and

$$\frac{C_n}{n} = E[\max(R_D^n + K - \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i, 0)]. \tag{2.23}$$

The strong law of large numbers dictates that $\frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i$ converges almost surely to $E(\tilde{y})$. Since $E(\tilde{y}) > K + R$, relation 2.22 shows that $\lim_n R_D^n = R$ (i.e., deposits are asymptotically riskless). Therefore, by 2.23

$$\lim_n \frac{C_n}{n} = \max(R + K - E(\tilde{y}), 0) = 0.$$

■

Several criticisms have been expressed regarding Diamond's theory. For instance, the use of nonpecuniary penalties that have to be modulated according to the cash flow reported by the borrower is not realistic: real-world nonpecuniary penalties (such as jail or the loss of reputation associated with failure) are more lump-sum. However, the delegated monitoring approach can also work in a context of costly state verification à la Townsend-Gale-Hellwig, in which the optimal contract is also a standard debt contract[19] but the deadweight loss C comes from auditing in case of low realizations of the cash flow.

Krasa and Villamil (1992) construct a model in which only monitoring ("auditing" in the terminology of Gale and Hellwig) 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, since only the bank will have to monitor the firms.

Another interesting contribution that builds on Diamond's (1984) model of delegated monitoring is the article by Cerasi and Daltung (1994), 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 (see the appendix to Chapter 3), 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, since 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).

2.5 Coexistence of Direct and Intermediated Lending

So far, the discussion has centered on why FIs exist, and it has therefore focused on the "uniqueness of bank loans" (to recall the title of the important

1987 article by James, who shows in particular that financial markets tend to react positively when they learn that a quoted firm has obtained a bank loan). However, direct finance 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 will analyze 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 will study two models that explain the coexistence of the two types of finance. Both are based on moral hazard, which prevents firms without enough assets from obtaining direct finance. These assets can be monetary, as in the Holmström-Tirole (1993) model (covered in subsection 2.5.3), or reputational, as in the Diamond (1991) model, (presented in subsection 2.5.2). Subsection 2.5.4 will give a brief account of several related contributions. Subsection 2.5.1 will present a simple model of the credit market with moral hazard that will be used in several chapters of this book.

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 one. The riskless rate of interest is normalized to zero. Firms have a choice between a "good" technology, which produces G with probability π_G (and zero otherwise), and a "bad" technology, which produces B with probability π_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 or the return. Therefore 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. Notice that $R_C < G < B$. From the lender's viewpoint, the probability p of repayment therefore depends on R:

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

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

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

Because of the assumptions that have been made, this is only possible when $\pi_G R_C > 1$, which means that moral hazard is not too important. If $\pi_G R_C < 1$, the equilibrium involves no trade, and the credit market collapses (this is because bad projects have a negative NPV).

Now a monitoring technology is introduced: at a cost C, 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:

1. The nominal value R_m of 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.

2. 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 $\pi_G (\pi_G \in \left[\frac{1+C}{G}, \frac{1}{R_C}\right])$ 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 $\pi_G \in [\frac{1+C}{G}, \frac{1}{R_C}]$ (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 (Adapted from Diamond 1991)

The objective will be to show, in a dynamic version 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 heterogenous: only some fraction f of them strategically choose between the two technologies. The rest have access only to a given technology (for example, the bad one). This discussion will show that 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.

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 π_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's 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 $R_S = \frac{1}{\pi_S}$. 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) also choose bank lending.

The symbol π_0 denotes the unconditional probability of success at $t = 0$ (when strategic firms choose the good technology):

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

The notion of reputation building comes from the fact that $\pi_U < \pi_0 < \pi_S$, i.e., the probability of repayment of the bank loan by the firm is initially π_0, but increases if the firm is successful (π_S) and decreases in the other case (π_U). Because of that, the critical level of debt (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

$$\pi_B(B - R) + \delta\pi_G[G - \pi_B R_S - (1 - \pi_B)R_U]$$
$$= \pi_G(G - R) + \delta\pi_G[G - \pi_G R_S - (1 - \pi_G)R_U],$$

which gives

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

The following is the complete result:

Result 2.6 Under the following assumptions:

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

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 rate $R_0 = \frac{1+C}{\pi_0}$.
2. At $t = 1$, successful firms issue direct debt at rate $R_S = \frac{1}{\pi_S}$, whereas the rest borrow from banks at rate $R_u = \frac{1+C}{\pi_U}$.

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

- Firms with a good reputation can issue direct debt.[20]
- Unsuccessful firms pay a higher rate than new firms ($R_U > R_0$).
- Moral hazard is partially alleviated by reputation effects ($R_C^0 > R_C$).

Proof It is necessary only to apply repeatedly result 2.5, after adjusting the parameters for the different cases:

Part 1 in 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 comes from parts 1 and 2 of result 2.5, since by assumption $\pi_G^S > \frac{1}{R_C}$ and $\frac{1+C}{G} < \pi_U < \pi_0 \leq \frac{1}{R_C^0} < \frac{1}{R_C}$. ■

2.5.3 Monitoring and Capital (Adapted from Holmström and Tirole 1993)

Holmström and Tirole (1993) consider a simple model that elegantly captures the notion of a substitutability between capital and monitoring, both at 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 assumption is the opposite of that of 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 (i.e., banks), represented by the index m; and (3) uninformed investors (depositors), represented by the index u. Each industrial project (owned by firms) costs I and returns R (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 Δp). Bad projects give a private benefit to the borrower: this is the source of moral hazard.[21] Monitoring the firm (which costs C) implies a reduction of this benefit from B (without monitoring) to b (with monitoring). Investors are risk neutral, are uninformed (i.e., they are not able to monitor firms), and have access to an alternative investment of gross expected return γ. It is assumed that only the good project has a positive Net Expected Present Value, even if the private benefit of the firm is included:

$$p_H R - \gamma I > 0 > p_L R - \gamma I + B.$$

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 will determine the total lending capacity of the banking industry. The following paragraphs examine the 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. As usual, the incentive compatibility constraint gives an upper bound on R_u:

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

Now the individual rationality constraint of uninformed investors implies an upper bound on I_u: