The Development and Structure of Financial Systems*

Shankha Chakraborty† Tridip Ray‡

Revised: September 2006

Abstract

Firms raise external finance via monitored bank loans and non-monitored borrowing in a dynamic general equilibrium model. Access to credit and each type of financing depend on the wealth distribution due to moral hazard. We study the depth of credit markets (financial development) and conditions under which the financial system relies more on either type of external finance (financial structure). Initial inequality, investment size and institutional factors determine the level of financial development, while financial structure is shaped by the investment technology and legal and financial institutions. The model's predictions are consistent with historical and recent development experience.

Keywords: Financial Development, Financial Structure, Bank Finance, Market Finance, Credit Frictions.

JEL Classification: E44, G20, G30, O15, O16

*This paper was written when the second author was at the Hong Kong University of Science and Technology. Financial support from Hong Kong RGC Competitive Earmarked Research Grant 2001-02 (Project No. HKUST6073/01H) is gratefully acknowledged. For helpful discussions and suggestions we thank Joydeep Bhattacharya, Nancy Chau, Satya P. Das, Sudipto Dasgupta, Claudian Kwok, Raoul Minetti, Kunal Sengupta, Steve Turnovsky, Yong Wang and seminar participants at various places this paper was presented. We are especially grateful to the co-editor, Carl Chiarella, an associate editor and two anonymous referees of this journal for comments and suggestions. All remaining errors are ours.

†Department of Economics, University of Oregon, Eugene, OR 97403-1285. Email: shankhac@uoregon.edu

‡Department of Economics (Planning Unit), Indian Statistical Institute, 7, S.J.S. Sansanwal Marg, Qutab Institutional Area, New Delhi 110016, India. Email: tridip@isid.ac.in

1 Introduction

Systematic evidence over the last decade has documented a robust and positive relationship between finance and economic development.[1] This positive relationship motivates our paper. But our interest goes deeper, into the roles various participants play in the financial system. We begin with the century-old debate on the efficacy of banks versus financial markets. Some commentators have cited Germany and Japan's experience to argue that banks are better at mobilizing savings, identifying good investment and exercising corporate control. Observers on the other side of the debate point to the UK and US as evidence that financial markets have an advantage over banks in information acquisition, corporate control and risk management.

This debate is as relevant today as it was historically. A global trend, favoring financial markets, has emerged over the last two decades. This is true even in traditionally bank-based developed countries like France and Japan which have increased the role of financial markets since the mid-eighties (Allen and Gale, 2000). In developing and transition countries, market-finance is finding increasing favor as banking crises have become widespread and as government interventions, particularly in the banking sector, have been thoroughly discredited (Allen and Gale, 2000, Holmstrom, 1996, Demirgüc-Kunt and Levine, 2001a).

Yet, as Allen and Gale (2000) point out, it is unclear why financial markets are suddenly seen as the panacea. Missing from the global policy debate is a clear understanding of the dynamic implications of each type of financial system and, more specifically, how the financial system itself evolves. So our primary goal is to write down and analyze an explicit dynamic model of financial structure. We see this as a necessary step to addressing policy issues raised by the bank-versus-market debate, especially in developing and transition economies.

What We Do

The benefits of a financial system depend on the degree to which external finance facilitates industrialization. But in a second-best world infested with credit frictions, access to credit is constrained by wealth levels and internal asset positions of individuals and firms. There is thus an intimate connection between the wealth distribution and financial development. We construct a dynamic model that incorporates this interdependence.

In the model, manufacturing requires large-scale investment which cannot be funded by internal assets only. Potential entrepreneurs may borrow using monitored bank loans (bank finance) or non-monitored sources like bonds and equities (market finance) or a combination

[1] See Levine (2005) for an up-to-date survey of this literature.

of the two. Credit frictions arise because owner-managers of manufacturing firms may choose an inferior technology in order to enjoy private benefits. The incentive to do so is greater the lower the personal stake an owner has in her investment project (Holmstrom and Tirole, 1997).

Under this incentive problem, wealth thresholds determine who invests and the kind of financial instruments they use. Poorer individuals do not obtain any funding since they cannot guarantee lenders the required return; they instead work for a wage. Others obtain loans and produce capital using a risky technology. When investment succeeds, these capitalists hire workers to operate the machinery and produce a final good. Among capitalists, wealth thresholds again determine how they borrow. Individuals of medium wealth levels are able to borrow only through a combination of intermediated (bank) and unintermediated (market) finance. Bank finance entails monitoring that partially eliminates the incentive problem; perceiving this, direct lenders are willing to lend.[2] Capitalists who are wealthy enough, on the other hand, do not have to be monitored and use only market finance.

In this way the wealth distribution determines access to credit markets (financial depth) and dependence on each type of external finance (financial structure). The wage earners who were rationed out of the credit market may ultimately accumulate enough assets to become capitalists. Whether or not this happens depends on their labor income, which in turn depends on the extent of industrialization. At the same time today's capitalists may find themselves denied access to credit in the future if they suffer losses on their current investment.

Main Findings

The dynamic interplay of the wealth distribution and borrowing choices determines the path of financial development and structure. We identify initial inequality, investment size and institutional factors as key determinants of financial development, while financial structure is shaped by investment technology and the nature of its legal and financial institutions.

An unequal distribution, specifically a large peasant population relative to capitalists, hurts financial development since few individuals obtain credit. The sparse industrialization that results from this prevents workers from accumulating enough to access credit markets in the future. Low to moderate degrees of inequality, however, see the emergence of developed financial systems. Even then, financial depth is negatively related to inequality, a prediction consistent with evidence from a cross-section of countries. Initial inequality also impacts an economy's historical financial structure – by reducing dependence on monitored finance

[2] Costly monitoring makes bank finance a more expensive, but necessary, alternative to market finance.

it results in more market-oriented financial systems, a pattern we observed during Western Europe's industrialization.

The investment technology plays a vital role. When investment requirements are too large relative to average wealth levels, fewer individuals obtain credit: industrialization and financial development remain low. Larger capital requirements also promote a bank-based financial system, at least during the initial stages of financial development. This outcome matches well Western Europe's historical experience. In particular, the British industrial revolution occurred in industries, textiles for example, that did not call for enormous investments (Landes, 1969). Germany, in contrast, was involved in heavy manufacturing and chemicals, both of which required large injections of external capital. Such technological differences could have played a key role in Britain's historical reliance on market finance and Germany's dependence upon its banking systems (Baliga and Polak, 2004).

We show that investment technologies that are riskier are more conducive to market-based systems because a larger proportion of capitalists can accumulate enough assets to shed their reliance on bank intermediation. Finally, institutional factors like agency costs shape a financial system in intuitively plausible ways. Bank-based systems result when bank monitoring is particularly efficient in resolving agency problems, although it depends upon monitoring costs as well. Our theoretical results lend credence to recent empirical evidence, LaPorta et al. (1997, 1998) and Levine (2002) in particular, that shows institutional and legal factors to be important determinants of financial structure.

Recent Related Literature

This paper contributes to the emerging literature pioneered by Banerjee and Newman (1993) and Galor and Zeira (1993) on the dynamic link between credit frictions and wealth distribution.[3] Similar to some of these papers, an important feature of our model is the dependence of factor prices on wealth distribution. In general, such dependence can give rise to complicated non-linear dynamics (see for example Aghion and Bolton (1997)). An advantage of our framework is its ability to circumvent this problem and flesh out interesting and empirically plausible predictions regarding financial development.

The key innovation we bring to this literature is a rich financial structure and a fuller understanding of the determinants of financial development. Treatment of financial systems in the existing literature is incomplete since its main goal has been to characterize distributional dynamics or study occupational choices. What is missing, specifically, is the variety

[3]Recent work in this area include Aghion and Bolton (1997), Piketty (1997) and Mookherjee and Ray (2002).

of financing choices that firms typically face. In contrast, our interest lies first and foremost in these choices and how they shape the development and structure of financial systems.[4]

At the other end of the spectrum lies the corporate finance literature on firm financing choices. This literature usually deals with static (often partial equilibrium) models of *developed* financial systems. Conclusions about *developing* societies are hard to infer, even though financial reforms in developing countries have borrowed extensively from the experience of developed systems.

We draw insights from Diamond (1991) and Holmstrom and Tirole (1997) which analyze the link between firm financing choices and some form of asset distribution. Diamond (1991) considers how a firm switches from expensive (and monitored) bank finance to cheaper forms such as public debt as it develops a better reputation, which of course is a form of asset. Holmstrom and Tirole (1997), whose incentive structure we adopt here, observe how incentive problems and access to different types of external finance depend on a firm's internal assets. Neither of these papers incorporate the feedback that macro-fundamentals have on financing choices. Once these are taken into account, we show how the typical life-cycle story of firm financing choices, as articulated by the finance literature, can fall apart unless initial conditions and policy parameters are appropriate for long-run financial development.

Finally, our paper is related to Baliga and Polak (2004) who analyze a static partial-equilibrium model in which all firms borrow using either monitored or non-monitored debt (but not both, as in this paper). They discuss how investment size and entrepreneurial wealth explain the historical development of the Anglo-Saxon and German financial systems. Our dynamic model shows that the relationship between firm financing choices and investment size (wealth) posited by Baliga and Polak is temporary, relevant only during the initial stages of development. The relationship disappears in the long-run unless investment size (wealth) gives rise to history-dependence, in which case the very process of financial development can be crippled.

The paper is organized as follows. The model is developed in section 2 and optimal

[4] Greenwood and Smith (1997) study a dynamic model of banks and equity markets. They are interested in how these two types of external financing emerge endogenously to facilitate growth through liquidity provision and specialization. The key differences with our paper are that banks exist here to address agency problems and our focus in on how the mix of two types of external finance – what we call financial structure – changes over time. We are also primarily interested in the demand for various types of external finance (the supply side is a sideshow) and abstract from growth issues (see Chakraborty and Ray, 2006).

Turnovsky (2000, chapter 9) discusses some early papers, focusing on the short-run, that incorporate the Modigliani-Miller theorem in a macro-environment.

financial contracts characterized in section 3. Section 4 looks at the static general equilibrium while section 5 analyzes the dynamics and its implications for financial development and structure. Section 6 contains further discussions on what our analysis adds to the literature and to the policy debate. Section 7 concludes.

2 The Model

Consider a small open economy populated by a continuum of agents of measure one. Time is continuous and successive generations are connected by a bequest motive.

An agent is born with an initial wealth, a, received as bequest from her parent and a labor time endowment of one unit. This labor can be either supplied inelastically to the labor market or used to oversee an investment project that produces capital. Inheritance is the sole source of heterogeneity among newly borns. We denote the cumulative distribution of agents at t by $G_t(a)$ and assume that the initial distribution G_0 is continuous and differentiable.

Preferences are given by the "warm-glow" utility function:

$$u_t = c_t^{\beta} b_t^{1-\beta}, \quad \beta \in (0,1),$$

where c denotes consumption and b denotes bequest left to offspring. Given a realized income z, optimal consumption and bequests are linear functions of z:

$$c_t = \beta z_t, \quad b_t = (1-\beta) z_t. \tag{1}$$

The indirect utility function is then also linear in income, $U_t = \varphi z_t$ with $\varphi \equiv \beta^{\beta}(1-\beta)^{1-\beta}$, implying that agents are risk-neutral.

Similar to Banerjee and Newman (1993), newly born agents become economically active only when they become 'mature'. Time to maturity, T, is distributed exponentially with the density function $h(T) = \eta e^{-\eta T}$, $\eta > 0$, across members of the same cohort. All economic activity occurs at the instant an agent becomes mature: she chooses her occupation and earns income accordingly, gives birth to one offspring, consumes, leaves bequests and dies. There is thus no population growth and members of a cohort do not all die at the same time. Without loss of generality we set $\eta = 1$ so that agents live for a unit length of time on average.

2.1 Production and Occupation

Whether or not an individual is a worker or a capitalist is determined by access to external finance. Production of capital requires an indivisible investment of size q. Only individuals

able to raise the requisite funds (from internal and external sources) become capitalists, the rest join the labor force.

A worker supplies her unit labor endowment to the labor market, earning a wage income w_t. A capitalist's income, on the other hand, is uncertain. In particular the investment project is risky – a successful project yields capital amounting to θq ($\theta > 1$), while failure yields nothing. Successful capitalists become producers of final goods by hiring workers to operate the capital. This capital depreciates completely upon use.

Markets for the final good and for labor are perfectly competitive. We assume an Arrow-Romer type technological spillover in the final goods sector. Specifically, for a successful capitalist j, the *private* technology for producing the unique consumption good is constant returns to scale in private inputs:

$$Y_t^j = \left(K_t^j\right)^{\alpha} \left(A_t N_t^j\right)^{1-\alpha}, \ \alpha \in (0,1), \tag{2}$$

Here A_t denotes time-dependent labor-efficiency that is common to all final goods producers. Labor efficiency A depends upon capital per worker, k, through a learning-by-doing externality:

$$A_t = \hat{A} k_t. \tag{3}$$

Productivity improvements in any particular firm spills over instantaneously to the rest of the economy, becoming public knowledge. The *social* (intensive-form) production function is thus an Ak type technology $y_t = Ak_t$, where $A \equiv \hat{A}^{1-\alpha}$.

It remains to characterize the investment decision facing a potential capitalist. An individual with assets $a_t < q$ can become a capitalist only if she is able to borrow the shortfall $q - a_t$. To obtain a rich financial structure, we introduce an agency problem similar to that in Holmstrom (1996) and Holmstrom and Tirole (1997).

Specifically, the probability of success of investment depends upon an unobserved action taken by the capitalist – her choice on how to spend q. She can spend it on an efficient technology that yields θq units of capital with probability π_G, but uses up all of q. Alternatively, she can spend it on one of two inefficient technologies. One of these technologies is a low moral hazard project, costing $q - vq$, leaving vq for the capitalist to appropriate. This project too yields θq units of capital when it succeeds, but it succeeds less often, with probability $\pi_B < \pi_G$. The other inefficient technology is a high moral-hazard project, costing $q - Vq$. This leaves Vq in private benefits.

Both inefficient technologies carry the same probability of success, π_B, but since $0 < v < V < 1$ by assumption, the capitalist would prefer the high moral-hazard project over the

low moral-hazard one. Only the efficient technology is, however, economically viable.[5] The table below summarizes these investment choices.

TABLE 1

PROJECT	GOOD	LOW MORAL HAZARD (low private benefit)	HIGH MORAL HAZARD (high private benefit)
Private Benefit	0	vq	Vq
Success Probability	π_G	π_B	π_B

3 The Financial Sector

Capital is perfectly mobile across borders so that this small open economy has free access to the international capital market. The time-invariant (gross) world rate of return on investment, r^*, is taken as given.

Supply of loans in the domestic financial sector comes from two sources: through financial intermediaries or banks, and directly from workers and international investors. Workers are indifferent between bank deposits, lending directly to capitalists and investing on the international capital market as long as all three yield an expected return of r^*. In other words, r^* is the return that banks promise their depositors as well as the expected return on direct lending.[6]

On the demand side, loans are obtained by individuals who invest in the production of capital; they invest their entire wealth, borrowing the remainder from the domestic financial sector. Credit-constrained agents work for the capitalists. They deposit their wealth with banks or lend directly to domestic capitalists or the international capital market.

Capitalists face a perfectly elastic supply curve of loanable funds since they can freely access the international capital market. Availability of domestic investable resources therefore does not concern us. What matters is *how* capitalists borrow. Direct borrowing from

[5] To ensure this we assume $\pi_G \alpha A\theta q - r^* q > 0 > \pi_B \alpha A\theta q - r^* q + Vq$. Here r^* is the world return on investment and we anticipate that a successful capitalist's return per unit capital produced, ρ_t, equals αA in equilibrium (see equation (11) below).

[6] Note that we do not allow direct investment (FDI) by foreign investors. Implicitly we are assuming TFP differences between the domestic and foreign economies ($A^* > A$) because of which these investors earn higher returns from directly investing in their home countries. There may be other reasons too. For example, foreign direct investors may face setup costs which domestic capitalists do not, and realistically they may also face expropriation risks or problems with repatriating profits.

domestic (workers) and foreign investors will be referred to as *direct* (or *market*) finance, and should be thought of as occurring through the purchase of one-period corporate bonds and equities. Borrowings intermediated by the banking sector will be called *indirect* (or *bank*) finance.

Bank finance plays a specific role. Banks have a monitoring technology that allows them to inspect a borrowing capitalist's activities and ensure that she conforms to the terms agreed upon in the financial contract (Hellwig, 1991; Holmstrom and Tirole, 1997). Direct lenders (workers and foreign investors) do not possess this technology. Thus, as in Diamond (1984, 1991), banks are the delegated monitors.

