance-efficient, then any multiple of this balance sheet is also mean-variance-efficient.

Finally, if the possibility of bank failure is taken into account, the symmetry between assets and liabilities breaks. It is no longer possible to assume that the rate of return on equity demanded by investors (the stockholders or the debtholders of the bank) is independent of the assets chosen by the bank because the latter affect the probability of failure of the bank. This question is examined in section 8.4.3, where an application of the portfolio model to the question of solvency ratios is developed.

### 8.4.3 An Application of the Portfolio Model: The Impact of Capital Requirements

Since January 1993 all commercial banks in the European Union[18] are submitted to a common solvency requirement, inspired by a similar requirement (the so-called Cooke ratio) adopted in December 1987 by the Bank of International Settlements. The portfolio model presented previously will allow an investigation of the consequences of such a regulation on the behavior of commercial banks. This section is inspired by Koehn and Santomero (1980), Kim and Santomero (1988), and Rochet (1992). The model is as follows.

At date 0 the bank chooses the composition of its portfolio of assets and invests the amounts $x_0, \ldots, x_n$ on $(n+1)$ securities, taking as given the random returns $\tilde{r}_i$ on these securities. Security zero is assumed to be riskless ($r_0$ is deterministic and normalized to zero). For simplicity, liabilities, deposits $D$, and equity capital $K$, are fixed. Deposits are remunerated at the riskless rate. At date 1, the bank is liquidated, and stockholders receive the difference between the value of the bank's assets and the value of deposits so that $D$ vanishes out of this expression:

$$\tilde{K}_1 = K + \sum_{i=1}^{n} x_i \tilde{r}_i.$$

The bank behaves as a portfolio manager and seeks to maximize

$$\Phi(x) = Eu(\tilde{K}_1), \tag{8.39}$$

where $u$ is a concave increasing Von Neumann–Morgenstern utility function. Notice that the bank's owner behaves as if the bank had full liability ($\tilde{K}_1$ can be negative). This is inconsistent with the main justification of capital requirements, namely, the prevention of bank failures. This point, raised initially by Keeley and Furlong (1990), is discussed later. For the moment, we focus on the original formulation of Kim and Santomero and show that capital requirements may severely distort the allocation of assets by banks.

To allow application of mean-variance analysis, assume that the joint distribution of returns is normal, with an invertible variance-covariance matrix $V$. $\rho$ denotes the vector of expected excess returns. Under this assumption, $\tilde{K}_1$ is itself a normal random variable of mean

$$\mu = E(\tilde{K}_1) = K + \langle x, \rho \rangle,$$

(where $\langle a, b \rangle$ denotes the scalar product of vectors $a$ and $b$), and variance

$$\sigma^2 = \mathrm{var}(\tilde{K}_1) = \langle x, Vx \rangle.$$

The normality assumption implies that $\Phi$ only depends on $\mu$ and $\sigma^2$:

$$\Phi(x) = U(K + \langle x, \rho \rangle, \langle x, Vx \rangle),$$

where $U$ is defined:

$$U(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u(\mu + t\sigma) \exp -\frac{t^2}{2} dt.$$

**The Behavior of a Full-Liability Bank in the Absence of a Solvency Regulation**

The behavior of a full-liability bank in the absence of a solvency regulation is characterized by the solution of

$$\mathscr{P}_1 \begin{cases} \max \Phi(x) \\ x \in \mathbb{R}^n. \end{cases}$$

The solution of this program, $x_1^*$, is such that

$$x_1^* = \lambda_1 V^{-1} \rho,$$

where

$$\lambda_1 = \frac{-(\partial U/\partial \mu)}{2(\partial U/\partial \sigma^2)} > 0.$$

As already noted, this formulation is inconsistent with previous assumptions, since the bank does not take into account its limited liability clause. However, failure occurs when $\tilde{K}_1 < 0$. The probability of this event is easily computed, since $\tilde{K}_1$ follows a Gaussian distribution of mean $\mu$ and variance $\sigma^2$. Therefore,

$$\frac{\tilde{K}_1 - \mu}{\sigma}$$

follows a normalized Gaussian distribution of cumulative function $N(\cdot)$, and

$$\text{Proba}[\tilde{K}_1 < 0] = \text{Proba}\left[\frac{\tilde{K}_1 - \mu}{\sigma} < -\frac{\mu}{\sigma}\right] = N\left(-\frac{\mu}{\sigma}\right).$$

Thus, the probability of failure of a bank choosing an asset portfolio $x^*$ and having initial net worth $K$ is

$$\text{Proba}[\tilde{K}_1 < 0] = N\left[-\frac{K + \langle x^*, \rho\rangle}{(\langle x^*, Vx^*\rangle)^{1/2}}\right].$$

A solvency ratio is usually computed as the ratio of the level of capital divided by a weighted sum of assets $\sum_{i=1}^{n} \alpha_i x_i^*$ (weights $\alpha_i$ are assumed to reflect the relative riskiness of assets; in particular, it is natural to take $\alpha_0 = 0$):

$$CR = \frac{K}{\langle \alpha, x^* \rangle}.$$

The solvency regulation imposes an upper bound on this ratio. The rationale for it is that if banks behave as described by program $\mathscr{P}_1$, then their probability of failure will be a decreasing function of the capital ratio. This is established in the following result.

**Result 8.8** In the absence of a solvency regulation, and if banks do not take into account the limited liability clause, the probability of banks' failure is a decreasing function of their capital ratio, independent of the non-negative weights used in the computation of the ratio.

***Proof*** Because of the mean-variance property, all banks choose colinear portfolios. Let $x_1^*(K)$ denote the portfolio chosen by a bank having net worth $K$:

$$x_1^*(K) = \sigma(K) x_M,$$

where $x_M$ is defined as the portfolio colinear to $V^{-1}\rho$ such that its return has a unitary variance. $\sigma(K)$ is a non-negative constant, equal to the standard deviation of the return of $x_1^*(K)$. Using a similar notation, $\mu(K)$ represents the expectation of $\tilde{K}_1$:

$$\mu(K) = K + \langle x_1^*(K), \rho \rangle = K + \sigma(K)\langle x_M, \rho \rangle.$$

As a consequence,

$$\text{Proba}(\tilde{K}_1 < 0) = N\left(-\frac{\mu(K)}{\sigma(K)}\right) = N\left(-\langle x_M, \rho\rangle - \frac{K}{\sigma(K)}\right),$$

whereas

$$CR(K) = \frac{K}{\langle x_M, \alpha\rangle \sigma(K)}.$$

Therefore,

$$\text{Proba}(\tilde{K}_1 < 0) = N(-\langle x_M, \rho\rangle - \langle x_M, \alpha\rangle CR(K)).$$

Since $\langle x_M, \rho\rangle$ is positive, the probability of failure is a decreasing function of $CR(K)$. ∎

**The Behavior of a Full-Liability Bank after Introduction of a Solvency Regulation**
Since a capital ratio is a good indicator of the failure risk of a bank, it may seem reasonable to impose a lower bound on this ratio to limit the risk of failure. However, introducing such a ratio may alter the asset allocation of the bank, since its behavior is now characterized by a new program (in the case of a full-liability bank):

$$\mathscr{P}_2 \begin{cases} \max\ \Phi(x) \\ \langle \alpha, x\rangle \leq K, \end{cases}$$

where, without loss of generality, the minimum capital ratio is normalized to 1:

$$CR = \frac{K}{\langle \alpha, x\rangle} \geq 1 \Leftrightarrow \langle \alpha, x\rangle \leq K.$$

If $v$ denotes the Lagrange multiplier associated with this constraint, the first-order condition of $\mathscr{P}_2$ becomes

$$\nabla\Phi(x_2^*) = \frac{\partial U}{\partial \mu}\rho + 2\frac{\partial U}{\partial \sigma^2} V x_2^* = v\alpha.$$

Therefore,

$$x_2^* = V^{-1}[\lambda_2 \rho + v_2 \alpha],$$

where

$$\lambda_2 = \frac{-(\partial U/\partial \mu)}{2(\partial U/\partial \sigma^2)} \quad \text{and} \quad \frac{v_2 = v}{2(\partial U/\partial \sigma^2)}.$$

Thus the following result has been proved.

**Result 8.9** If $\alpha$ is not colinear to $\rho$, and if the solvency constraint is binding, the bank will choose an inefficient portfolio; $x_2^*$ will not be colinear to $V^{-1}\rho$.

Therefore, in general (if $\alpha$ is not colinear to $\rho$), introducing a solvency regulation will entail an inefficient asset allocation by banks. The total volume of their risky portfolio will decrease, but its composition will be distorted in the direction of more risky assets. Kim and Santomero (1988) showed an example in which the probability of failure *increases* after the capital ratio is introduced. The explanation is that the adverse structure effect (recomposition of the risky portfolio) dominates the direct volume effect.

However, there is a simple way (in theory) to suppress this adverse recomposition effect.

**Result 8.10** If the weights $\alpha_i$ used in the capital ratio are proportional to the systematic risks $\beta_i$ of the risky assets, the solvency regulation becomes efficient. All banks choose efficient portfolios, and their probability of failure decreases.

***Proof*** If $\alpha$ is colinear to $\beta$ (or to $\rho$, since the CAPM implies that vectors $\beta$ and $\rho$ are themselves colinear), the first-order condition of $\mathscr{P}_2$ becomes

$$x_2^* = (\lambda_2 + \nu_2)V^{-1}\rho.$$

Therefore, $x_2^*$ is mean-variance-efficient. Moreover, the probability of failure is a decreasing function of $CR$ (as in $\mathscr{P}_1$). This implies that imposing a capital ratio (with correct weights, that is, proportional to the market evaluation of risk given by the $\beta_i$'s) is instrumental for limiting the failure risk of banks. ■

To conclude this discussion of the portfolio model applied to banks' solvency ratios, return to the criticism of Keeley and Furlong (1990). What happens when the limited liability option is correctly taken into account by the bank? Rochet (1992) has studied this question. He shows that the mean-variance approach can still be used, but that the indirect utility function of the bank has a different expression, $U_{LL}(\mu, \sigma^2)$. The bank's decision problem becomes

$$\mathscr{P}_3 \begin{cases} \max \psi(x) \\ \langle \alpha, x \rangle \leq K, \end{cases}$$

where

$$\psi(x) = U_{LL}(K + \langle \mu, x \rangle, \langle x, Vx \rangle),$$

and

$$U_{LL}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \int_{-\mu/\sigma}^{+\infty} u(\mu + t\sigma) \exp -\frac{t^2}{2} dt.$$

$U_{LL}$ is the indirect utility function under limited liability. Rochet shows that $U_{LL}$ is not always decreasing in $\sigma^2$. For low levels of $K$, the bank chooses a portfolio with maximal risk and minimum diversification. As a result, a solvency regulation (even with correct weights) is not sufficient for taking care of moral hazard. Rochet suggests introducing an additional regulation, namely, a minimum level of capital, independent of the size of the banks' assets.

## 8.5 Problems

### 8.5.1 The Model of Prisman, Slovin, and Sushka

Recall that the expected profit of the bank studied in section 8.2.2 is given by

$$\Pi(r_L, r_D) = \Pi_0(r_L, r_D) - C(R, \theta),$$

where

$$\Pi_0(r_L, r_D) = (r_L - r)L(r_L) + (r - r_D)D(r_D)$$

is the (gross) profit obtained in the Monti-Klein model (assuming $D' > 0$ and $L' < 0$),

$$R = D(r_D) - L(r_L)$$

is the level of reserves, and

$$C(R, \theta) = r_p E[\max(0, \tilde{x} - R)]$$

is the expected cost of liquidity shortages. Here $\theta$ represents any parameter that will influence this cost; it can be $r_p$, the penalty rate, or the variance of $\tilde{x}$. The first-order conditions, which determine the optimal values $r_L^*$ and $r_D^*$, are thus

$$\frac{\partial \Pi}{\partial r_L}(r_L, r_D) = \frac{\partial \Pi_0}{\partial r_L}(r_L, r_D) + \frac{\partial C}{\partial R}(R, \theta)L'(r_L) = 0, \tag{8.40}$$

$$\frac{\partial \Pi}{\partial r_D}(r_L, r_D) = \frac{\partial \Pi_0}{\partial r_D}(r_L, r_D) - \frac{\partial C}{\partial R}(R, \theta)D'(r_D) = 0. \tag{8.41}$$

Part 1 can be skipped by readers familiar with the techniques of convex analysis.

1. Compute the matrix of the second-order derivatives of $\Pi$ at $(r_L^*, r_D^*)$, and show that it is definite negative under the following assumptions:

$$DD'' - 2D'^2 < 0, \qquad LL'' - 2L'^2 < 0.$$

2. By applying the implicit function theorem to equations (8.40) and (8.41) and using the inequalities given in part 1, show that

$$\frac{dr_L^*}{d\theta} \quad \text{and} \quad \frac{dr_D^*}{d\theta}$$

both have the same sign as

$$\left(-\frac{\partial^2 C}{\partial R \partial \theta}\right).$$

In other words, all changes that decrease the marginal cost of reserves also increase $r_L^*$ and $r_D^*$. (*Hint*: Use

$$\frac{\partial^2 \Pi}{\partial r_D^2} L' + \frac{\partial^2 \Pi}{\partial r_D \partial r_L} = \left(r - r_D - \frac{\partial C}{\partial R}\right) D'' L' - 2D'L'$$

and the first-order conditions.)

3. Prove the first part of result 8.4 by showing that $\partial C / \partial R$ decreases with $r_p$.

4. Suppose that $\tilde{x}$ equals $\sigma \tilde{x}_0$, where $\tilde{x}_0$ has zero mean and unit variance. Show that $\partial C / \partial R$ decreases with $\sigma$ (where $\sigma > 0$) if and only if the optimal level of reserves is positive. Deduce the second part of result 8.4, namely, $dL/d\sigma < 0$ if $R > 0$.

### 8.5.2 The Risk Structure of Interest Rates

This problem is adapted from Merton (1974). Recall that the market value of a risky debt in the Merton model studied in section 8.1.2 is given by

$$D_0 = VN(h_1) + De^{-rT} N(h_2) \tag{8.42}$$

with

$$\begin{cases} h_1 = \dfrac{1}{\sigma\sqrt{T}} \log d - \dfrac{1}{2}\sigma\sqrt{T}, \\ h_2 = -h_1 - \sigma\sqrt{T}, \end{cases}$$

and

$$d = \frac{De^{-rT}}{V}$$

is the nominal debt-to-asset ratio.

1. Show that (8.42) can be rewritten as

$$\frac{D_0}{De^{-rT}} = \min_x \left\{ \frac{N(x)}{d} + N(-x - \sigma\sqrt{T}) \right\}. \tag{8.43}$$

2. The interest spread $H$ is defined by

$$H = -\frac{1}{T} \log \frac{D_0}{De^{-rT}}.$$

By applying the envelope principle to equation (8.43), show that $H$ increases with $d$, $H$ increases with $\sigma$, and $HT$ increases with $T$.

### 8.5.3 Using the CAPM for Loan Pricing

A bank computes the nominal interest rate $r_L$ that it charges on a certain type of loan by the following formula:

$$\frac{(1-\delta)(1+r_L) - \gamma_L - (1-\alpha)(1+r)}{\alpha} = 1 + r + \pi,$$

where

$\delta =$ the proportion of defaulting loans, assuming the proceeds of a defaulting loan are zero.

$\gamma_L =$ the management cost per unit of loan.

$r =$ the interbank rate, taken as the riskless rate.

$\pi =$ the risk premium demanded by stockholders.

$\alpha =$ the capital coefficient required for this type of loan.

1. Compute the expected return on a loan, and show that the preceding pricing formula is closely related to the CAPM approach.
2. Assuming that the bank has a monopoly power on the loans side but faces a competitive market on the liabilities side, compute what should be the modified pricing formula, for given $\delta$ and $\beta$, as a function of the elasticity of the demand for loans.
3. How should the formula be modified if the bank faces a tax on profits at a rate $\tau$?

## 8.6 Solutions

### 8.6.1 The Model of Prisman, Slovin, and Sushka

1. Compute the second-order derivatives of $\Pi$:

$$\frac{\partial^2 \Pi}{\partial r_L^2} = \frac{\partial^2 \Pi_0}{\partial r_L^2} + \frac{\partial C}{\partial R} L'' - \frac{\partial^2 C}{\partial R^2} L'^2,$$

$$\frac{\partial^2 \Pi}{\partial r_L \partial r_D} = \frac{\partial^2 C}{\partial R^2} L' D', \quad \text{and} \quad \frac{\partial^2 \Pi}{\partial r_D^2} = \frac{\partial^2 \Pi_0}{\partial r_D^2} - \frac{\partial C}{\partial R} D'' - \frac{\partial^2 C}{\partial R^2} D'^2.$$

Using the first-order conditions, at the optimum

$$\frac{\partial C}{\partial R}(R,\theta) = -\frac{1}{L'}\frac{\partial \Pi_0}{\partial r_L}(r_L^*, r_D^*) = \frac{1}{D'}\frac{\partial \Pi_0}{\partial r_D}(r_L^*, r_D^*).$$

Thus,

$$\frac{\partial^2 \Pi}{\partial r_L^2} = \left(\frac{\partial^2 \Pi_0}{\partial r_L^2} - \frac{L''}{L'}\frac{\partial \Pi_0}{\partial r_L}\right) - \frac{\partial^2 C}{\partial R^2}L'^2,$$

$$\frac{\partial^2 \Pi}{\partial r_D^2} = \left(\frac{\partial^2 \Pi_0}{\partial r_D^2} - \frac{D''}{D'}\frac{\partial \Pi_0}{\partial r_D}\right) - \frac{\partial^2 C}{\partial R^2}D'^2.$$

But now a simple computation yields

$$\begin{cases} \dfrac{\partial^2 \Pi_0}{\partial r_L^2} - \dfrac{L''}{L'}\dfrac{\partial \Pi_0}{\partial r_L} = 2L' - \dfrac{L''L}{L'} \\ \dfrac{\partial^2 \Pi_0}{\partial r_D^2} - \dfrac{D''}{D'}\dfrac{\partial \Pi_0}{\partial r_D} = -2D' + \dfrac{D''D}{D'}. \end{cases}$$

By assumption, these quantities, denoted respectively by $\alpha$ and $\beta$, are negative. Since $C$ is convex in $R$, then

$$\frac{\partial^2 \Pi}{\partial r_L^2} \quad \text{and} \quad \frac{\partial^2 \Pi}{\partial r_D^2}$$

are also negative.

It remains to show that the determinant of the second-order derivative of $\Pi$ is positive:

$$\begin{aligned} H &= \frac{\partial^2 \Pi}{\partial r_L^2}\frac{\partial^2 \Pi}{\partial r_D^2} - \left(\frac{\partial^2 \Pi}{\partial r_L \partial r_D}\right)^2 \\ &= \left(\alpha - \frac{\partial^2 C}{\partial R^2}L'^2\right)\left(\beta - \frac{\partial^2 C}{\partial R^2}D'^2\right) - \left(\frac{\partial^2 C}{\partial R^2}\right)^2 L'^2 D'^2 \\ &= \alpha\beta - \beta\frac{\partial^2 C}{\partial R^2}L'^2 - \alpha\frac{\partial^2 C}{\partial R^2}D'^2. \end{aligned}$$

Since $\alpha$ and $\beta$ are negative, $H$ is positive.

2. A total differentiation of (8.40) and (8.41) with respect to $\theta$ yields

$$\begin{cases} \dfrac{\partial^2 \Pi}{\partial r_L^2}\dfrac{dr_L}{d_\theta} + \dfrac{\partial^2 \Pi}{\partial r_L \partial r_D}\dfrac{dr_D}{d_\theta} = -\dfrac{\partial^2 C}{\partial R \partial \theta}L' \\ \dfrac{\partial^2 \Pi}{\partial r_L \partial r_D}\dfrac{dr_L}{d\theta} + \dfrac{\partial^2 \Pi}{\partial r_D^2}\dfrac{dr_D}{d\theta} = \dfrac{\partial^2 C}{\partial R \partial \theta}D', \end{cases}$$

from which can be deduced

$$\frac{dr_L}{d\theta} = \frac{1}{H}\frac{\partial^2 C}{\partial R \partial \theta}\left[-\frac{\partial^2 \Pi}{\partial r_D^2}L' - \frac{\partial^2 \Pi}{\partial r_D \partial r_L}D'\right],$$

$$\frac{dr_D}{d\theta} = \frac{1}{H}\frac{\partial^2 C}{\partial R \partial \theta}\left[\frac{\partial^2 \Pi}{\partial r_L^2}D' + \frac{\partial^2 \Pi}{\partial r_L \partial r_D}L'\right].$$

Since $H$, the Hessian determinant computed previously, is positive, it remains to show that the terms between the brackets are negative:

$$\frac{\partial^2 \Pi}{\partial r_D^2}L' + \frac{\partial^2 \Pi}{\partial r_D \partial r_L}D' = \left(r - r_D - \frac{\partial C}{\partial R}\right)D''L' - 2D'L'.$$

Using the first-order condition again,

$$r - r_D - \frac{\partial C}{\partial R} = \frac{D}{D'},$$

the following is obtained:

$$\frac{\partial^2 \Pi}{\partial r_D^2}L' + \frac{\partial^2 \Pi}{\partial r_D \partial r_L}D' = L'\left[\frac{D''D}{D'} - 2D'\right] > 0.$$

Similarly,

$$\begin{aligned} -\frac{\partial^2 \Pi}{\partial r_L \partial r_D}L' - \frac{\partial^2 \Pi}{\partial r_L^2}D' &= -\left(r_L - r + \frac{\partial C}{\partial R}\right)L''D' - 2L'D' \\ &= D'\left[\frac{L''L}{L'} - 2L'\right] > 0. \end{aligned}$$

3.–4.

$$\frac{\partial C}{\partial R} = -r_p \text{ Proba}[\tilde{x} \geq R] = -r_p \text{ Proba}\left[\tilde{x}_0 \geq \frac{R}{\sigma}\right].$$

This obviously decreases with $r_p$; therefore part 3 is proved. Moreover, $\partial C/\partial R$ decreases with $\sigma$ if and only if $R > 0$. When $R > 0$, then $drL/d\sigma > 0$ and therefore $dL/d\sigma < 0$.

### 8.6.2 The Risk Structure of Interest Rates

1. Let $\varphi(x)$ denote the term between brackets in (8.43):

$$\varphi(x) = \frac{N(x)}{d} + N(-x - \sigma\sqrt{T}).$$

A straightforward computation shows that $\varphi$ has a unique minimum for $x = h_1$, which establishes part 1 by comparison with (8.42).

2. It is immediate from (8.43) that $D_0/De^{-rT}$ is a decreasing function of $d$ and $\sigma$, as a minimum of decreasing functions of $d$ and $\sigma$. Therefore, the interest spread $H$ is also decreasing in $d$ and $\sigma$. The dependence in $T$ is more complex. According to the envelope theorem,

$$\frac{\partial}{\partial T}\left[\frac{D_0}{De^{-rT}}\right] = -\frac{\sigma}{2\sqrt{T}} N'(h_2) < 0.$$

Therefore, $HT$ is an increasing function of $T$. However, the dependence in $T$ is not clear-cut.

### 8.6.3 Using the CAPM for Loan Pricing

1. The gross expected return on a loan is

$$E(1 + \tilde{\rho}) = (1 - \delta)(1 + r_L) - \gamma_L.$$

The proposed pricing formula is thus equivalent to

$$E(\tilde{\rho}) - r = \alpha\pi.$$

The CAPM approach requires

$$E(\tilde{\rho}) - r = \beta\pi.$$

Hence, the two formulas are equivalent provided that

$$\beta = \alpha.$$

2. The monopolist bank's program is

$$\max_{r_L}[(1 - \delta)(1 + r_L) - \{\gamma_L + (1 - \alpha)(1 + r) + \alpha(1 + r + \pi)\}]L(r_L).$$

The first-order condition implies

$$\frac{(1 - \delta)(1 + (1 - 1/\epsilon_L)r_L) - \gamma_L - (1 - \alpha)(1 + r)}{\alpha} = 1 + r + \pi,$$

where

$$\epsilon_L = -\frac{r_L L'(r_L)}{L(r_L)}$$

is the elasticity of the demand for loans. (It is assumed that the second-order condition is satisfied.)

3. Simply replace $r + \pi$ by $(r + \pi)/(1 - \tau)$.

## Notes

1. In addition, *transformation risks* occur because banks' assets typically have a longer maturity than banks' liabilities; banks transform short deposits into long loans.

2. See, for example, Altman (1983) or, more specifically, the classic textbook of Hempel and Simonson (1991).

3. The recovery rate (proportion of nominal debt that is recovered by the lender when the borrower defaults) is also important. In this section, we make the simplifying assumption of a zero recovery rate.

4. The exponential discount factor $e^{-r}$ is used here because it simplifies the formulas.

5. For the definition, properties, and pricing formulas of a call option, see Ingersoll (1987) or Huang and Litzenberg (1988).

6. A standard Wiener process is a Gaussian process with continuous trajectories such that $E[Z(t)] = 0$, and for all $t$ and $s$, $E[Z(t)Z(s)] = \min(t, s)$.

7. See also Ingersoll (1987).

8. This ratio uses the book value of debt. The marked-to-market evaluation of the debt-to-asset ratio $(D_0/V)$ is endogenous and cannot be taken as a parameter.

9. Note that interpreting $sT$ as a global risk premium (because $e^{r_L T} - e^{rT} \sim sT$) is valid only when $T$ is small.

10. See, for example, Dewatripont and Tirole (1994) for a very clear account of these regulations.

11. Recall that this is equivalent to maximizing a constant absolute risk aversion (CARA) or exponential utility function when the returns are normally distributed.

12. For simplicity we adopt a continuous actualization. The yield to maturity with discrete actualization, $R(t, t+\tau)$, is defined by the more familiar formula

$$B(t, t+\tau) = \frac{1}{[1 + R(t, t+\tau)]^{\tau}}.$$

13. It also reflects implicitly the aggregate risk aversion of market participations through the use of the risk-adjusted probability $Q$.

14. See also Cox, Ingersoll, and Ross (1981; 1985) for a complete account of the different theories and models of the term structure.

15. The idea is to assimilate assets and liabilities to huge portfolios of coupon bonds and measure their sensitivities by using their durations (see, for example, Ingersoll, Skelton, and Weil 1978).

16. The negative sign in the definition of $\lambda$ is introduced for convenience; $\lambda$ should be positive.

17. However, liquidity risk is disregarded so as to emphasize the credit and interest rate risk the bank assumes.

18. Many countries have also adopted similar regulations.

## References

Altman, E. I. 1983. *Corporate financial distress: A complete guide to predicting, avoiding and dealing with bankruptcy risks.* New York: Wiley.

Baltensperger, E., and H. Milde. 1976. Predictability of reserve demand, information costs and portfolio behavior of commercial banks. *Journal of Finance* 31 (3): 835–843.

Bierman, H., and J. Hass. 1975. An analytic model of bond risk differentials. *Journal of Financial and Quantitative Analysis* 10 (5): 757–773.

Bizer, D., and P. DeMarzo. 1992. Sequential banking. *Journal of Political Economy* 100 (1): 41–61.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81 (3): 637–654.

Burgard, J. J. 1988. *La Banque en France.* Paris: Presses de la Fondation Nationale des Sciences Politiques.

Cox, J. C., J. E. Ingersoll, and S. A. Ross. 1981. A re-examination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 41 (3): 616–630.

———. 1985. A theory of the term structure of interest rates. *Econometrica* 53 (2): 385–407.

Décamps, J. P. 1996. Integrating the risk and term structure of interest rates. *European Journal of Finance* 2: 219–238.

Dewatripont, M., and J. Tirole. 1994. *The prudential regulation of banks.* Cambridge, Mass.: MIT Press.

Fabozzi, F., and A. Konishi. 1991. *Asset liability management*. Chicago: Probus.

Franks, J., and W. Torous. 1994. A comparison of financial recontracting in distressed exchanges and chapter 11 reorganizations. *Journal of Financial Economics* 35 (3): 349–370.

Geske, R. 1977. The valuation of corporation liabilities as compound options. *Journal of Financial and Quantitative Analysis* 12 (4): 541–552.

Gordy, M. B. 2003. A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation* 12 (3): 199–232.

Hart, O., and D. Jaffee. 1974. On the application of portfolio theory of depository financial intermediaries. *Review of Economic Studies* 41 (1): 129–147.

Hempel, G. H., and D. G. Simonson. 1991. *Bank financial management: Strategies and techniques for a changing industry.* New York: Wiley.

Ho, T., and A. Saunders. 1981. The determinants of bank interest margins: Theory and empirical evidence. *Journal of Financial and Quantitative Analysis* 16 (4): 581–600.

Ho, T., and H. Stoll. 1980. On dealer markets under competition. *Journal of Finance* 35 (2): 259–267.

Huang, C. F., and R. Litzenberger. 1988. *Foundations for financial economics.* Amsterdam: North-Holland.

Ingersoll, J. E. 1987. *Theory of financial decision making.* Totowa, N.J.: Rowan and Littlefield.

Ingersoll, J. E., J. Skelton, and R. L. Weil. 1978. Duration 40 years later. *Journal of Financial and Quantitative Analysis* 13 (4): 627–650.

Jappelli, T., and M. Pagano. 1993. Information sharing in credit markets. *Journal of Finance* 48 (5): 1693–1718.

Keeley, M., and F. Furlong. 1990. A reexamination of mean-variance analysis of bank capital regulation. *Journal of Banking and Finance* 14 (1): 69–84.

Kim, D., and A. M. Santomero. 1988. Risk in banking and capital regulation. *Journal of Finance* 43 (5): 1219–1233.

Klein, M. 1971. A theory of the banking firm. *Journal of Money, Credit and Banking* 3: 205–218.

Koehn, M., and A. M. Santomero. 1980. Regulation of bank capital and portfolio risk. *Journal of Finance* 35 (5): 1235–1244.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7 (1): 77–91.

Merton, R. 1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29 (2): 449–476.

Modigliani, F. R., and R. Sutch. 1966. Innovations in interest rate policy. *American Economic Review* 56: 178–197.

Prisman, E., M. Slovin, and M. Sushka. 1986. A general model of the banking firm under conditions of monopoly, uncertainty and recourse. *Journal of Monetary Economics* 17 (2): 293–304.

Pyle, D. 1971. On the theory of financial intermediation. *Journal of Finance* 26 (3): 737–747.

Rochet, J. C. 1992. Capital requirements and the behaviour of commercial banks. *European Economic Review* 36 (5): 1137–1170.

Schwartz, E. S., and W. N. Torous. 1992. Prepayment, default, and the valuation of mortgage pass-through securities. *Journal of Business* 65 (2): 221–239.

Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Smith, C., and J. Warner. 1979a. Bankruptcy, secured debt, and optimal capital structure: Comment. *Journal of Finance* 34 (1): 247–251.

———. 1979b. On financial contracting: An analysis of bond covenants. *Journal of Financial Economics* 7 (2): 117–161.

Stanhouse, B. 1986. Commercial bank portfolio behavior and endogenous uncertainty. *Journal of Finance* 41 (5): 1103–1114.

Vasicek, O. A. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5 (2): 177–188.

Yawitz, J. 1977. An analytical model of interest rate differentials and different default recoveries. *Journal of Financial and Quantitative Analysis* 12 (3): 481–490.

# 9 The Regulation of Banks

Banks are regulated in virtually every country with a well-developed banking system. This has an important effect both on the behavior of banks' managers and on the specific characteristics of the banking industry. In fact, it is practically impossible to study the theory of banking without referring to banking regulation. This is why the preceding chapters have already studied the effects of several aspects of bank regulation.

Bank regulation has a long history. The production of (private) money has always been taxed, the seigniorage or monopoly premium on coins being the property of the government. Contemporary banking regulation contemplates more complex problems because the set of regulatory instruments has become richer, and the regulators have set more ambitious macroeconomic and prudential objectives. Since the scope of this book is restricted to microeconomic banking theory, this discussion addresses exclusively issues related to the safety and soundness of the banking system. Traditionally, one distinguishes between the regulation of structure and the regulation of conduct. The former establishes which firms are qualified to develop a certain type of activity; the latter concerns the permitted behavior of firms in their chosen activities (see, e.g., Kay and Vickers 1988). Both are relevant for the study of banking regulation.

Safety and soundness regulatory instruments in use in the banking industry could be classified into six broad types:

- Deposit interest rate ceilings
- Entry, branching, network, and merger restrictions
- Portfolio restrictions, including reserve requirements
- Deposit insurance
- Capital requirements
- Regulatory monitoring and supervision (including closure policy)

Except for entry and merger restrictions, these regulatory instruments are specific to the banking industry. The absence of other, classical instruments could be explained by the constraints that limit regulatory actions.[1]

Banking regulation appears to involve diverse issues, all of them worth devoting effort to, but so heterogeneous that no model can encompass the main issues. Also, it is important to view this area as being in full evolution, where many issues remain unsolved (see Bhattacharya, Boot, and Thakor 1998 for an assessment).[2]

Section 9.1 discusses the justification for banking regulation, and in particular the existence of a Central Bank. Section 9.2 examines the question of banking regulation within the general framework of the theory of regulation. Then the discussion examines the six main regulatory instruments. Since the first three were addressed in chapter 3, the discussion here considers the three remaining points: deposit insurance (section 9.3), capital requirements (section 9.4), and bank closures (section 9.5). Finally, section 9.6 discusses market discipline.

## 9.1 The Justification for Banking Regulation

We could think of addressing banking regulation prima facie as an application of a general theory of public regulation to the specific problems of banking. But in fact this would be misleading. It is worth devoting some effort to understanding why banking regulation raises some questions that are not addressed within the general theory of public regulation. Although some instruments and models of the theory of regulation can be adapted to cope with issues in banking regulation, there are exceptions. The discussion examines the justification for regulation, considering first the general argument relating regulation and market failure, and then analyzing what is specific to banks: their inherent fragility and the fact that they borrow money from their customers. We examine the similarities and differences between the general theory of regulation and banking regulation from three perspectives: their justification, their scope, and the regulatory instruments.

### 9.1.1 The General Setting

In general, public regulation is justified by market failures that can come from market power, externalities, or asymmetric information between buyers and sellers.

In the context of financial intermediation, contemporary banking theory offers a series of explanations for the emergence of banks. These institutions arise to relax the problems due to asymmetric information (see chapter 2). Still, it is quite possible that banks do not completely solve the associated market failure or even that they create a new market failure. The Diamond-Dybvig framework allowed us to illustrate this point. Banks emerge because they provide liquidity insurance solving the market failure owing to the absence of contingent markets. Still, they create a new market failure because a bank run equilibrium exists, therefore requiring regulation.

From the point of view of efficiency, regulation results from a market failure in the banking industry, and therefore its justification relies on the identification of some market failure. The subsequent discussion establishes the existence of two market failures that make banking regulation necessary. However, once regulation is in place, politicians are often tempted to exploit it for their own benefit. First, the regulator is able to generate fiscal revenues by obliging banks to hold nonremunerated reserves. Second, by setting compulsory minimum ratios of investment, the regulator may channel credit to politically sensitive sectors, such as housing, exports, small businesses, or even less favored regions. Finally, some banking regulations may result from an effort to control other activities, for instance, money laundering.

In the next two sections we focus on the two main market failures of the banking industry: the fragility of banks because of their illiquid assets and liquid liabilities, and the fact that depositors are not armed to monitor the management of their bank.

### 9.1.2 The Fragility of Banks

The history of banking shows that bank panics are as old as the fractional reserve system. In other words, as soon as banks started to finance illiquid loans through demand deposits, most recessions were accompanied by losses of confidence by the public in the banking system, often leading to bank panics. Soon banks privately developed cooperative systems for protecting their collective reputation. These systems were later taken on and transformed by Central Banks when governments decided to impose control on the banking systems of most developed countries. However, several free banking episodes (such as in Scotland and the United States) showed that completely unregulated banking industries are conceivable.

For example, Calomiris (1993) compares the panics in the United States during the U.S. national banking era (1863–1913) and the banking collapse of the 1930s. He argues that during the national banking era, few banks actually failed, and panics were limited by temporary suspensions of convertibility during which bank notes circulated as a substitute for currency. This did not happen in the 1930s, which may explain the large number of bank failures that occurred then. Calomiris argues that the risk of runs can be reduced dramatically when banks are allowed to form large networks (as in the Scottish free banking era) and to enter into voluntary coinsurance and other cooperative arrangements with other banks (like the Suffolk system; see, e.g., Calomiris and Kahn 1996).[3]

Chapter 7 explained the role of banks in providing liquidity insurance to households, as modeled by the Bryant (1980) and Diamond and Dybvig (1983) models.[4] We also explained the unique role played by banks in screening and monitoring borrowers who cannot obtain direct finance from financial markets. As mentioned in chapter 7, it is the combination of these two functions that generates the fragility of banks. As Klausner and White (1993) argue, it is the nature of these core bank services to depositors and borrowers that explains the financial structure of banks (liquid liabilities and illiquid assets), which in turn explains the vulnerability of banks to runs.

An important argument in favor of this financial structure was expressed by Black (1975) and was reformulated by Fama (1985) as follows: "The ongoing history of a borrower as a depositor provides information that allows a bank to identify the risks of loans to depositors and to monitor the loans at lower cost than other lenders. . . . Two facts tend to support these arguments. First, banks usually require that borrowers maintain deposits (often called compensating balances). Second, banks are the dominant suppliers of short-term inside debt. The inside debt or private placements offered by insurance and finance companies (which do not have the monitoring information provided by ongoing deposit histories) are usually much longer-terms than bank loans" (38). Nakamura (1993) tests this conjecture on a large data set of U.S. banks and finds that, at least for small banks, scope economies exist between deposits and loans. However, this does not seem to be the case for large banks, which mainly lend to large firms. An explanation may be that large firms deal with multiple banks, which decreases the value of the information held by each of these banks.

### 9.1.3 The Protection of Depositors' and Customers' Confidence

In a free banking world, bank failures may be very costly, especially to the creditors of the failing bank (such as depositors, shareholders, and other banks) and, to a lesser extent, to borrowers who had previously developed a close relationship with the failing bank. Moreover, a bank failure may spread to other banks (interbank loans account for a significant proportion of banks' balance sheets) and similarly endanger the solvency of nonfinancial firms. Also, a bank failure may temporarily harm the payments system, since the finality of the payments managed by the failing bank just before its failure may be reconsidered. Therefore, the official justifications for banks' solvency regulations (given by regulators themselves), namely, protection of the public (essentially depositors) and safety of the payments system, appear prima facie quite reasonable.

However, two simple counterarguments can be found. First, there is no *qualitative* difference between the failure of a bank and that of a nonfinancial firm. All the negative externalities caused by bank failure are also present when a nonfinancial firm fails. Solvency regulations exist essentially for financial intermediaries and are absent in nonfinancial industries. Then, to quote the title of a famous article by Fama (1985), "What's Different about Banks?" Second, unless dishonest behavior is suspected on the part of banks' managers, they should not have any interest in provoking the failure of their own bank. How is it justified that the staff of the regulatory authority (who has a priori less competence, less inside information, and fewer incentives than a bank's manager) should decide on the solvency ratio of a commercial bank?

A partial answer to the first question has already been discussed. The distinguishing characteristic of banks (and more generally of financial intermediaries) is that their *creditors* are also their *customers*. In contrast to nonfinancial firms, whose debt is mostly held by professional investors (banks, venture capitalists, or informed private investors), the debt of banks (and insurance companies) is held in large part by uninformed, dispersed, small agents (mostly households) who are not in a position to monitor the banks' activities. It is true that large corporations are also financed by the public; stocks and bonds issued by large companies are indeed widely diffused. However, there are two differences: *these securities are not used as a means of payment* (which moderates the free-rider problem involved in bank monitoring), and *the debt-to-asset ratio is substantially higher* for financial intermediaries than for nonfinancial firms. Therefore, the free-rider problem involved in the monitoring of widely held firms seems to be quantitatively much more serious in the case of banks and insurance companies.

As for the second question (Why shouldn't banks' managers themselves choose the optimal solvency ratio?), the answer is given by the important observation (initially from Jensen and Meckling 1976) that there are *conflicts of interest* inside firms, between managers, stockholders, and bondholders. Consider, for instance, the case of a bank whose capital is held by a small number of equity holders (insiders) who manage the bank themselves. As shown by Jensen and Meckling (1976), these owner-managers will tend to choose an investment policy that is more risky than depositors would like. Since these depositors are not in a position to control the banks' activities (or to bargain with the owners), their interests must be defended by some institution (leading to another form of delegated monitoring than in section 2.4). This institution can be either a public regulator that is given the mission of maximizing the utility of depositors, or a deposit insurance company whose objective is to minimize the expected costs of insuring depositors.

Another important case of conflict of interest is that of a large bank whose capital is widely held. In that case, the most important conflict is between the bank's managers and the outside financiers (depositors and stockholders). It is more difficult then to understand why the financial structure (debt-to-asset ratio) of the bank matters, since it is a priori unrelated to the relevant dimension, namely managerial incentives.

This puzzle can be solved by introducing the incomplete contract paradigm (see section 4.5). If no contract can be written (and enforced) that specifies the actions of the manager, the only way to discipline this manager is to threaten him with external intervention.

Dewatripont and Tirole (1993) have developed a general theory of corporate structure along these lines. Their most spectacular result is that debt and equity are precisely the adequate instruments for inducing optimal managerial performance. The intuition for their result is that debt and equity generate a separation of tastes and tasks among financiers. Indeed, equity gives a payoff function that is convex with respect to the liquidation value of the firm (because of limited liability). Therefore, equity holders tend to favor risky decisions by the manager. It is thus appropriate to give them the control rights when the firm performs well. On the contrary, debtholders have a concave payoff function; they tend to be more risk-averse. It is thus appropriate to give them the control rights when the firm performs badly (bankruptcy). This model can be adapted to the framework of bank regulation, as in Dewatripont and Tirole (1994) (see section 9.4.3).

### 9.1.4 The Cost of Bank Failures

The preceding discussion points to three characteristics of banks that make their failure generate a strong negative externality on other economic agents.

First, banks emerge to solve an asymmetric information problem or a market imperfection. A bank's failure may therefore leave its customers facing the market imperfection the bank was supposed to solve. A defaulting bank's potential borrowers will be left without credit and will have to prove their creditworthiness to another bank. All investment previously made by the defaulting bank in screening and monitoring will be lost, as will the investment in a relationship (see Slovin et al. 1999 for the cost of the Continental Illinois failure).

Second, financial fragility generates financial contagion, as clearly illustrated by the models presented in chapter 7. Because a bank failure may signal a weakness in bank assets, it may cause depositors to question the solvency of all the other banks. So a bank failure may produce a perfectly rational Bayesian updating of the assessment of any other bank risk and a generalized withdrawal of deposits. The precise way in which contagion develops, because of a change in depositors' expectations or banks' financial interdependence resulting from the net of reciprocal claims that are generated, is immaterial. The illiquid assets and liquid liabilities characteristics of banks' financial fragility foster contagion.

Finally, the cost of default inflicted upon small depositors may be politically inacceptable.

## 9.2 A Framework for Regulatory Analysis

We first consider a general framework for the analysis of regulation, as inherited from industrial organization theory (fig. 9.1). Absent this preliminary work we "may

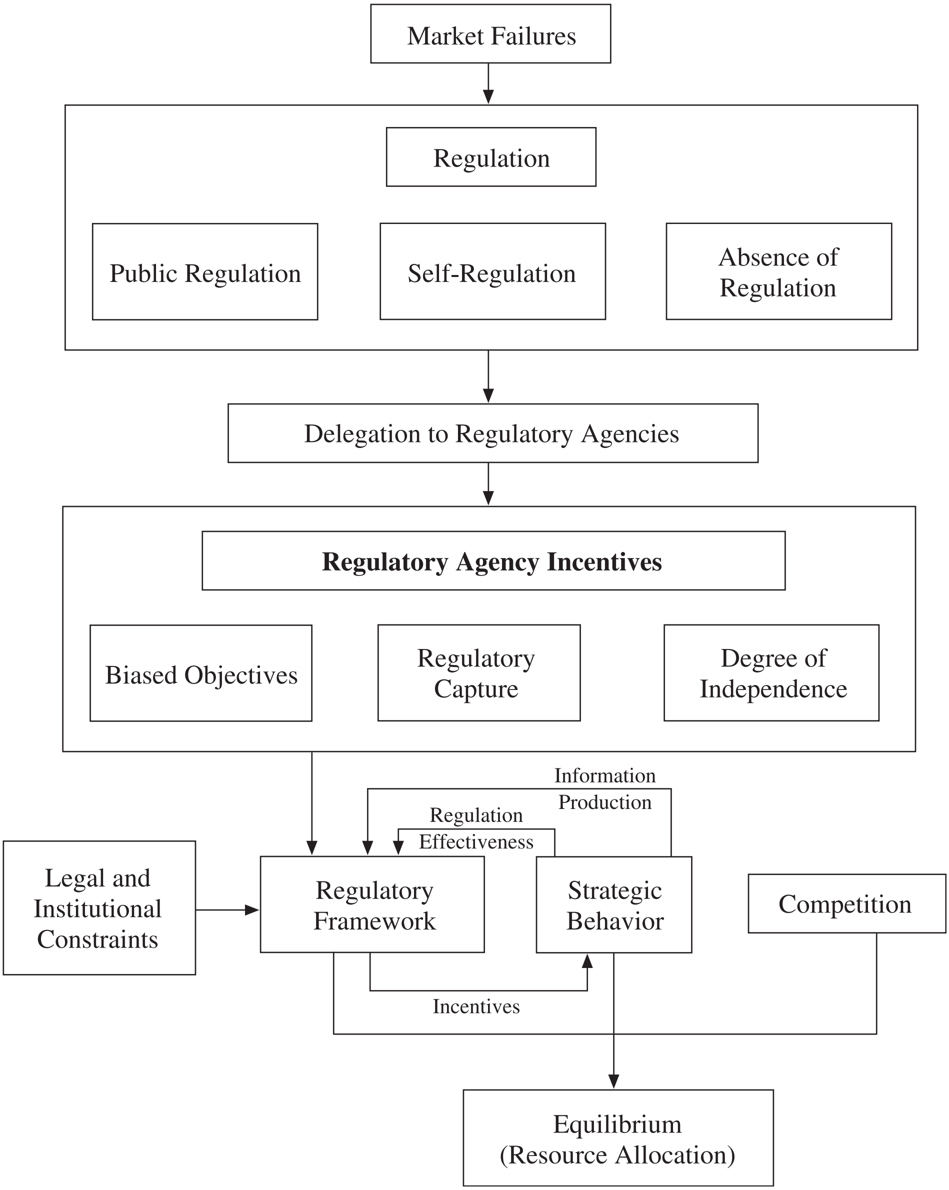



**Figure 9.1**
Banking regulation in perspective.

fall prey to a naive view of the world," as Freixas and Santomero (2001) state, "[where] powerful regulators act in the best interest of society and the regulated banks will submissively abide by the regulation." Instead, regulators may deviate from their ideal objective function, and banks may react strategically.

Regulation may be implemented by one or several regulatory agencies with different responsibilities,[5] which complement each other but sometimes compete with each other. Some of these agencies may have a mandate of implementing regulation on behalf of the public interest, while others may be in charge of the long-run interests of the industry (self-regulation). Nevertheless, as in any principal-agent framework, the agency may have biased objectives. This is a well-known characteristic of regulation, as acknowledged by the tendency of regulators to act in the interest of the government (lack of independence of Central Banks) or in the interest of the industry (as in the case of regulatory capture).

The regulatory agency, acting within the framework of current legislation, will enact a series of regulatory rules that constitute the regulatory framework. When confronted with these rules, banks will adapt their strategy both in terms of the actions (observable and nonobservable) they undertake and the information they supply to the regulator. Competition among banks will then take place within this framework, yielding an equilibrium for the banking industry characterized by level of risk, monitoring, and so on.

General regulation theory is concerned with the design of the optimal regulatory rules.[6] It is therefore mainly *normative*. However, only a minor part of the literature on banking regulation follows this regulation design approach. The main strand takes a *positive* approach: regulation analysis. Its aim is to analyze the consequences of a given regulation that either exists or is under study by the regulatory authorities. Regarding, for instance, capital adequacy requirements, those following this approach would ask questions such as, Will this regulation succeed in attaining its objectives (reduction of the bank failure risk)? Will it induce more risk taking by banks? Will it change the equilibrium rates in the credit market?

There are two lessons to be drawn from this abstract exercise.

First, banking regulation is costly, both directly (salaries of supervisors, administrative costs for banks) and indirectly (through the distortions it generates). Also, it may generate rents for banks. Therefore, if regulators are self-interested (see Boot and Thakor 1993), they may be captured by the banking industry. For all these reasons, advocates of free banking prefer an imperfectly competitive market to an imperfectly regulated banking sector.

Second, the level and characteristics of regulation are to be derived in equilibrium, so that regulatory rules are related to each other. Thus, for instance, the optimal level of regulatory capital depends on the bankruptcy rules that apply to the banking sector and the probability of bankruptcy spreading to other banks.

## 9.3 Deposit Insurance

To avoid bank panics and their social costs, governments have established deposit insurance schemes.[7] Under such schemes each bank pays a premium to a deposit insurance fund, such as the Federal Deposit Insurance Corporation (FDIC) in the United States. In exchange depositors have their deposits insured up to a fixed limit in case their bank fails.

In the United States deposit insurance mechanisms were developed by the Fed as a response to the Great Depression bank panics. They were later adopted by most developed countries with different modalities. Insurance may be compulsory or voluntary, it may be implemented by one or by several funds, it may cover only principal or principal plus interest, and the limits may differ widely (from \$100,000 in the United States to \$20,000 in Europe). Before implementing explicit deposit insurance systems, some European countries had implicit deposit insurance systems based on direct government intervention to pay depositors, sharing the losses with the country's other main banks ("survivors pay" principle).

In most cases, deposit insurance schemes are public, although some economists have advocated recourse to private insurance systems. Such systems have in fact been reintroduced in some U.S. states with mixed success (see Mishkin 1992). The potential advantage of private systems is that competition provides incentives for information extraction and accurate pricing. There are also important drawbacks. Because of systemic risks, private insurance systems lack credibility unless they are backed by the government, which in turn casts doubt on the incentives of deposit insurance funds to look for an accurate pricing of deposit insurance. Also, since Central Bank interventions and closures of commercial banks are public decisions, private insurance schemes can function only if the government establishes explicit contingent closure policies, which is a difficult task (see Benston et al. 1986).

Chapter 7 showed how deposit insurance could provide a solution to bank runs (Diamond and Dybvig 1983). This section examines several other aspects of deposit insurance: the moral hazard issue (section 9.4.1), risk-based pricing (section 9.4.2), and incomplete information problems (section 9.4.3).

### 9.3.1 The Moral Hazard Issue

Before developing the well-known arguments related to the moral hazard consequences of deposit insurance, this section briefly describes the simple model used here. It is a static model with only two dates. At $t = 0$ the deposit insurance premium is paid by the bank. At $t = 1$ the bank is liquidated, and depositors are compensated whenever the bank's assets are insufficient. For simplicity, the riskless rate and the deposit rate are normalized to zero. The balance sheets of the bank are as follows:

| Assets | Liabilities |
|---|---|
| Loans $L$ | Deposits $D$ |
| Insurance premiums $P$ | Equity $E$ |
| Loan repayments $L$ | Deposits $D$ |
| Insurance payments $S$ | Liquidation value $V$ |
| $t = 0$ | $t = 1$ |

At date 1 the stockholders share the liquidation value of the bank:

$$\tilde{V} = \tilde{L} - D + \tilde{S}, \tag{9.1}$$

where $\tilde{S}$ is the payment received from deposit insurance:

$$\tilde{S} = \max(0, D - \tilde{L}). \tag{9.2}$$

Using the balance sheet at date 0, we can express $D$ as $L + P - E$, and $\tilde{V}$ can also been written as

$$\tilde{V} = E + (\tilde{L} - L) + (\max(0, D - \tilde{L}) - P). \tag{9.3}$$

Thus the shareholders' value of the bank equals the sum of its initial value, the increase in the value of loans, and the net subsidy (positive or negative) received from the deposit insurance.

Suppose, for instance, that $\tilde{L}$ can take only two values: $X$ with probability $\theta$ (success) and 0 with probability $(1 - \theta)$ (failure). The expected gain for the bank's shareholders is

$$\Pi \stackrel{\text{def}}{=} E(\tilde{V}) - E = (\theta X - L) + ((1 - \theta)D - P), \tag{9.4}$$

where the first term represents the net present value of the loans and the second term is the net subsidy from the deposit insurance system. If deposit insurance is fairly priced, this term is nil ($P = (1 - \theta)D$), and the strong form of the Modigliani-Miller result obtains: the total value of the bank, $E(\tilde{V}) + D$, is independent of its liability structure.[8]

The moral hazard problem is easily captured from formula (9.4). Suppose that $P$ is fixed and that banks are free to determine the characteristics $(\theta, X)$ of the projects they finance in a given feasible set. Then, within a class of projects with the same net present value (NPV) ($\theta X - L =$ constant), the banks will choose those with the lowest probability of success $\theta$ (or the highest risk). This comes from the fact that the premium rate $P/D$ is given and does not depend on the risk taken by the bank. Such flat-rate deposit insurance pricing was in place in the United States until December 1991, when the U.S. Congress legislated a new system involving risk-related insurance premiums. The following section shows how such pricing rules can be designed theoretically, and whether they provide a solution to the moral hazard problem.

### 9.3.2 Risk-Related Insurance Premiums

As can be seen from formula (9.2), the deposit insurance payment $\tilde{S}$ is identical to a put option on a bank's assets $\tilde{L}$ at a strike price $D$.[9] This was originally observed by Merton (1977), who proposed using the arbitrage pricing method for finding the appropriate pricing policy for deposit insurance. This method requires the existence of complete (and perfect) financial markets, on which the deposit insurance (or option) contract can be duplicated by a portfolio of tradable securities. In the absence of arbitrage opportunities, the price of such a contract can be computed as its expected NPV under some risk-adjusted or martingale probability measure (which incorporates market corrections for risk). Suppose, for instance, that the value of the bank's assets at date $t$ follows a geometric random walk,

$$\frac{d\tilde{L}}{\tilde{L}} = \mu\, dt + \sigma\, dZ, \tag{9.5}$$

where $Z(t)$ is a standard Brownian motion. If the riskless rate $r$ and deposit rate $r_D$ (measured in continuous terms) are constant, and if $T$ denotes the time between two examination dates, the Black-Scholes formula (1973) applies and the no-arbitrage (or actuarial) price of deposit insurance is given by

$$P^* = De^{(r_D - r)T} N(h_2) - LN(h_1), \tag{9.6}$$

where $N(\cdot)$ is the standard Gaussian cumulative distribution function (c.d.f.) and

$$h_1 = \frac{1}{\sigma\sqrt{T}} \log \frac{De^{(r_D - r)T}}{L} - \frac{1}{2}\sigma\sqrt{T}, \tag{9.7}$$

$$h_2 = h_1 + \sigma\sqrt{T}. \tag{9.8}$$

Homogeneity of these formulas allows for focusing on the (actuarial) premium rate $P^*/D$ as a function of the deposit-to-asset ratio $D/L$ and the volatility of assets $\sigma$. Classical properties of the Black-Scholes formula imply the following (unsurprising) result.

**Result 9.1** The actuarial rate $P^*/D$ of deposit insurance is an increasing function of the deposit-to-asset ratio $D/L$ and of the volatility $\sigma$ of the bank's assets (Merton 1977).

Marcus and Shaked (1984) and Ronn and Verma (1986) tried to estimate the difference between these theoretical premiums $P^*$ and those actually paid by U.S.

banks, in an attempt to evaluate the importance of implicit subsidies to the banking industry. Buser, Chen, and Kane (1981) argue that the implicit subsidy allows estimating the value of banks' charters. On the other hand, result 9.1 has been extended to take into account audit costs (Merton 1978), liquidation costs (Mullins and Pyle 1991), and interest rate risk (McCullough 1981; Kerfriden and Rochet 1993).

Another extension of Merton's option pricing model for pricing deposit insurance is put forth by Pennacchi (1987), who analyzes the effect on bank insurance pricing and bank failure resolution. In particular, Pennacchi contrasts the consequences of a purchase and assumption transaction with a policy of making direct payments to depositors. In the latter case, even if deposit insurance is fairly priced, banks will tend to take excessive risks. In the former case, however, sufficient monopoly rents (charter values) would induce banks to prefer to increase their capital.

If the authorities can close the bank *before* it is actually insolvent, deposit insurance becomes analogous to a *callable put option*, and dynamic considerations have to be introduced. Acharya and Dreyfus (1988) develop a model along these lines. At each date banking authorities receive a report $X$ on the value of the bank's assets. Using this information, the optimal closure policy is determined (simultaneously with the price of deposit insurance) as the minimum cost policy for the deposit insurer. Acharya and Dreyfus show that the insurer will optimally close the bank whenever

- the net increase in the insurer (discounted) liability exceeds the immediate cost of reorganizing the bank; or
- the bank's current asset value is too low for the insurer to be able to charge an actuarially fair premium.

In a competitive banking industry, bank closure will always happen with a positive probability. Thus Acharya and Dreyfus's results imply a more involved formula for the market value of deposit insurance. They do not alter the nature of the result. If the deposit insurance company is able to observe perfectly the bank's risk characteristics ($D/V$ and $\sigma$), then it is theoretically possible to price deposit insurance in an actuarially fair way. However, this is more complex under asymmetric information.

### 9.3.3 Is Fairly Priced Deposit Insurance Possible?

The title of this section is taken from an article by Chan, Greenbaum, and Thakor (1992), who show that when asymmetric information is present, fairly priced deposit insurance may not be feasible. A first issue is timing. Even if the portfolio decisions of banks are perfectly observable, there is a time lag between these decisions and the subsequent premium adjustments by the regulator or the insurer. Therefore, if the bank is seriously undercapitalized, its managers may decide to gamble for resurrection during this time lag even if they know that later they may have to pay for it. Also, increasing insurance premiums may increase this incentive to gamble for resurrection because the bank's stockholders know that they will not be liable in case the bank fails.

A second issue, examined in more detail by Chan, Greenbaum, and Thakor, is adverse selection. Consider the simple model developed in section 9.3.1 and suppose that $\theta$, the probability of repayment of the bank's loan, is private information of the bank. Fairly priced deposit insurance can nevertheless be possible if there exists a (nonlinear) premium schedule $P(D)$ such that premiums equal expected losses,

$$P[D(\theta)] = (1 - \theta)D(\theta),$$

where $D(\theta)$ is the profit-maximizing level of deposits for a bank of characteristic $\theta$. Namely, $D(\theta)$ realizes $Max_D\, \Pi(D, \theta)$, where by definition

$$\Pi(D, \theta) = (\theta X - L) + (1 - \theta)D - P[D(\theta)].$$

The first-order condition of this problem is

$$\frac{\partial \Pi}{\partial D}(D(\theta), \theta) = 0 = (1 - \theta) - P'[D(\theta)].$$

Differentiating the fair pricing condition yields

$$P'[D(\theta)]D'(\theta) = (1 - \theta)D'(\theta) - D(\theta).$$

Multiplying the first equation by $D'(\theta)$ and comparing it to the second equation gives $D(\theta) \equiv 0$, which is, of course, absurd. Chan, Greenbaum, and Thakor conclude that fairly priced deposit insurance is not viable because of asymmetric information. Freixas and Rochet (1995) show that, in a more general case, fairly priced deposit insurance may in fact be viable under asymmetric information, but that it will never be completely desirable from a general welfare viewpoint. The reason is that cross-subsidies between banks are Pareto-improving in an adverse selection context. However, these cross-subsidies may also lead to an artificial survival of inefficient banks, thus generating a trade-off between static and dynamic efficiency.

Bond and Crocker (1993) study the consequences of linking deposit insurance premiums to the capitalization of banks, in an interesting model based on the costly state verification paradigm of Townsend (1979) and Gale and Hellwig (1985). In this model, inspired also by Diamond (1984), banks attract the funds of risk-averse depositors and invest them in industrial projects. These are small banks, managed by their owners (the bankers) and prevented from diversification by organization costs (as in Cerasi and Daltung 2000). The return $\tilde{x}$ on a bank portfolio is observable only by its manager, except if depositors pay an audit cost. The optimal deposit contract is therefore a standard debt contract (transposed to the deposit side). Depositors receive $\min(\tilde{x}, R)$, where $R$ is the nominal rate of deposits, and pay the audit cost when $\tilde{x} < R$. Bond and Crocker start by analyzing the competitive equilibrium of the banking sector in the absence of deposit insurance. Banks determine the capital level $K^*$ and the deposit rate $R^*$ that maximize the depositors' expected utility under the constraint that banks break even. Banking capital is useful in this context, because it provides partial insurance to risk-averse depositors against fluctuations in banks' portfolio returns. Bond and Crocker then show that introducing actuarial deposit insurance provides banks with an additional tool for insuring depositors. They find that complete deposit insurance would be suboptimal in this context because it suppresses the incentive of depositors to require that banks self-protect through capitalization. Finally, Bond and Crocker study the optimal deposit insurance plan, in which insurance premiums paid by banks depend on the banks' capitalization.

### 9.3.4 The Effects of Deposit Insurance on the Banking Industry

The positive approach to banking regulation is concerned with the effects of regulation on equilibrium in the deposit and credit markets. The complexities of this approach stem from the fact that for regulation to be justified (for free banking not to be optimal), an imperfection of capital markets must be introduced. Since there is no general consensus in the literature on which imperfection is the crucial one, there are multiple models, which also differ as to the way regulation is introduced. This discussion therefore does not try to survey the different approaches to modeling the effects of banking regulation but instead focuses on some contributions that illustrate the main types of results that can be obtained.

The simplest way to model the effects of regulation is to disregard the imperfections of the financial markets. This approach has been used to analyze the effect of flat-rate deposit insurance, that is, insurance for which the premium is just proportional to the volume of deposits and thus does not depend on the level of risk of the bank's assets. As mentioned in section 9.3.1, flat-rate deposit insurance gives banks an incentive to take too much risk. But the consequences on the equilibrium level of deposit and loan margins are not obvious. The analysis of these effects has been the subject of two articles by Suarez (1993a; 1993b). Assuming risk neutrality and limited liability of banks, Suarez shows that the banks' portfolio problem has solutions of the bang-bang type, which are quite intuitive: high margins on deposits will lead the banks to assume a lower risk. Still, even if the margin on loans is negative, the banks may be interested in lending (provided they have sufficient leverage through deposits), simply because they obtain a subsidy via deposit insurance. Similar bang-bang results could be obtained in a dynamic equilibrium in which the banks' level of risk taking has an effect on the probability of bankruptcy (the present value of future profits forgone) but increases the value of the banks' claim on the deposit insurance company. If future profits are low, the banks will choose to take a maximum amount of risk; if instead the banks have some market power, they will take less risk (Suarez 1993b).

Gennote and Pyle (1991) also consider the effect of deposit guarantees, although they focus on the banks' portfolio of loans. They show that deposit guarantees will lead to inefficient investment and that increases in bank capital requirements could not compensate for the increase in risk (see section 9.4.1 for other results contained in the same article).

Among the few contributions that explicitly introduce a capital market imperfection that makes free banking inefficient are two articles by Matutes and Vives (1996; 2000). In the first, they use Hotelling's (1929) model of horizontal differentiation and obtain situations of market failure, in which free banking is not viable. In this context, they show that deposit insurance is desirable because it prevents market collapse and allows an increase in the market size by restoring the confidence of depositors. However, as a result of deposit insurance, the banks will compete more fiercely, which increases the expected cost of failure. In Matutes and Vives (2000), a model of imperfect competition in the presence of a social cost of failure is considered, and flat-rate deposit insurance regulation results in excessive risk taking. One of the implications of this model is that deposit rate regulation may be desirable.

## 9.4 Solvency Regulations

### 9.4.1 The Portfolio Approach

The portfolio approach, developed originally by Kahane (1977) and Kareken and Wallace (1978), and examined later by Crouhy and Galai (1986), Kim and Santomero (1988), and Koehn and Santomero (1980), is parallel to the literature presented in section 8.4. The main idea is that if banks behave as portfolio managers when they choose the composition of their portfolio of assets and liabilities, then it is important to use risk-related weights for the computation of the capital-to-asset ratio. Like Crouhy and Galai, Kareken and Wallace use a complete markets framework and show that in that context, capital regulations are dominated by risk-related insurance premiums as an instrument for solving the moral hazard problem. Of course, as has been repeatedly argued in this book, the complete market setting is not really appropriate for modeling banks.

As a proxy for incomplete markets, Kim and Santomero introduce risk aversion in the bank's objective function.[10] This is legitimate in the case of a small bank, owned and managed by the same agent, who cannot completely diversify the risk. Using a mean-variance model, Kim and Santomero compare the bank's portfolio choice before and after a solvency regulation is imposed. They show that in general, the solvency regulation will entail a recomposition of the risky part of the bank's portfolio in such a way that its risk is increased. As a consequence, even if the global size of this risky portfolio decreases (because of the solvency regulation), the probability of the bank's failure may *increase* after the solvency regulation has been imposed, which is rather ironic (this has already been pointed out by Kahane). Kim and Santomero, and later Rochet (1992a), show that this distortion in the banks' asset allocation disappears when regulators use correct (market-based) measures of risk in the computation of the solvency ratio.

Keeley and Furlong (1990) and Rochet criticize the Kim-Santomero approach for inconsistency, in the sense that the limited liability option is not introduced in the bank's objective function. Rochet shows that when this option is properly accounted for, the efficiency of solvency regulations is jeopardized even more. Even when market-based risk weights are used, it may be necessary to require an additional regulation in the form of an additional minimum capital requirement for banks (in absolute terms), independent of their size.

Gennotte and Pyle (1991) revisit the analysis of the effect of capital regulations on bank risk by assuming that banks can invest in projects that have a positive NPV. As mentioned, if all banks' assets have zero present value (because, for instance, these assets are traded on perfect capital markets), not only do banks have no social value but also the only reason their portfolio decisions may be relevant is that they enjoy rents (which may come from underpriced deposit insurance or from imposed restrictions on competition). This is a caricatural vision of banking. On the other hand, in Gennotte and Pyle, banks have a social utility because they screen and monitor industrial projects that could not be directly financed by capital markets. By investing $v$ in a project of risk characteristic $\sigma$, a bank generates an NPV, denoted $J(\sigma, v)$. If banks could completely self-finance, they would optimally choose the size $v^*$ and the risk characteristic $\sigma^*$ that jointly maximize $J(\sigma, v)$. Since in fact they finance themselves in part by attracting insured deposits $D$, they benefit from the option value associated with limited liability. Therefore, the bank's objective function is distorted in the direction of excessive risk taking.

Finally, it is worth mentioning that the macroeconomic implications of solvency regulation have begun to be explored (see, e.g., Blum and Hellwig 1995; Bolton and Freixas 2006).

### 9.4.2 Cost of Bank Capital and Deposit Rate Regulation

Hellman, Murdock, and Stiglitz (2000) develop an argument in favor of deposit rate regulation, which we briefly sketch in a static framework. In their model, deregulation and competition lead to investment in inefficient risky projects, whereas deposit rate regulation, by providing a sufficiently high charter value, gives banks incentives to invest in safe efficient projects, thus guaranteeing the stability of the banking industry. Their argument is based on the effect of a bank's charter value as a counterbalancing force to the risk-taking incentives generated by deposit insurance. They point out that if the cost of capital $\beta$ is particularly high, then an increase in capital requirements could erode the bank charter value.