Bank monitoring partially resolves the moral hazard problem and reduces a capitalist's opportunity cost of being diligent. By choosing to monitor borrowers, banks are able to eliminate the high moral-hazard project but not the low moral-hazard one. For instance, a bank could stipulate conditions that prevent the firm from implementing the high moral-hazard project when it negotiates a loan contract with the bank. But such monitoring is privately costly for the bank and requires it to spend a nonverifiable amount γ per unit invested by the capitalist. Evidently, bank monitoring will be an optimal arrangement only if the gains from resolving the incentive problem is commensurate with monitoring costs.

3.1 Optimal Contracts

Faced with the incentive problem outlined above, a capitalist will behave diligently to the extent that she receives an incentive-compatible expected payoff and whether or not she is monitored.

Consider the financing options a prospective capitalist faces in borrowing from banks or from the market. Since banks monitor firms while outside investors do not, we shall refer to the former as *informed* investors. We consider optimal contracts that induce capitalists to invest in the good project.

Direct Finance

An optimal contract between a capitalist and direct financiers has a simple structure. Capitalist-i invests her entire wealth, a_t^i, in her own project since that yields a return strictly higher than she would otherwise obtain. Direct lenders provide the remaining, $q - a_t^i$. Neither party is paid anything if the investment fails. When the project succeeds, the capitalist earns an amount $x_t^C > 0$ while uninformed investors are paid $x_t^U > 0$. Denote a successful capitalist's rate of return per unit capital produced by ρ_t. Since a successful project produces θq units

of capital, we have $x_t^C + x_t^U = \rho_t \theta q$.[7]

In order to invest in the good project, capitalist-i must earn an incentive compatible expected income. Moreover, the contract should satisfy each lender's participation constraint, that is, lenders should be guaranteed at least as much as they would earn on the international capital market. Combining these two constraints, we can show (see Appendix A.1 for details) that only capitalists with wealth exceeding $\overline{a}_t$ would be able to obtain direct finance, where

$$\overline{a}_t \equiv \frac{q}{r^*}\left[\frac{\pi_G}{\pi_G - \pi_B}V - \{\pi_G \rho_t \theta - r^*\}\right]. \tag{4}$$

Indirect Finance

Indirect or *intermediated* finance entails three parties to the contract: the bank, uninformed investors and the capitalist. An optimal contract here too stipulates that no one earns anything when the project fails. In case of success, total returns, $\rho_t \theta q$, are divided up as $x_t^C + x_t^U + x_t^B = \rho_t \theta q$, with x_t^B denoting the bank's returns.

Besides the incentive compatibility constraint of the capitalist and the participation constraint of the uninformed investors, we have to take into account an additional incentive compatibility constraint, that for bank monitoring. Moreover the loan size has to be chosen optimally to maximize bank profits subject to the capitalist's incentive constraint and the bank's incentive and resource constraints. Finally, the banking sector earns zero profits in a competitive equilibrium.

Together these have the following implications (see Appendix A.1): (i) bank finance is relatively more expensive than direct finance (due to monitoring costs), that is, the (gross) return on bank loans, r_t^L, is greater than r^*/π_G, the return promised to direct lenders in case of success

$$r_t^L = \frac{r^*}{\pi_B} > \frac{r^*}{\pi_G}, \tag{5}$$

and (ii) capitalists with wealth $\overline{a}_t > a_t^i \geq \underline{a}_t$, where

$$\underline{a}_t \equiv \frac{q}{r^*}\left[\frac{\pi_G v}{\pi_G - \pi_B} - \{\pi_G \rho_t \theta - (1+\gamma) r^*\}\right], \tag{6}$$

[7]Since project returns are observable and verifiable, optimal contracts between direct financiers and capitalists may be interpreted either as debt or as outside equity. For an equity contract, the capitalist sells a share s_t of her project return, $x_t^U = s_t(\rho_t \theta q)$. For a debt contract, the capitalist borrows $q - a_t$, promising to repay a return of r^*/π_G in case of success. The implicit return on equity when projects succeed has to be r^*/π_G for both assets to be held simultaneously, that is, $s_t(\rho_t \theta q) = r^*(q - a_t)/\pi_G$. Again, what matters is that neither of these is monitored lending.

are able to convince uninformed investors to supply the remaining funds for the investment project only after the bank lends an amount (and agrees to monitor)[8]

$$l_t^i = \gamma \left(\frac{\pi_B}{\pi_G - \pi_B} \right) q. \tag{7}$$

Capitalists with wealth level below $\underline{a}_t$ cannot obtain any external finance, direct or indirect.