Consider a risk-neutral, zero-interest-rate economy where two banks $(i = 1, 2)$ compete. Banks are financed by insured deposits in limited supply and by capital.

Banks are able to invest in two types of projects, a safe project that returns $G$ with probability $p_G$ and zero otherwise, and a risky project that returns $B$ with probability $p_B$ and zero in case of failure. We assume, as we have done before, that $p_G G > p_B B$ and that $B > G$, so that $G$ is the efficient investment choice. Still, we assume here that $B$ is a positive net present value project, so that the credit market does not collapse if in equilibrium the $B$ project is chosen. The banks' choice of project is unobservable, so that the safe project will only be chosen if it yields higher profits.

The cost of capital $\beta$ is assumed to be high enough because it satisfies $\beta > p_G G$.

Let $k$ be the amount of capital per unit of deposit, $r_D^i$ the interest rate paid on deposits, and $D_i(r_D^i, r_D^{-i})$ the amount of insured deposits, where $-i$ stands for $i$'s competitor. The bank's assets will then be $(1 + k)D_i(r^i, r_D^{-i})$.

A bank will prefer to invest in the safe project if the profits it generates per unit of deposit are larger, that is:

$$p_G[G(1+k) - r_D] - k\beta \geq p_B[B(1+k) - r_D] - k\beta. \tag{9.9}$$

After simplification, this condition sets a cap on the interest rate $r_D$:

$$r_D \leq \widehat{r(k)} \equiv \frac{p_G G - p_B B}{p_G - p_B}(1+k). \tag{9.10}$$

The deposit insurance premium has no effect on the bank's project choice provided it is set before the bank's decision (or set as flat-rate deposit insurance). In the same way, the fact that deposits are insured could be dispensed with provided that deposit rates do not depend upon the project choice, as in Matutes and Vives (1996). If deposit rates reflect the riskiness of the project, the market is perfect, with $p_G r_D = 1$ or $p_B r_D = 1$, and the Modigliani-Miller theorem implies that project $G$ will always be chosen.

**Market Equilibrium**

To begin with, the discussion focuses on the equilibrium allocation for an unregulated market. It is characterized by the bank's choices regarding $k$ and the type of project as well as the equilibrium deposit rates. Regarding the choice of $k$, the inequality $\beta > p_G G$ implies that the optimal choice for the bank is to set the capital level to its minimum level, which in an unregulated setting is $k = 0$. Regarding the choice of project, it is characterized by the implementation of the $B$ project.

To see this, notice that Bertrand competition and constant returns to scale imply that in equilibrium banks' profits are zero. As a consequence, the safe project will never be implemented in an unregulated equilibrium, because if $r_D = p_G G$, it is profitable for a bank to implement the risky inefficient project and attract all depositors by setting a higher interest rate. Consequently, the unregulated equilibrium leads to a zero capital level for the banking industry and the choice of the risky inefficient project.

**Optimal Regulation**

In order to decrease banks' incentives to gamble, it seems natural to impose minimum capital requirements. Still, the regulator has at its disposal another regulatory instrument, deposit rate regulation, so that the use of capital requirements should be compared with setting caps on deposit remunerations.

Expression (9.10) establishes for each level of capital $k$ the maximum deposit rate $\widehat{r(k)}$ that leads the bank to invest in the safe project. When there is a charter value, $V$, for the continuation of the bank's activity, expression (9.10) is simply modified by adding a term $(p_G - p_B)V$ to the left-hand side of (9.10):

$$p_G[G(1+k) - r_D] + (p_G - p_B)V(k, r_D) \geq p_B[B(1+k) - r_D].$$

The bank's charter value will depend upon the equilibrium interest rates on deposits and on capital requirements. We have that

$$\frac{\partial V}{\partial k} \leq 0 \quad \text{and} \quad \frac{\partial V}{\partial r_D} \leq 0$$

because higher levels of both $k$ and $r_D$ diminish future profits.

Taking into account the effect of the charter value changes the shape of the $\widehat{r(k)}$ frontier. Computing

$$\frac{d\hat{r}}{dk}$$

leads to

$$\frac{d\hat{r}}{dk} = \frac{(p_G G - p_B B) + (p_G - p_B)(\partial V/\partial k)}{(p_G - p_B)(1 - \partial V/\partial r)}.$$

Consequently, for high negative values of

$$\frac{\partial V}{\partial k},$$

we find that

$$\frac{d\hat{r}}{dk}$$

becomes decreasing. When this is the case, a higher level of capital is associated with lower ceilings on deposit rates, so that it not only results in an increase of the bank's costs of capital but also leads to additional distortions in the deposit market.

### 9.4.3 The Incentive Approach

It is natural to assume that banks have better information regarding their own risks and returns than the regulator does. Modeling this issue with techniques similar to those developed by Laffont and Tirole (1986; 1993) has led to a new approach to solvency regulations, initiated by Giammarino, Lewis, and Sappington (1993), Rochet (1992b), Bensaid, Pagès, and Rochet (1993), and Freixas and Gabillon (1999).

In this approach, solvency regulations are modeled as a principal agent problem between a public insurance system (operated, say, by the Central Bank) and a private bank. The latter is run by managers who carry out risky projects (loans) and invest in safe assets (reserves). Both activities are financed by cash collected from depositors and by capital raised among outside shareholders. When there is no conflict of interest between shareholders and the managers of the bank, the regulator simply attempts to minimize the expected loss of deposit insurance under the individual rationality constraint of both managers and shareholders. Since insurance is costly (overhead, deadweight loss of taxation) the cost of public funds will in the end determine the optimal trade-off faced by the regulator between the cost of banking capital and that of insuring depositors. The main results obtained by Giammarino, Lewis, and Sappington and by Bensaid, Pagès, and Rochet are as follows.

- Functional separation between deposit and loan activities is in general inefficient. In this setup, the proper allocation of capital is consistent with some risk transformation, and this always implies some positive probability of failure. In this respect, free and narrow banking appear as special cases of optimal regulation when the cost of public funds is respectively zero or infinite.
- The optimal incentive scheme may be decentralized through a solvency requirement that induces banks to internalize the cost of the deposit insurance system. The appropriate regulation imposes the capital-to-asset ratio which, at the margin, leaves unaffected the expected cost of deposit insurance.
- Efficient regulation should be risk-adjusted. Under the 1988 international Basel agreement, assets are essentially risk-weighted according to the institutional nature of the borrower.[11] Here another dimension of risk adjustment is stressed, according

to the size of the bank's portfolio. The risk brought about by a marginal increase in loans of any credit category should be larger than the average risk of that category. Hence, the marginal capital-to-asset ratio is set above the average ratio. This may be achieved through a system of lump-sum deductibles, under which equity is not strictly proportional to the assets outstanding.

• Finally, the capital-to-asset ratio should be contingent on the quality of banks' assets, measured, for instance, by ratings performed by independent agencies.

The social costs of an insured failure is a matter of concern that could justify regulation of the banks' capital ratios. Still, when deposits have a utility, or when capital is costly, the trade-off between capital requirements and the cost of bank failure must be considered. In a perfect information setting, this could be determined simply as a marginal condition. But this issue becomes particularly relevant in an adverse selection setting in which mechanism design theory can be used to obtain the optimal regulatory scheme. Freixas and Gabillon (1999) consider mechanisms that combine the amount of risk-free assets (reserves) that banks are bound to hold, the amount of capital they are required to have, and the deposit insurance premium they have to pay. The banks are assumed to have private information on the initial value of their portfolio of loans, but this value follows the Merton formula for pricing the claims on the deposit insurance company. Using this framework, Freixas and Gabillon characterize the optimal mechanism, that is, the one that maximizes social surplus when deposits and loans have a social value, and when there is a social cost of bankruptcy, constraining the mechanisms to be incentive-compatible and respecting an individual rationality constraint for the banks. The result they obtain is that if loans have a positive NPV, banks will never hold reserves, and the deposit insurance premium will have to be decreasing with the bank's capital.

### 9.4.4 The Incomplete Contract Approach

It is clear that most large modern banks are owned by many small investors. The previous approach, which assumed that banks were owned and managed by the same agent (the banker) does not fit this empirical evidence. In reality, bank managers own (at most) a small fraction of their banks' capital. Therefore, it may be reasonable to concentrate on the incentive scheme of these managers rather than on that of the stockholders. It is then more difficult to understand why banks' solvency (the financial structure of banks) matters, since there is no obvious relation between this financial structure and the performance of managers. In particular, if complete contracts can be written between the owners of a bank and its managers, the Modigliani-Miller theorem applies, and financial structure is irrelevant.

Therefore, the only possibility for reintroducing the relevance of banks' solvency in this context is to consider that contracts are incomplete in the sense that some decisions cannot be prespecified. Thus the allocation of the control rights on the bank becomes important. It is the financial structure of the firm (here the solvency regulation of the bank) that determines the allocation of these control rights among claim holders, and in particular when and how these claim holders can intervene in management. The approach that is followed now (taken from Dewatripont and Tirole 1994) is an application to the banking sector of a general theory of the financial structure of firms (also due to Dewatripont and Tirole 1993), and directly connected to Aghion and Bolton's (1992) general approach to bankruptcy as a mechanism for transferring control rights between claim holders. See also Tirole (1994) for a survey of the main issues this approach may be used to explain.

This discussion briefly describes the main features of the model used by Dewatripont and Tirole in applying their theory to the banking sector. The model is very simple, with three dates.

At date 0 the initial balance sheet of the bank is given. Deposits $D_0$ and equity $E_0$ are used to finance loans $L_0 = D_0 + E_0$. The manager can improve the quality of these loans by exerting some effort, which costs $c$. The problem is to provide the manager with the incentives to exert this effort, which is assumed here to be the efficient solution. The manager's incentives will be closely related to the allocation of control rights between the regulator (which represents the depositors) and the stockholders.

At date 1 a first repayment $v$ is obtained from the loans, and a signal $u$ is observed about their future liquidation value $\eta$ at date 2; $u$ and $v$ are independent, but both are related to the level of effort. Suppose that $v$ is reinvested at a riskless rate normalized to zero. The final (overall) performance of the bank (the liquidation value of its assets) will therefore be $v + \eta$. After observing $u$ and $v$, the controlling party (the board of directors on behalf of the stockholders, or the regulator representing the depositors) decides if the bank will continue to operate (action $C$ for "continuing") or if it will be reorganized (action $S$ for "stopping"). This action determines the distribution of $\eta$, conditionally on $u$; it is denoted $H_A(\eta|u)$, where $A \in \{C, S\}$.

At date 2 the liquidation value $v + \eta$ is observed.

The crucial point is that the action $A$ is noncontractible. Therefore the determination of the controlling party at $t = 1$ will be fundamental. This is the role of the solvency regulation. For simplicity, it is assumed that monetary incentives cannot be given to the manager. Incentives for managerial effort can be given only indirectly through the threat of reorganizing the bank, in which case the manager will be fired and will lose the private benefit $B$ attached to running the bank.

Since $u$ and $v$ are independent, the optimal action under complete information depends only on $u$. The expected profit $D(u)$ from continuing (instead of stopping) at $t = 1$ (conditionally on $u$) is computed as

$$D(u) \stackrel{\text{def}}{=} E[\eta|u, C] - E[\eta|u, S],$$

which is equal to

$$D(u) = \int_0^{+\infty} \eta \, dH_C(\eta|u) - \int_0^{+\infty} \eta \, dH_S(\eta|u),$$

or, after integrating by parts,

$$D(u) = \int_0^{+\infty} \{H_S(\eta|u) - H_C(\eta|u)\} \, d\eta.$$

Continuing is optimal under complete information if and only if $D(u)$ is non-negative. To fix ideas, Dewatripont and Tirole assume that $D(\cdot)$ is increasing, so that the best rule can be described as follows: *continue when* $u \geq \hat{u}$; *stop when* $u < \hat{u}$. The threshold $\hat{u}$ is defined by

$$D(\hat{u}) = 0.$$

From now on, assume that the effort level of the manager, which can take only two values: $e = \underline{e}$ (insufficient) or $e = \bar{e}$ (correct), is not observable by others. However $u$ and $v$ are positively correlated with $e$. Higher realizations of $u$ (or $v$) indicate a greater likelihood that $e = \bar{e}$. If $f(u|e)$ and $g(v|e)$ denote the conditional densities of $u$ and $v$, this means that

$$\frac{f(\cdot|\bar{e})}{f(\cdot|\underline{e})} \quad \text{and} \quad \frac{g(\cdot|\bar{e})}{g(\cdot|\underline{e})}$$

are both increasing functions. Let $x(u, v)$ denote the probability of continuing when $(u, v)$ is observed. The second-best decision rule is obtained by maximizing the expected (incremental) profit from continuing $\iint x(u, v) D(u) f(u|\bar{e}) g(v|\bar{e}) \, du \, dv$ under the incentive compatibility constraint:

$$B \iint x(u, v) \{f(u|\bar{e}) g(v|\bar{e}) - f(u|\underline{e}) g(v|\underline{e})\} \, du \, dv \geq c,$$

which means that the expected loss from shirking is higher than the cost of effort. The Lagrangian of this problem is

$$L = \iint x(u, v) \{(D(u) + \mu B) f(u|\bar{e}) g(v|\bar{e}) - \mu B f(u|\underline{e}) g(u|\underline{e})\} \, du \, dv - \mu c,$$

where $\mu$ is the multiplier associated with the incentive constraint. Pointwise maximization of $L$ with respect to $x(u, v) \in [0, 1]$ gives the second-best decision rule:

$$x(u,v) = \begin{cases} 1 & \text{if } D(u) + \mu B \geq \mu B \dfrac{f(u|\underline{e})g(v|\underline{e})}{f(u|\bar{e})g(v|\bar{e})}, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, continuing is optimal under incomplete information if and only if

$$\frac{f(u|\bar{e})}{f(u|\underline{e})}\left\{1 + \frac{D(u)}{\mu B}\right\} \geq \frac{g(v|\underline{e})}{g(v|\bar{e})}. \tag{9.11}$$

Let $u^*(v)$ be defined as the value of $u$ such that condition (9.11) is satisfied with equality for a given value of $v$. Because the left side of (9.11) is increasing in $u$, continuing will be optimal if and only if $u \geq u^*(v)$. Moreover, the right side of (9.11) is decreasing in $v$; therefore the function $u^*(\cdot)$ is itself decreasing.

Let $\hat{v}$ be defined implicitly by

$$u^*(\hat{v}) = \hat{u}.$$

Figure 9.2 illustrates the differences between the best and the second-best decision rules. The shaded areas correspond to the two regions of ex post inefficiency. For $v > \hat{v}$, there are values of $u(u \in [u^*(v), \hat{u}])$ for which the bank is allowed to continue, although ex post efficiency would imply closing it (inefficient inaction). For $v < \hat{v}$, there are values of $u(u \in [\hat{u}, u^*(v)])$ for which the bank is stopped, although ex post efficiency would imply continuing (inefficient intervention).

The crucial step in the Dewatripont-Tirole theory of financial structure is to show that a convenient combination of debt and equity can provide outsiders with the appropriate incentives to implement this ex post inefficient decision rule. As is well known, the payoff of equity is a convex function of the profit of the bank, which implies that equity holders tend to favor risky decisions. Symmetrically, the payoff of a deposit is a concave function of the profit of the bank, which implies that depositors tend to favor less risky decisions. Therefore, under the reasonable assumption that closing the bank is less risky than continuing, stockholders (resp. depositors) will have a tendency to excessive forbearance (resp. liquidation). It is then not surprising that stockholders (resp. depositors) should be given the control rights of the bank when the first-period performance is good, $v \geq \hat{v}$ (resp. when this performance is bad, $v < \hat{v}$). Exact implementation of the second-best optimal decision rule can then be obtained by several alternative means: composite claims, net worth adjustments, or voluntary recapitalization (see Dewatripont and Tirole 1994, 81–84).