3.2 Occupational Incomes

Denote individual-i's income by z_t^i. Consider the case where i does not obtain any external finance since her wealth is too small, $a_t^i \leq \underline{a}_t$. Her income consists of labor earnings and returns on investment in the domestic and/or international capital market

$$z_t^i = w_t + r^* a_t^i.$$

Next consider those who borrow both from banks and the market, that is, using *mixed* finance. For these capitalists with $a_t^i \in [\underline{a}_t, \overline{a}_t)$, equations (5) and (7) imply that income from a successful project would be

$$z_t^i = \rho_t \theta q - r_t^L l_t^i - \hat{r}^* [q - l_t^i - a_t^i] = [\rho_t \theta - (1+\gamma)\hat{r}^*]\, q + \hat{r}^* a_t^i,$$

where $\hat{r}^* \equiv r^*/\pi_G$. Failure gives them zero returns.

Finally, capitalists with adequate wealth, $a_t^i \geq \overline{a}_t$, borrow only from the market and earn

$$z_t^i = \rho_t \theta q - \hat{r}^* \left[q - a_t^i \right] = (\rho_t \theta - \hat{r}^*) q + \hat{r}^* a_t^i,$$

from a successful project. Of course, we are assuming that the rate of return from the project ($\rho_t \theta$) is high enough for the capitalist's participation constraint to be satisfied. That is, we require that a capitalist's expected income $\pi_G z_t^i$ is greater than $r^* a_t^i$, what she could earn for sure by investing her entire wealth on the domestic and international capital markets. This will be true under appropriate restrictions on θ and the final goods technology (α, A).

[8] In order that the loan size does not exceed investment size, that is $l_t^i \leq q$, monitoring costs should not be so high as to make it impossible for bank intermediation to resolve agency problems. Hence, we restrict monitoring cost to $\gamma \leq (\pi_G - \pi_B)/\pi_B$.

It is also natural to assume that $\overline{a}_t > \underline{a}_t$, or else there will be no demand for intermediation since monitoring would be too costly to be socially useful. This is true as long as the expected gain from monitoring exceeds its cost: $\pi_G(V - v)/(\pi_G - \pi_B) \geq \gamma r^*$.

Earnings for each type of economic agent are thus given by

$$z_t^i(a_t^i) = \begin{cases} w_t + r^* a_t^i, & \text{for } a_t^i \in [0, \underline{a}_t) \\ \left.\begin{array}{l} [\rho_t \theta - (1+\gamma)\hat{r}^*]\, q + \hat{r}^* a_t^i, \text{ with prob. } \pi_G \\ 0, \text{ otherwise} \end{array}\right\} & \text{for } a_t^i \in [\underline{a}_t, \overline{a}_t) \\ \left.\begin{array}{l} (\rho_t \theta - \hat{r}^*) q + \hat{r}^* a_t^i, \text{ with prob. } \pi_G \\ 0, \text{ otherwise} \end{array}\right\} & \text{for } a_t^i \in [\overline{a}_t, \infty) \end{cases} \tag{8}$$

To summarize properties of optimal loan contracts and external financing choices, we note that given $q > a_t^i$ and the wealth distribution G_t,

(i) individuals with $a_t^i < \underline{a}_t$ are unable to obtain any external finance and work as laborers;

(ii) individuals with $a_t^i \in [\underline{a}_t, \overline{a}_t)$ obtain external finance from banks as well as households: they borrow an amount l_t^i from banks at the loan rate r_t^L, given by (5) and (7) above, agree to being monitored, and raise the remaining $(q - l_t^i - a_t^i)$ directly from investors at the rate $\hat{r}^*$; optimal contracts guarantee these capitalists incentive compatible payments such that they behave diligently;

(iii) individuals with $a_t^i \geq \overline{a}_t$ borrow only from investors, paying them a return of $\hat{r}^*$; here too, incentive compatible payments to these capitalists ensure that investments occur in the good project; and

(iv) income in each case is given by (8) above.