Note that this theory is very general. It can be applied as well to managerial corporations (with bondholders or creditors replacing depositors). The special characteristic of banks is that their creditors (depositors) are small and uninformed. Thus the depositors are not in a position to monitor the bank's manager. The role of the

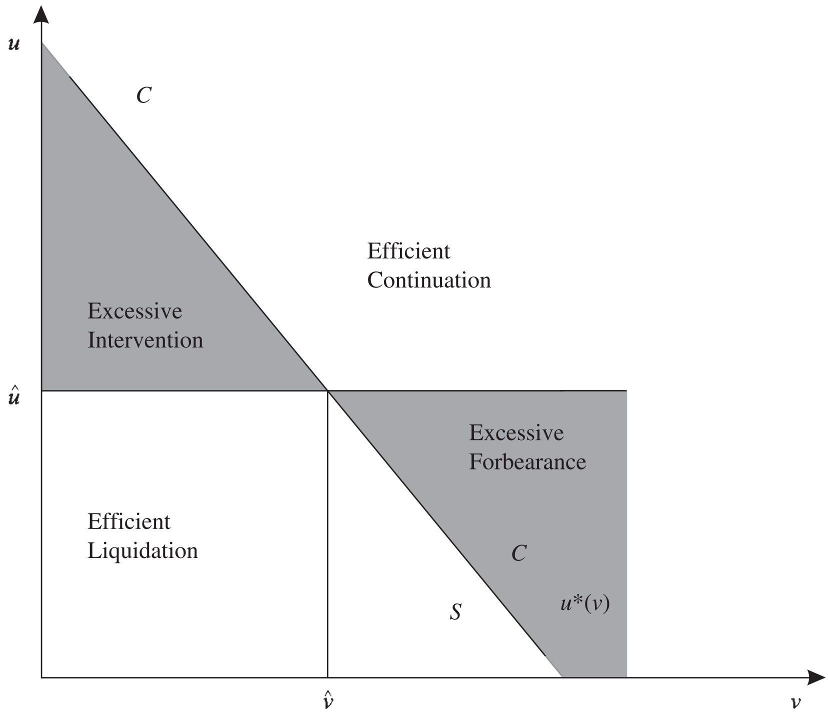



**Figure 9.2**
Best and second-best decision rules (Dewatripont and Tirole 1994, 8.66).

regulator is to represent their interest and act on their behalf. Therefore, the solvency regulation of banks brings in a rule specifying the conditions under which the stockholders remain in control of the bank or the regulator comes to the fore. A detailed discussion of the practice of solvency regulation in light of this theory is given by Dewatripont and Tirole.

### 9.4.5 The Three Pillars of Basel II

The Basel Accord of 1988 set a simple standard for harmonizing solvency regulations for internationally active banks of the G-10 countries. However, it was criticized for having provoked a "credit crunch" (at the beginning of the 1990s), and it led to regulatory arbitrage activities by banks. After a long bargaining process of more than ten years with the banking profession, the Basel Committee on Banking Supervision set a new regulation (Basel II) that is supposed to rely on three pillars: a more complex capital ratio, proactive supervision, and market discipline. Rochet (2004) proposes a formal analysis of Basel II and suggests that the first pillar might be overemphasized with respect to the two others, generating some imbalance.

## 9.5 The Resolution of Bank Failures

In a Modigliani-Miller environment, in which the liability structure of a bank is irrelevant, bank closures would occur only when they are efficient (when the risk-adjusted NPV of continuation is negative). But when depositors are insured (and deposit insurance premiums cannot be adjusted in real time to the investment decisions of the banks), moral hazard appears. Bankers have incentives to take too much risk and to keep operating (at the expense of the deposit insurance fund) in situations in which liquidation would be efficient. The reason is that limited liability gives the bank owners the equivalent of a call option on the bank's assets. This option appreciates when the volatility of these assets increases, and it keeps a positive value even when the NPV of continuation is negative.

In such a context, banking regulations can be efficient only when they include closure policies that prevent such behavior (gambling for resurrection). More generally, regulation must address the question of resolution of banks' distress, which matters, for instance, when a profitable bank has liquidity problems. This section begins with a discussion of the instruments and policies that can be used for the resolution of banks' distress. It then examines the issue of information revelation by regulated banks, and the question of how to delegate the decision of bank closure to some organization (typically a regulatory body) that should be provided with adequate incentives to do its job properly.[12]

### 9.5.1 Resolving Banks' Distress: Instruments and Policies

Goodhart and Schoenmaker (1993) present an interesting study surveying 104 bank failures in 24 countries between 1970 and 1992.[13] They classify the resolution methods used by the banking authorities for dealing with these failures into four categories: (1) a rescue package (which may include emergency aid by the Central Bank and a recapitalization by the stockholders), (2) a takeover by other banks (under the "purchase and assumption" regime), (3) the creation of a special regime administered by the government or the deposit insurance fund (in case of chain failures, such as the savings and loan debacle in the United States, or the banking crises in Scandinavia and Japan), and (4) liquidation of the financial institution. Among the lessons drawn by Goodhart and Schoenmaker are the following conclusions:

- Bank failures are not uncommon, nor are they limited to a few countries.
- Authorities have been reluctant to see such failures end in straightforward liquidation (only 31 out of 104 instances).
- Separation of authority between monetary and supervisory agencies is less likely to lead to involvement by taxpayers or by other commercial banks in the form of financing rescues ("survivors pay" principle).[14]

If recapitalization is not possible or not desirable, the deposit insurance company takes over the bank. It then has the option of either liquidating the bank (payoff and liquidation), in which case uninsured depositors will not receive full payment, or keeping it in operation and selling it as a going concern (under the purchase and assumption procedure).[15] In this latter case, the deposit insurance company can obtain a higher price by auctioning off the bank (because bidders will value the bank's goodwill), but it will also be bound to make a full payment to all uninsured depositors (and even to all other depositors if a no-preference clause is present). Note that bank regulation usually allows the regulator to dismiss a bank's manager, but this does not always occur.

The banking authorities must also choose the procedure to be used for solving a bank's distress. In particular, they can adopt rigid rules that condition closure upon verifiable criteria, or delegate the decision to the monetary authorities (the Central Bank) or to the deposit insurance fund. The preferences of these two institutions are of course different, and delegation may result in excessive forbearance or excessive intervention. The following sections examine the relation between closure policy and the banks' incentives to provide truthful information, and the contributions of Repullo (1993) and Kahn and Santos (2005), who model the bank closure issue in an incomplete contract setup.

### 9.5.2 Information Revelation and Managers' Incentives

Mitchell (2000) and Aghion, Bolton, and Fries (1999) point out that in order to close down a bank it is first necessary to obtain evidence that the bank is in financial distress. But this accurate information is privately held by bank management. As a consequence, the criticism of banks bailouts as being the result of lenient regulatory bodies unwilling to exert discipline and provide the right incentives to the bank managers might prove incorrect. On the contrary, bank closures may not be feasible except if the regulator is lenient enough to give managers the incentive to truthfully reveal this information. Mitchell and Aghion-Bolton-Fries acknowledge that bank managers have the ability to misreport the extent of their banks' loan losses simply by extending credit to bad borrowers. This will be the case if the revelation of too large loan losses implies that bank managers will be dismissed. The incentives to forbear are here in cascade: the threat of bank closure triggers a bank's forbearance, and a bank's forbearance gives incentives to the firms' managers to forbear on their bad projects.

To illustrate this point, consider a zero-interest-rate, risk-neutral world with a population of firms that obtain zero in case of liquidation and $y$ with probability $p$ if they continue their projects. The bank faces a proportion $\theta$ of solvent firms that repay $R$ on their loans, the remaining $1-\theta$ being insolvent and having a liquidation value of $L$, larger than the expected value of continuation: $L > yp$.

The bank has private information on $\theta$, which can take two values, $\theta \in \{\theta_L, \theta_H\}$, with $\theta_L < \theta_H$. We assume that the bank's loan liquidation can be observed. This leads to a disymmetric situation because for legal reasons a $\theta_H$ bank will never be able to liquidate solvent firms and will have no degree of freedom, whereas a $\theta_L$ bank will be able to roll over loans of the insolvent firms and hide its loan losses in this way.

As in Dewatripont and Tirole (1994), the bank manager exerts an effort $e$, with $e \in \{e_L, e_H\}$, and $e_L < e_H$, which we identify without loss of generality with the project's probability of success. The difference in the cost of implementing the high level of effort is $c$.

Assume the bank manager's objective function is a fraction of the bank's profit, provided she retains her position, so that ex ante the bank manager's expected profit is proportional to $x\Pi$, where $x$ is the probability of remaining in charge and the bank's profit $\Pi$.

We assume that a bank facing a proportion $\theta_L$ of solvent firms is itself insolvent, whereas a bank facing a proportion $\theta_H$ is solvent. If $D$ is the bank amount of debt, this is equivalent to

$$\theta_L R + (1 - \theta_L)L < D < \theta_H R + (1 - \theta_H)L.$$

We drastically simplify Aghion-Bolton-Fries model by assuming that the regulator has no choice but to provide the required level of capital.[16] The regulator's policy is then summarized in the probability $x(\hat{\theta})$ of the bank manager's retaining her position as a function of the declared proportion of loan losses, $\hat{\theta}$. Since these losses can only take two values, we simplify the notation and write $x(\theta_L) = x_L$, and $x(\theta_H) = x_H$.

As the regulator faces imperfect information, it has to design an incentive policy that will result in truthful disclosure of the $\theta_L$ loan losses. This is given by

$$x_L[\theta_L R + (1 - \theta_L)L] \geq x_H[\theta_L R + (1 - \theta_H)L + (\theta_H - \theta_L)yp], \tag{9.12}$$

where the left-hand side gives the bank manager's profit from truthful revelation, and the right-hand side gives the profit from announcing $\theta_H$ and rolling over a proportion $(\theta_H - \theta_L)$ of bad loans.

The incentive constraint (9.12) can also be written as

$$x_L \geq \alpha x_H,$$

where

$$\alpha = \frac{[\theta_L R + (1 - \theta_H)L + (\theta_H - \theta_L)yp]}{[\theta_L R + (1 - \theta_L)L]} < 1.$$

Assume now that the regulator acknowledges the information asymmetry and defines a policy satisfying (9.12). Then, after simplification, the incentives to exert a high level of effort for a policy satisfying (9.2) are given by

$$x_H(\theta_H R + (1-\theta_H)L) - x_L(\theta_L R + (1-\theta_L)L) \geq c, \tag{9.13}$$

an expression that can be written as

$$x_L \leq \beta x_H - \gamma, \tag{9.14}$$

where

$$\beta = \frac{(\theta_H R + (1-\theta_H)L)}{(\theta_L R + (1-\theta_L)L)} > 1, \qquad \gamma = \frac{c}{(\theta_L R + (1-\theta_L)L)}.$$

If (9.12) is not satisfied, this expression holds with different parameters, $\beta'$ and $\gamma'$.

The economic interpretation of (9.12) is brought about by considering the effect of a tough regulatory policy of closing down bad ($\theta_L$) banks, which supposedly would have the right incentives on the effort level. This would be defined as $x(\theta_L) = 0$, and $x(\theta_H) = 1$. The incentive constraint (9.12) is not satisfied, and as a consequence, not only bad projects are continued but also the incentives to exert a high level of effort are low because the probability for the manager to remain in place, given his revelation strategy (that is, to lie) is affected only by the amount $(\theta_H - \theta_L)(L - yp)$.

On the other hand, for the moral hazard constraint (9.13) to be satisfied, the regulator has to commit to a significant difference in the probability of continuation. With $x_H = x_L$, it is obvious that (9.13) will only hold for very small values of $c$.

Thus, since the two conditions work in opposite directions, they may or may not be compatible, that is, expressions (9.12) and (9.13) may or may not have an intersection for $0 \leq x_L \leq 1$; $0 \leq x_H \leq 1$. It may be that no policy exists satisfying both the incentive compatibility and the moral hazard constraints. If both constraints are satisfied, the regulator will be able to obtain revelation and implement the best level of effort. Still, the price to pay will be to give sufficient compensation to managers who truthfully report their loan losses, which could be misinterpreted as a lenient policy.

The analysis can be extended by assuming that the level of liquidation is not observable. In this case, the regulator has to consider in addition to the preceding constraints the incentives of the $\theta_H$ banks to misrepresent their types. This is the case if the regulator offers a recapitalization of bankrupt banks. As Aghion-Bolton-Fries note, too lenient a policy may lead $\theta_H$ banks to misrepresent their types by reporting a large level of loan losses.

### 9.5.3 Who Should Decide on Banks' Closure?

Repullo (2000) and Kahn and Santos (2005) analyze the problem of optimal delegation of bank closure decisions in a model inspired by Dewatripont and Tirole's (see section 9.4.4). This closure decision can depend on two variables: a verifiable signal $v$ (assumed to give withdrawals on the bank's deposits at the interim date) and a nonverifiable signal $u$ that stands for the bank's probability of success. Since $u$ is nonverifiable, the allocation of control between the Central Bank and the deposit insurance fund (DIF) can depend only on $v$. Of course, once this control has been allocated, the controlling party can also base its decision on $u$. In fact, it will make the decision that maximizes its own preferences. The main result of Repullo is that if the Central Bank is a junior creditor with respect to the DIF, it is optimal to allocate control to the Central Bank when withdrawals are small and to the DIF when they are large.

Depositors are modeled as in chapter 7. Each of them invests 1 at $t = 0$ and decides to withdraw $v$ at $t = 1$, and $1 - v$ at $t = 2$ (the interest rate is normalized to zero). The bank invests $1 - I$ in liquid assets and $I$ in a long-run risky technology that yields $Y$ with probability $u$ and zero with the complementary probability.

The bank's investments can also be liquidated for a value $L$ $(L < 1)$ at $t = 1$. The signal $u$ on future returns is publicly observed at $t = 1$. From an ex ante viewpoint, both $u$ and $v$ are random.

The closure of the bank has a cost $c$, which is assumed to be the same at $t = 1$ and $t = 2$.[17] We focus on the case of an illiquid bank (such that $v > 1 - I$) that cannot obtain a loan from the market. Assuming the interbank market is perfectly competitive, so that interbank loans have a zero expected return, a loan with nominal repayment $R$ $(R < IY)$ is possible provided that

$$u > \frac{v - (1 - I)}{IY}.$$

We assume this condition is not fulfilled, implying that the bank cannot find liquidity support on the market.[18]

In order to analyze the closure decision, consider first the efficient closure decision where continuation is chosen provided it leads to a higher expected return:

$$u[IY + (1 - I)] + (1 - u)[(1 - I) - c] \geq IL + (1 - I) - c, \tag{9.15}$$

or

$$u \geq u^* = \frac{IL}{IY + c}. \tag{9.16}$$

As expected, the decision is independent of the liquidity shock. The following discussion compares different mechanisms that delegate the closure decision to a regulatory body.

**A Unique Regulator**

Consider the case of a single regulator that is able to provide liquidity and is responsible for deposit insurance. In case of failure, all depositors are fully reimbursed.

The regulator is supposed to choose closure decisions that minimize its total cost. So, assuming the loan bears no interest, and $R$ is the amount repaid on the loan

$v - (1 - I)$ in case of success, the regulator will choose continuation of the bank provided that

$$uR + (1 - u)(-(1 - v) - c) - (v - (1 - I)) > IL + (1 - I) - 1 - c, \tag{9.17}$$

which, using $R = v - (1 - I)$ simplifies to

$$u \geq \hat{u} = \frac{IL}{I + c}. \tag{9.18}$$

Since $Y > 1$, we observe that even in the case of a unique regulator closure decisions are inefficient. This inefficiency corresponds to the fact that the regulator has a tendency to close the bank excessively often, as $\hat{u} > u^*$. The cause of this inefficiency is that the regulator is concerned only with downside risk. The upper tail of the distribution $IY - R$ does not enter its objective function.

Two remarks are in order. First, the regulator may overestimate the bankruptcy cost if its reputation is at stake. In this case, for sufficiently large perceived cost, the result may be modified, and the regulator may allow for too much forbearance. Second, the inefficiency may be eliminated if we allow the regulator to seize all the bank's assets, so that with $R = IY$ it captures all the surplus.

**Decentralizing with Two Regulatory Agencies**

Decentralization in the present context implies allocating the power to close down the bank to the DIF or to the Central Bank, operating in its lender-of-last-resort capacity.[19]

We assume bankruptcy costs are shared between the two institutions and denote $\beta$ the fraction of cost that is contributed by the Central Bank.

If the DIF is in charge of closure decisions, its choice of continuation will be exactly the same as when there is a unique regulator, that is, given by (9.18). The only change is that the bankruptcy cost is now shared. This choice does not depend upon the withdrawal volume $v$:

$$u \geq u_{DI} = \frac{IL}{R + (1 - \beta)c}.$$

On the other hand, if the Central Bank is in charge of closure decisions, it will not bear the cost of payments to depositors, so it will provide a loan of nominal repayment $R$ (and thus choose continuation) provided that

$$uR + (1 - u)(-\beta c) - (v - (1 - I)) \geq -\beta c.$$

If the loan bears no interest rate, $R = v - (1 - I)$ and the Central Bank will lend provided that

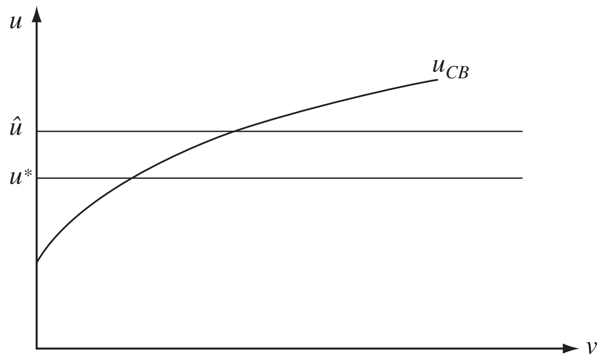



**Figure 9.3**
Closure policies.

$$u \geq u_{CB}(v) = \frac{v - (1 - I)}{v - (1 - I) + \beta c}.$$

The function $u_{CB}(\hat{v})$ is increasing and satisfies $u_{CB}(1) > u^*$ ($\beta < 1$ and $L < 1$).