4 General Equilibrium in Period t

Figure 1 outlines the decisions facing a representative agent ('firm') born with assets a. Once she becomes economically active, the agent makes occupational and financing decisions depending on how her assets compare to the wealth thresholds $\underline{a}$ and $\bar{a}$. Individuals are sorted into three categories: those who work (no credit access), those who become capitalists using intermediated and unintermediated loans (*mixed-finance* capitalists), and those who become capitalists by borrowing solely from the market (*market-finance* capitalists). Once investment outcomes are realized and final goods are produced, occupational incomes are determined according to (8). Wealth transfers are subsequently made to offsprings ('continuing firms') as specified by (1).

Parametric assumptions we make below ensure that workers earn a strictly lower income than either type of capitalist. Moreover, market-finance capitalists earn a higher expected income than mixed-finance capitalists (see equation (8)). Which of the three occupations an individual falls into is hence solely determined by her wealth. If wealth were not a constraint, all agents would want to become market-finance capitalists.

Consider the economy at time t. Denote the fractions of the three types of agents by (f_{1t}, f_{2t}, f_{3t}), where

$$f_{1t} = G_t(\underline{a}_t),\ f_{2t} = G_t(\overline{a}_t) - G_t(\underline{a}_t),\ f_{3t} = 1 - G_t(\overline{a}_t).$$

There are f_{1t} workers and $1 - f_{1t}$ capitalists at any instant t. Given the law of large numbers, π_G proportion of these capitalists succeed in producing capital, amounting to $K_t^j = \theta q$ each, where j denotes a successful capitalist. The aggregate capital stock is then $K_t = \pi_G \theta q(1 - f_{1t})$ and the workforce $N_t = f_{1t}$. Capital per worker is, thus,

$$k_t = \pi_G \theta q \left[\frac{1 - f_{1t}}{f_{1t}}\right]. \tag{9}$$

Since all successful capitalists produce the same amount of capital, given w_t, they hire the same number of workers

$$N_t^j = \frac{f_{1t}}{\pi_G(1 - f_{1t})}.$$

Note the private technology (2). In equilibrium, substituting for the labor augmenting technological progress ((3) and (9)) into this production function gives output produced by a successful capitalist as $Y_t^j = A\theta q$.

Under competitive markets, the equilibrium wage rate is given by the private marginal product of capital,

$$w_t = (1 - \alpha)Ak_t = (1 - \alpha)\pi_G A\theta q \left[\frac{1 - f_{1t}}{f_{1t}}\right]. \tag{10}$$

A successful capitalist then earns the income $\widehat{Y}_t^j = \alpha A\theta q$, net of wage payments $w_t N_t^j$, from her capital θq. The (gross) rate of return on capital, which we defined as ρ_t above, is clearly equal to αA, the private marginal product of capital, that is,

$$\rho_t = \alpha A. \tag{11}$$

Due to overall constant returns to capital, this return is time-invariant. Since all successful capitalists earn the same return on capital, we assume, without loss of generality, that they produce final goods using only their own capital.

Using the equilibrium return on capital from (11) the cutoff wealth levels defined by (6) and (4) now do not depend upon time:

$$\underline{a} = \delta_1 q, \quad \overline{a} = \delta_2 q, \tag{12}$$

where[9]

$$\begin{aligned} \delta_1 &\equiv [\upsilon \pi_G/(\pi_G - \pi_B) - \{\pi_G \alpha \theta A - (1+\gamma) r^*\}]/r^*, \\ \delta_2 &\equiv [V \pi_G/(\pi_G - \pi_B) - \{\pi_G \alpha \theta A - r^*\}]/r^*. \end{aligned}$$

It remains to check whether or not a worker earns lower income than a capitalist. This is by no means guaranteed. For instance, when there are "too few" workers, the marginal product of labor may be so high that even individuals who could have obtained external finance choose to work. It turns out that this happens when the proportion of credit-constrained agents falls below $\widetilde{f}_1$, where $\widetilde{f}_1$ satisfies

$$(1-\alpha)\pi_G \theta A \left[\frac{1-\widetilde{f}_1}{\widetilde{f}_1}\right] = \alpha \pi_G \theta A - (1+\gamma) r^*.$$

We restrict ourselves to empirically plausible distributions, those that are positively skewed. We assume hence that G_0 satisfies $f_{10} > \widetilde{f}_1$. This ensures that occupational "choice" is stable over time and we can simply focus on the proportions of the three types of agents without having to worry about the effect of income on occupational dynamics.