Figure 9.3 summarizes the different closure policies. The best closure rule and the one chosen by the DIF are independent of $v$. The inequality $u_{DI} > u^*$ obtains for values of $\beta$ and $R$ such that $R + (1 - \beta)c < IY + C$; otherwise the reverse obtains.

Using this framework, Repullo (2000) shows that under some reasonable assumptions on the distribution of $u$, it is optimal that the Central Bank be responsible for dealing with small liquidity needs, whereas the decision for large liquidity needs should be in hands of the DIF. Kahn and Santos (2005) consider a modified version of Repullo's model. They show that centralizing the functions of lender of last resort and bank supervision within the same institution may lead to excessive forbearance and to suboptimal investment.

## 9.6 Market Discipline

The important role of market discipline in the regulatory framework has been widely acknowledged. In Basel II it is viewed as the third pillar of regulation, the other two being capital adequacy and supervision.

Market discipline should reduce the bank manager moral hazard problem of excessive risk taking by making the bank pay the actual cost of its risk taking.

To some extent, the discussion about market discipline has been led astray by lack of precision. At least the notion of market discipline should be clarified by distinguishing ex ante market discipline and interim market discipline, which reflect the type of monitoring market discipline may provide. Ex ante market discipline implies that the market fully reflects the risk taken by the manager so that the bank pays the cost of its liabilities. Interim market discipline implies that if the bank manager chooses an action that is detrimental to the market value of liability holders, the latter are able to discipline it by liquidating the bank or taking control over it.

Typically this action may be the redemption at par of some type of liabilities. Alternatively, the action may be bank supervision, and interim market discipline may trigger it by leading to too large a spread in the secondary market for subordinated debt.

### 9.6.1 Theoretical Framework

To better explain the distinction between ex ante and interim market discipline, we resort to the model of 9.3.1 but replace $\theta$, a measure of the probability of success, with $\sigma$, a measure of the riskiness of the bank investment project, in line with the models of Suarez (1993a; 1993b) and Matutes and Vives (1996) (see section 3.5.1).

Because the discussion regarding market discipline has focused on subordinated debt, it is convenient to model three types of liabilities: deposits $D$, subordinated debt $S$, and equity $E$. Subordinated debt is junior to deposits, so that denoting by $D'$ and $S'$ the nominal repayments on deposits and on subordinated debt, the effective repayments are given by $\min(x, D')$ for deposits, and $\min([x - D']_+, S')$ for subordinated debt, as a function of $x$, the liquidation value of the bank's assets. In a zero-interest-rate, risk-neutral framework we have

$$\int_0^{D'} x\, dF(x, \sigma) + \int_{D'}^{\infty} D'\, dF(x, \sigma) = D, \tag{9.19}$$

$$\int_{D'}^{D'+S'} (x - D')\, dF(x, \sigma) + \int_{D'+S'}^{\infty} S'\, dF(x, \sigma) = S. \tag{9.20}$$

The bank's expected profit is

$$\Pi(\sigma) = \int_{D'+S'}^{\infty} (x - D' - S')\, dF(x, \sigma). \tag{9.21}$$

If subordinated debt holders and depositors have perfect information or no information, the choice of the level of risk $\sigma$ is determined as in section 3.5.1. Absent a charter value there is indeterminacy in the first case, and in the second case the level of risk is maximum. The interesting case occurs when depositors are uninformed and subordinated debt holders are informed. This is a reasonable assumption if we consider that small depositors do not have an incentive to invest in information, whereas large subordinate debt holders do.

In this case, $S'$ is a function of $\sigma$, but $D'$ is independent of $\sigma$. This implies that replacing (9.19) and (9.20) in (9.21), we obtain

$$\Pi(\sigma) = E[\max(x, D')] - S.$$

As a consequence, the bank's choice of risk will depend upon the bank's charter value $V$. There exists a threshold $\hat{V}(D')$ such that for $V > \hat{V}(D')$ the bank will take the minimum level of risk; for $V < \hat{V}(D')$ the bank will take the maximum level of risk. From the ex ante market discipline point of view, subordinated debt and equity play exactly the same roles.

Interim market discipline is best modeled using models such as Chari and Jagganathan (1988) or Calomiris and Kahn (1991) (see section 7.3). The informed liability holders observe an interim signal on the bank's return and exercise their option to liquidate the bank if the observed signal predicts a low return. The model can be adapted to monitoring the level of risk because subordinated debt is negatively affected by risk. It may alternatively trigger a bank examination if a market for subordinated debt exists.

In any case, three conditions are required for subordinate debt holders to exert market discipline:

- The bank has to have incentives to issue subordinated debt at the initial stage.[20]
- The volume of informed liability debt holders has to be large enough to trigger liquidation.
- Liability debt holders have to have the incentives to monitor the bank's risk (in spite of the existing free-rider problem).

### 9.6.2 Empirical Evidence

The empirical literature on subordinated debt has examined three issues:

- To what extent does the market price for subordinated debt reflect a bank's future expected returns? In other words, the issue is to assess whether the market for subordinated debt is informationally efficient.
- Should the regulator consider subordinated debt spreads to complement its information and trigger its action?
- Does subordinated debt implement market discipline?

The third issue has only been slightly touched upon. Bliss and Flannery (2000) "find no prima facie support for the hypothesis that bond holders or stockholders influence day-to-day managerial actions in a prominent manner consistent with their own interests." Absent any liquidation threat there is no possible justification for market discipline, so that the absence of empirical evidence for interim market discipline is not surprising.

Regarding the first issue (subordinated debt and informational efficiency), it seems that subordinated debt prices (in the United States) did not contain a risk premium before the 1980s but have incorporated such premia since then (Flannery and Sorescu 1996). This divergence could be due to the existence of implicit government guarantees during the early period. Sironi (2000) obtains similar evidence for Europe, with different levels of government support in the different countries.

Concerning the use of market spreads as a regulatory instrument, empirical results show that increases in the spreads on subordinated debt may help predict CAMEL downgrades (Evanoff and Wall 2000). Regarding whether the information on subordinated debt is better than the one provided by the stock market, Berger, Davies, and Flannery (2000) show that subordinated debt and bonds are better predictors than stocks. Martinez-Peria and Schmukler (2001) examine the efficiency of market discipline by using a panel of Latin American banks during the last banking crises in Argentina, Chile, and Mexico. They show that uninsured depositors do punish their banks when the banks take too much risk.

## 9.7 Suggestions for Further Reading

Besanko and Thakor (1992) use a spatial differentiation model à la Salop (1979) to model the implications of banking deregulation, and more specifically, of relaxing entry barriers. Borrowers and depositors are uniformly located on a circle, and choose the bank that offers them the best combination of interest rate and proximity. Since there is no interbank market in this model, banks must use equity to match the (possible) gap between loans and deposits. It is assumed that banks' shareholders are risk-neutral and therefore try to maximize expected profit. Because of transportation costs, banks extract surplus from borrowers and lenders, and bank charters have a positive value.[21] Besanko and Thakor study the decision made by banks at equilibrium. They show that increasing the number of banks implies increasing the deposit rate and decreasing the loan rate but also decreasing the equity-debt ratio. When a capital requirement is introduced, it has the effect of decreasing interest rates on both deposits and loans, which means that borrowers benefit from the capital requirement, whereas depositors are hurt.

Campbell, Chan, and Marino (1992) formalize the notion of susbstitutability between capital requirements and monitoring in controlling the behavior of bank managers. This control can be exerted directly by regulators or indirectly by giving adequate incentives to stockholders to do so. An interesting aspect of their article is that they also explicitly study the incentives of depositors to monitor their banks.

Kane (1990) has convincingly argued that the adverse incentives of regulators were one of the main explanations for the problems of U.S. depository institutions in the 1980s. In particular, if regulators have long-run career concerns, they have an interest in delaying the disclosure of difficulties encountered by the institutions under their supervision. Campbell, Chan, and Marino consider three versions of their model:

• Monitoring of banks' assets is impossible, and the regulator uses capital requirements to prevent excessive risk taking by the bank.

• Monitoring is feasible, and the regulator is benevolent. There is substitutability between bank capital and monitoring effort. At the optimum, capital requirements are less stringent and simultaneously the banks take less risk.

• Monitoring is still feasible, but the regulator is self-interested. The crucial limitation to the incentive scheme that the depositors have to design for the monitor is the limited liability of the monitor. In that case, the penalty that can be inflicted on a shirking monitor is limited; this induces distortions on the levels of capital and monitoring that were obtained in second version. As expected, more capital will be needed and less monitoring effort will be required.

Boot and Greenbaum (1993) analyze the interaction of reputation and financial structure on the risk-taking and monitoring behavior of commercial banks. They distinguish three sources of finance for banks: inside equity, insured deposits, and outside finance raised on capital markets (such as outside equity, wholesale deposits, or subordinated debt). Since the cost of outside finance is related to the bank's reputation, the bank has an incentive to avoid risk and increase monitoring to improve its reputation, thus improving efficiency. These reputational benefits are therefore a substitute for the rents that banks may obtain from imposed restrictions on entry or competition in the banking sector. Boot and Greenbaum show that these reputational benefits are negligible when the bank invests in safe assets but important when the bank's assets are risky. Given that efficiency improves with reputational benefits, they see this as a new support for the narrow banking proposal of investing all insured deposits into safe assets, whereas uninsured deposits (and more generally outside finance collected on capital markets) can be invested in risky assets. Their results are in line with the general perception that outside (short-term) uninsured finance improves market discipline, an idea that is also developed by Rey and Stiglitz (1993).

Finally, Smith (1984) explores the structure of banking competition in a Diamond-Dybvig environment. He considers several banks that compete for deposits by offering first- and second-period interest rates. When information is perfect, the optimal contract is obtained. But when there are two types of depositors, each of them characterized by a different probability of withdrawing early, under adverse selection a classical nonexistence problem first pointed out by Rothschild and Stiglitz (1976) may be faced (see section 5.4 for a similar development by Bester 1985). The failure of equilibrium to exist is due to the fact that the equilibrium contracts, either separating or pooling, are destroyed by the existence of positive profit contracts that are addressed specifically to a segment of depositors. Smith interprets this nonexistence of equilibrium as an instability of the deposit market (a point she does not develop explicitly in a dynamic model) and argues that regulating the deposit rate is the appropriate response.

## 9.8 Problem

### 9.8.1 Moral Hazard and Capital Regulation

Consider an economy where banks could invest either in a safe project that yields $G$ with probability $P_G$ and zero otherwise, or in a high-risk project that yields $B$ with probability $P_B$ and zero otherwise. The project has constant returns to scale and satisfies $G < B$ and $P_G G > P_B B > 1$. To develop project $G$ requires additional effort with a unitary cost $c$.

Banks are financed by short-term unsecured deposits with a return $r_D$ per unit of deposit. Depositors are risk-neutral and will require an expected return equal to the risk-free rate, which is normalized to zero. We assume the participation constraint for banks is satisfied.

Capital is costly because equity holders require an expected return of $\rho > 0$.

1. Describe the competitive equilibrium in the absence of bank capital, and determine under what conditions the safe project, $G$, or the risky one, $B$, will be implemented.

2. Assume that in the absence of bank capital the only equilibrium obtained is characterized by implementing the $B$ project. Determine the minimum level of capital a bank needs in order to restore the possibility of an equilibrium where the safe project $G$ is preferred by banks.

3. Assume depositors observe banks' capital. What will be the amount of capital a profit-maximizing bank will choose?

## 9.9 Solution

### 9.9.1 Moral Hazard and Capital Regulation

1. The bank will implement the technology if

$$P_G(G - R) - c \geq P_B(B - R),$$

which is equivalent to

$$R \leq R^* = \frac{P_G G - P_B B}{P_G - P_B} - c.$$

If the banks hold zero capital, and the good project is chosen, competitive equilibrium in the deposit market yields $R = 1/P_G$, implying

$$\frac{1}{P_G} \leq R^* = \frac{P_G G - P_B B}{P_G - P_B} - c.$$

On the other hand, if the bad project is chosen, $R = 1/P_B$, implying

$$\frac{1}{P_B} > R^* = \frac{P_G G - P_B B}{P_G - P_B} - c.$$

Consequently, for $R^* > 1/P_G$, the safe project will be chosen in equilibrium and $R = 1/P_G$; for $R^* < 1/P_G$ the risky project will be chosen and $R = 1/P_B$; finally, for

$$R^* \in \left[\frac{1}{P_G}, \frac{1}{P_B}\right]$$

there are multiple equilibria, each equilibrium being realized with self-fulfilling expectations about the project choice.

2. Assume the bank holds an amount of capital $k$ for each unit of investment. Then, if the safe project is chosen, the repayment is

$$R = \frac{1-k}{P_G}.$$

$\rho$ does not play any role in the technology choice because it cancels out. The incentives to implement the safe project are restored if

$$\frac{1-k}{P_G} \leq R^* = \frac{P_G G - P_B B}{P_G - P_B} - c.$$

3. Since depositors observe the bank's capital, if the bank chooses a level of capital such that $(1-k)/P_G > R^*$, the repayment to depositors will be $(1-k)/P_B$. The bank will therefore choose the level of $k$ that maximizes its profits,

$$\max_k \; p_B\left(B - \frac{1-k}{P_B}\right) - (1+\rho)k$$

that is, $k = 0$, and its profits will be $\Pi = p_B B - 1$.

If, instead, the bank chooses the minimum level of capital such that $G$ is credibly implemented, $(1-k)/P_G = R^*$ the bank's profits will be $\Pi = p_G G - 1 - \rho(1 - P_G R^*)$.

So, for a low cost of capital $\rho$, $(\rho \leq P_G G - p_B B - \rho^*(1 - P_G R^*))$ banks prefer to choose a capital level $(1 - P_G R^*)$, while for a larger cost of capital they prefer to hold no capital.

## Notes

1. Laffont and Tirole (1993) distinguish (1) *informational constraints*, which limit regulation because the relevant information is held by the firm; (2) *transactional constraints*, which limit the possibility of writing contingent contracts; and (3) *administrative and political constraints*, which impose limits on the scope of regulation as well as on the available instruments.

2. Bhattacharya, Boot, and Thakor (1998) establish a list of the five main unsolved issues:

- Are demand deposits important for investors' welfare?
- Is the safety net of deposit insurance necessary?
- What should be the goal of financial regulation?
- What role, if any, should the government play in coping with liquidity shocks?
- What portfolio restrictions should be imposed on banks?

3. Other economists adopt similar views. For example, Dowd (1992) challenges the view that fractional reserve banking is inherently liable to runs and crises, and Kaufman (1994) argues that the likelihood of contagion in a properly established system and the size of externalities in case of banking failures are not greater in banking than in other industries.

4. A contribution by Holmström and Tirole (1996) offers a new rationale for the superiority of banks over financial markets in the provision of liquidity insurance. Their model focuses on the liquidity needs of firms. They show that when moral hazard is present, financial markets are dominated by banks in the provision of liquidity insurance. A variant of this model is used by Rochet and Tirole (1996) to model interbank lending.

5. Three in the United States: the Comptroller of the Currency, the Federal Reserve System, and the Federal Deposit Insurance Corporation for national chartered commercial banks.

6. As mentioned, there is a competing school of thought that sees regulation as the result of opportunistic behavior of politicians and considers that regulators are captured by the industry they officially supervise in the interest of the general public.

7. In the United States deposit insurance schemes were privately developed prior to the creation of the Fed.

8. This is directly related to result 3.6 in section 3.5.

9. A put option on a security entitles its owner to sell the security at a future date at a prescribed price. For details, see Ingersoll (1987).

10. The articles discussed in this paragraph and the next are discussed in detail in chapter 8 as an application of the portfolio model to banking.

11. The subsequent regulation of market risks (including exchange rate, interest rate, or other off-balance-sheet asset risks) is ignored; see the Basel Proposal on Banking Supervision (Dermine 1993).

12. Of course, this question of rules versus discretion has a much broader relevance in economics. For example, Fischer (1994) provides an interesting discussion of rules versus discretion in the determination of monetary policy.

13. The main purpose of this article is to study whether the two main functions of Central Banks, namely, monetary policy and banking supervision, should be separately provided by distinct agencies.

14. The empirical evidence obtained by Ioannidou (2005) shows that when the same agency (e.g., the Fed) is in charge of both supervision and monetary policy, its bank supervisory behavior is altered.

15. We do not describe exhaustively all the procedures that can be used for solving banks' failures. The reader interested in a more complete approach to these institutional aspects is referred to Bovenzi and Muldoon (1990).

16. Aghion et al. (1999) consider a regulator who can impose recapitalization and replace the bank manager.

17. This is a restrictive assumption because an early closure at time $t = 1$ may have a lower social cost than a chaotic failure following forbearance and gambling for resurrection.

18. Notice, though, that this framework allows the analysis of an imperfect interbank market as well.

19. An alternative approach, followed by Kahn and Santos (2005), considers that the two institutions coexist. In that case, we might have banks that are not closed down by the DIF but that fail because the Central Bank does not provide its lending facility.

20. This may be difficult for small banks.

21. Even with free entry, however, this charter value would have to be positive to compensate for entry costs.

## References

Acharya, S., and J. F. Dreyfus. 1988. Optimal bank reorganization policies and the pricing of federal deposit insurance. *Journal of Finance* 44 (5): 1313–1334.

Aghion, P., and P. Bolton. 1992. An incomplete contract approach to financial contracting. *Review of Economic Studies* 59: 473–494.

Aghion, P., P. Bolton, and S. Fries. 1999. Optimal design of bank bailouts: The case of transition economies. *Journal of Institutional and Theoretical Economics* 155 (1): 51–70.

Akerlof, G. 1970. The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84 (3): 488–500.

Bensaid, B., H. Pagès, and J. C. Rochet. 1993. Efficient regulation of banks' solvency. Institut d'Economie Industrielle, Toulouse, France. Mimeograph.

Benston, G., R. Eisenbeis, P. Horvitz, E. Kane, and G. Kaufman. 1986. *Perspectives on safe and sound banking: Past, present and future*. Cambridge, Mass.: MIT Press.

Berger, A., S. Davies, and M. Flannery. 2000. Comparing market and regulatory assessments of bank performance: Who knows what when? *Journal of Money, Credit and Banking* 32 (3): 641–670.

Bernanke, F., and M. Gertler. 1990. Financial fragility and economic performance. *Quarterly Journal of Economics* 105 (1): 87–114.

Besanko, D., and A. V. Thakor. 1992. Banking deregulation: Allocational consequences of relaxing entry barriers. *Journal of Banking and Finance* 16: 909–932.

Bester, H. 1985. Screening vs rationing in credit markets with imperfect information. *American Economic Review* 75 (4): 850–855.

Bhattacharya, S., A. Boot, and A. V. Thakor. 1998. The economics of bank regulation. *Journal of Money, Credit and Banking* 3: 745–770.

Bhattacharya, S., and A. V. Thakor. 1993. Contemporary banking theory. *Journal of Financial Intermediation* 3: 2–50.

Black, F. 1975. Bank funds management in an efficient market. *Journal of Financial Economics* 2: 323–339.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–659.

Bliss, R., and M. Flannery. 2000. Market discipline in the governance of U.S. bank holding companies: Monitoring vs influencing. Working paper 2000-3, Federal Reserve Bank of Chicago.

Blum, J., and M. Hellwig. 1995. The macroeconomic implications of capital adequacy requirements for banks. *European Economic Review* 39 (3-4): 739–749.

Bolton, P., and X. Freixas. 2006. Corporate finance and the monetary transmission mechanism. *Review of Financial Studies* 19 (3): 829–870.

Bond, E., and K. Crocker. 1993. Bank capitalization, deposit insurance, and risk categorization. *Journal of Risk and Insurance* 60 (3): 547–569.

Boot, A. 1995. Challenges to competitive banking. Discussion paper, University of Amsterdam.

Boot, A., and S. Greenbaum. 1993. Bank regulation, reputation, and rents: Theory and policy implications. In *Capital markets and financial intermediation*, ed. C. Mayer and X. Vives. Cambridge: Cambridge University Press.

Boot, A., and A. V. Thakor. 1993. Self-interested bank regulation. *American Economic Review* 83 (2): 206–212.

Bovenzi, J. F., and M. E. Muldoon. 1990. Failure-resolution methods and policy considerations. *FDIC Banking Review* 3 (1): 1–11.

Bryant, J. 1980. A model of reserves, bank runs, and deposit insurance. *Journal of Banking and Finance* 4: 335–344.

Buser, S., A. Chen, and E. Kane. 1981. Federal deposit insurance, regulatory policy and optimal bank capital. *Journal of Finance* 36: 51–60.

Calomiris, C. 1993. Regulation, industrial structure, and instability in U.S. banking: An historical perspective. In *Structural change in banking*, ed. M. Klausner and L. White. New York: New York University.

Calomiris, C., and C. Kahn. 1996. The efficiency of self-regulated payment systems: Learning from the Suffolk system. *Journal of Money, Credit and Banking* 28 (4): 766–797.

Campbell, T. S., Y. S. Chan, and A. M. Marino. 1992. An incentive-based theory of bank regulation. *Journal of Financial Intermediation* 2: 255–276.

Capie, F., C. Goodhart, and N. Schnadt. 1995. The development of central banking. In *The future of central banking: The tercentenary symposium of the Bank of England*, ed. F. Capie, C. Goodhart, S. Fischer, and N. Schnadt. Cambridge: Cambridge University Press.

Cerasi, V., and S. Daltung. 2000. The optimal size of a bank: Costs and benefits of diversification. *European Economic Review* 44 (9): 1701–1726.

Chan, Y. S., S. I. Greenbaum, and A. V. Thakor. 1992. Is fairly priced deposit insurance possible? *Journal of Finance* 47: 227–245.

Chari, V. V., and R. Jagannathan. 1988. Bankingpanics, information, and retional expectations equilibrium. *Journal of Finance* 43 (3): 749–761.

Crouhy, M., and D. Galai. 1986. An economic assessment of capital requirements in the banking industry. *Journal of Banking and Finance* 10: 231–241.

———. 1991. A contingent claim analysis of a regulated depository institution. *Journal of Banking and Finance* 15: 73–90.

Dermine, J. 1993. The evaluation of interest rate risk: Some warnings about the Basle proposal. *Financial Markets and Portfolio Management* 7 (2): 141–149.

Dewatripont, M., and J. Tirole. 1993. Efficient governance structure: Implications for banking regulation. In *Capital markets and financial intermediation*, ed. C. Mayer and X. Vives. Cambridge: Cambridge University Press.

———. 1994. *The prudential regulation of banks*. Cambridge, Mass.: MIT Press.

Diamond, D. W. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51 (3): 393–414.

Diamond, D. W., and P. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91 (3): 401–419.

Dowd, K. 1992. *The experience of free banking*. London: Routledge.

Evanoff, D., and L. Wall. 2000. Subordinated debt and bank capital reform. Working paper 2000-24, Federal Reserve Bank of Atlanta.

Fama, E. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6 (1): 39–57.

———. 1985. What's different about banks? *Journal of Monetary Economics* 15: 29–40.

Fischer, S. 1994. Modern central banking. In *The future of central banking: The tercentenary symposium of the Bank of England*, ed. F. Capie, C. Goodhart, S. Fischer, and N. Schnadt. Cambridge: Cambridge University Press.

Flannery, M., and S. Sorescu. 1996. Evidence of bank market discipline in subordinated debt debenture yields, 1983–1991. *Journal of Finance*, Vol. 51.

Freixas, X., and E. Gabillon. 1999. Optimal regulation of a fully insured deposit banking system. *Journal of Regulatory Economics* 16 (2): 111–134.

Freixas, X., C. Giannini, G. Hoggarth, and F. Soussa. 2000. Lender of last resort: What have we learnt since Bagehot? *Journal of Financial Services Research* 18 (1): 63–87.

Freixas, X., and C. Holthausen. 2001. Interbank market integration under asymmetric information. Working paper 74, European Central Bank.

Freixas, X., and J. C. Rochet. 1995. Fairly priced deposit insurance: Is it possible? Yes. Is it desirable? No. Finance and banking discussion paper 4, Universitat Pompeu Fabra, Barcelona.

Freixas, X., and A. M. Santomero. 2001. An overall perspective on banking regulation. Working paper 02-1, Federal Reserve Bank of Philadelphia.

Friedman, M. 1960. A program for monetary stability. New York: Fordham University Press.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

Gennote, G., and D. Pyle. 1991. Capital controls and bank risk. *Journal of Banking and Finance* 15 (4-5): 805–824.

Giammarino, R. M., T. R. Lewis, and D. Sappington. 1993. An incentive approach to banking regulation. *Journal of Finance* 48: 1523–1542.

Glasner, D. 1989. *Free banking and monetary reform*. Cambridge: Cambridge University Press.

Goodhart, C. 1987. Why do banks need a central bank? *Oxford Economic Papers* 39: 75–89.

———. 1988. *The evolution of central banks*. Cambridge: Cambridge University Press.

Goodhart, C., and D. Schoenmaker. 1993. Institutional separation between supervisory and monetary agencies. In *Prudential regulation, supervision and monetary policy*, ed. F. Bruni. Milano: Universita Bocconi.

Gorton, G. 1993. Bank regulation, reputation and rents: Theory and policy implications: Discussion. In *Capital markets and financial intermediation*, ed. C. Mayer and X. Vives. Cambridge: Cambridge University Press.

Gorton, G., and G. Pennacchi. 1993. Money market funds and finance companies: Are they the banks of the future? In *Structural change in banking*, ed. M. Klausner and L. White. New York: Irwin.

Hayek, F. 1978. *The denationalisation of money*. London: Institute for Economic Affairs.

Hellmann, F. T., K. C. Murdock, and J. E. Stiglitz. 2000. Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough? *American Economic Review* 90 (1): 147–165.

Holmström, B., and J. Tirole. 1996. Modelling aggregate liquidity. *American Economic Review* 86 (2): 187–191.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–45.

Ingersoll, J. E., Jr. 1987. *Theory of financial decision making*. Totowa, N.J.: Rowan and Littlefield.

Ioannidou, V. P. 2005. Does monetary policy affect the central bank's role in bank supervision. *Journal of Financial Intermediation* 14 (1): 58–85.

Jensen, M., and W. R. Meckling. 1976. Theory of the firm, managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

John, K., T. John, and A. Saunders. 1994. *Journal of Banking and Finance* 18 (2): 307–323.

Kahane, Y. 1977. Capital adequacy and the regulation of financial intermediaries. *Journal of Banking and Finance* 1: 207–218.

Kahn, C., and J. Santos. 2005. Allocating bank regulatory powers: lender of last resort, deposit insurance and supervision. *European Economic Review* 49 (8): 2107–2136.

Kane, E. 1990. Principal agent problems in S&L salvage. *Journal of Finance* 45 (3): 755–764.

Kareken, J. H. 1986. Federal bank regulatory policy: A description and some observations. *Journal of Business* 59: 3–48.

Kareken, J. H., and N. Wallace. 1978. Deposit insurance and bank regulation: A partial equilibrium exposition. *Journal of Business* 51: 413–438.

Kaufman, G. G. 1994. Bank contagion: A review of the theory and the evidence. *Journal of Financial Services Research* 8 (2): 123–150.

Kay, J., and J. Vickers. 1988. Regulatory reform in Britain. *Economic Policy* 7: 286–343.

Keeley, M. C., and F. T. Furlong. 1990. A reexamination of mean-variance analysis of bank capital regulation. *Journal of Banking and Finance* 14: 69–84.

Kerfriden, C., and J. C. Rochet. 1993. Actuarial pricing of deposit insurance. *Geneva Papers on Risk and Insurance Theory* 18 (2): 111–130.

Kim, D., and A. M. Santomero. 1988. Risk in banking and capital regulation. *Journal of Finance* 43: 1219–1233.

Klausner, M., and L. White. 1993. Bank regulatory reform and bank structure. In *Structural change in banking*, ed. M. Klausner and L. White. New York: New York University.

Koehn, M., and A. M. Santomero. 1980. Regulation of bank capital and portfolio risk. *Journal of Finance* 35: 1235–1244.

Kroszner, R. S., and G. R. Rajan. 1994. Is the Glass-Steagall Act justified? A study of the U.S. Experience with universal banking before 1933. *American Economic Review* 84 (4): 810–832.

Laffont, J. J., and J. Tirole. 1986. Using cost observations to regulate firms. *Journal of Political Economy* 94: 614–641.

———. 1993. *A theory of incentives in procurement and regulation*. Cambridge, Mass.: MIT Press.

Litan, R. E. 1987. *What should banks do?* Washington, D.C.: Brookings Institution.

Mailath, G., and L. Mester. 1994. A positive analysis of bank closure. *Journal of Financial Intermediation* 3 (3): 272–299.

Marcus, A., and I. Shaked. 1984. The valuation of the FDIC deposit insurance using option-pricing estimates. *Journal of Money, Credit and Banking* 16: 446–460.

Martinez-Peria, M. S., and S. L. Schmukler. 2001. Do depositors punish banks for bad behavior? Market discipline, deposit insurance, and banking crises. *Journal of Finance* 56 (3): 1029–1051.

Matutes, C., and X. Vives. 1996. Competition for deposits, fragility, and insurance. *Journal of Financial Intermediation* 5 (2): 184–216.

———. 2000. Imperfect competition, risk taking, and regulation in banking. *European Economic Review* 44 (1): 1–34.

McCullough, H. 1981. Interest rate risk and capital adequacy for traditional banks and financial intermediaries. In *Risk and capital adequacy in commercial banks*, ed. S. J. Maisel. Chicago: University of Chicago Press.

Merton, R. 1977. An analytic derivation of the cost of deposit insurance and loan guarantees. *Journal of Banking and Finance* 1: 3–11.

———. 1978. On the cost of deposit insurance when there are surveillance costs. *Journal of Business* 51: 439–452.

Mishkin, F. 1992. *The economics of money, banking and financial markets*. London: Scott, Foresman.

Mitchell, J. 2000. Bad debts and the cleaning, transferring of banks' balance sheets: An application to transition economies. *Journal of Financial Intermediation* 10 (1): 1–27.

Mullins, H., and D. Pyle. 1991. Risk-based bank capital. University of California, Berkeley. Mimeograph.

Nakamura, C. 1993. Commercial bank information implications for the structure of banking. In *Structural change in banking*, ed. M. Klausner and L. White. New York: New York University.

Pennacchi, G. 1987. Alternative forms of deposit insurance, pricing and bank incentive issues. *Journal of Banking and Finance* 11: 291–312.

Rajan, R. G. 1992. Insiders and outsiders: The choice between informed and arm's-length debt. *Journal of Finance* 47 (4): 1367–1400.

Repullo, R. 1993. Who should decide on bank closures? An incomplete contract model. Working paper, Centro de Estudios Monetarios y Financieros, Madrid.

———. 2000. Who should act as a lender of last resort? An incomplete contracts model. *Journal of Money, Credit and Banking* 32 (3, pt 2): 580–605.

Rey, P., and J. Stiglitz. 1993. Short-term contracts as a monitoring device. NBER working paper W4514.

Rochet, J. C. 1992a. Capital requirements and the behaviour of commercial banks. *European Economic Review* 36: 1137–1170.

———. 1992b. Towards a theory of optimal banking regulation. *Cahiers Economiques et Monétaires de la Banque de France* 40: 275–284.

———. 2004. Rebalancing the three pillars of Basel 2. Federal Reserve Bank of New York. *Economic Policy Review* 10 (2).

Rochet, J. C., and J. Tirole. 1996. Interbank lending and systemic risk. *Journal of Money, Credit and Banking* 28 (4): 733–762.

Rolnick, A., and W. Weber. 1983. New evidence on the free banking era. *American Economic Review* 73: 1080–1091.

———. 1984. The causes of free bank failures. *Journal of Monetary Economics* 14: 267–291.

Ronn, E. I., and A. K. Verma. 1986. Pricing risk-adjusted deposit insurance: An option-based model. *Journal of Finance* 41: 871–895.

Rothschild, M., and J. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics* 90 (4): 630–649.

Salop, S. 1979. Monopolistic competition with outside goods. *Bell Journal of Economics* 10 (1): 141–156.

Sargent, T., and N. Wallace. 1982. The real bills doctrine vs the quantity theory: A reconsideration. *Journal of Political Economy* 90: 1212–1236.

Selgin, G. 1988. *The theory of free banking*. Totowa, N.J.: Rowan and Littlefield.

Sironi, A. 2000. Testing for market discipline in the European banking industry: Evidence from subordinated debt issues. Working paper, Bocconi University, Milan, Italy.

Slovin, M., M. E. Sushka, and J. Polonchek. 1999. An analysis of contagion and competitive effects at commercial banks. *Journal of Financial Economics* 54 (2): 197–225.

Smith, V. 1984. *The rationale of central banking and the free banking alternative*. Indianapolis: Liberty Press.

Suarez, J. 1993a. Banking regulation in an equilibrium model. Discussion paper 9308, Centro de Estudios Monetarios y Financieros, Madrid.

———. 1993b. Closure rules, market power and risk-taking in a dynamic model of bank behaviour. Discussion paper, Universidad Carlos III, Madrid.

Tirole, J. 1994. On banking and intermediation. *European Economic Review* 38 (3-4): 469–487.

Tobin, J. 1986. Financial innovation and deregulation in perspective. In *Proceedings of the Second International Conference of the Institute for Monetary and Economic Studies of the Bank of Japan*, ed. Y. Suzuki and H. Yomo. Tokyo: University of Tokyo Press.

Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21 (2): 265–293.

White, L. 1984. *Free banking in Britain: Theory, experience and debate 1800–1845*. Cambridge: Cambridge University Press.

Williamson, S. D. 1992. Pricing free bank notes. Discussion paper, Wharton School, University of Pennsylvania, Philadelphia.

# Index