5 Dynamics of Financial Development and Structure

The financial system, by which we mean the degree to which an economy relies upon external finance in general, and bank and market-finance in particular, is determined by access to credit markets. Drawing upon the instantaneous equilibrium from the previous section, we now consider the evolution of this financial system.

Given an initial wealth distribution G_0, wealth thresholds $\underline{a}$ and $\overline{a}$ determine the proportion of individuals able to borrow and the relative composition of bank- and market-finance in aggregate investment. These investment choices lead to income realizations that determine the subsequent distribution through bequests. The process continues recursively, with changes in the financial system tracking the wealth distribution through time.

[9] Parametric restrictions in footnotes 5 and 8 ensure that $\delta_1 < \delta_2 < 1$.

Substituting labor and capital's equilibrium returns into (8), and using optimal bequests (1), we obtain the intergenerational law of motion:

$$b_t = \begin{cases} (1-\beta)\left[r^* a_t + (1-\alpha)\pi_G A\theta q\left(\frac{1-f_{1t}}{f_{1t}}\right)\right], & \text{for } a_t \in [0, \underline{a}) \\ \left.\begin{array}{l}(1-\beta)\left[\hat{r}^* a_t + \{\alpha A\theta - (1+\gamma)\hat{r}^*\}q\right], \text{ with prob. } \pi_G \\ 0, \text{ otherwise}\end{array}\right\} & \text{for } a_t \in [\underline{a}, \overline{a}) \\ \left.\begin{array}{l}(1-\beta)\left[\hat{r}^* a_t + (\alpha A\theta - \hat{r}^*)q\right], \text{ with prob. } \pi_G \\ 0, \text{ otherwise}\end{array}\right\} & \text{for } a_t \in [\overline{a}, \infty) \end{cases} \tag{13}$$

Figure 2 depicts this wealth dynamics for various possibilities (see below); the dotted lines represent expected income from investment.

Note that all three regimes of (13) are piecewise linear. For this mapping to converge we need to rule out the possibility that a dynasty can get arbitrary rich over time by simply reinvesting its wealth. This is ensured by $(1-\beta)\hat{r}^* < 1$.

We would also like to rule out a dynasty from being able to self-finance its entire investment. When investment succeeds, the fixed-point of the mapping for $a_t \in [\overline{a}, \infty)$ is given by $a^U = (1-\beta)(\alpha A\theta - \hat{r}^*)q/[1-(1-\beta)\hat{r}^*]$. For this to be less than q, we assume that

$$(1-\beta)\alpha A\theta < 1.$$

Since investment is undertaken only if the return from it ($\alpha A\theta$) is greater than the return to be paid to lenders, ($\hat{r}^*$), this is sufficient to ensure $(1-\beta)\hat{r}^* < 1$. We maintain it henceforth and, without loss of generality, restrict ourselves to distributions on the domain $[0, a^U]$.

It is important to note the nonlinearity of the distributional dynamics. The current wealth distribution and threshold $\underline{a}$ determine the size of the working class (f_{1t}) which then determines equilibrium wages through (10). This endogeneity of the wage rate gives rise to nonlinear dynamics since the future wealth distribution depends upon wages via optimal bequests. The dynamic behavior of such systems can be quite complicated (see for example Aghion and Bolton, 1997). Fortunately it is sufficient for our purpose to simply track the evolution of (f_{1t}, f_{2t}, f_{3t}) instead of conducting a complete characterization of the wealth dynamics.[10] A couple of features of our model simplifies the task. In the first place, there is no feedback from the wealth distribution to $\underline{a}$ and $\overline{a}$, which are independent of

[10] We are interested in two features of a financial system, its depth and structure. Financial depth is captured by $(1-f_{1t})$, the proportion of unconstrained borrowers, while the financial structure is characterized by the relative measure of capitalists relying on bank-finance (f_{2t}) and market-finance (f_{3t}).

time.[11] Secondly, constant returns to capital at the aggregate level ($\rho_t = \alpha A$) guarantees that recursion dynamics for wealth levels exceeding $\underline{a}$ is independent of time.[12] Specifically, wealth dynamics for the two upper categories are not affected by the endogeneity of the wage rate which impacts only working-class dynamics. By exploiting the feature of the investment technology that failure yields zero returns, here too we are able to precisely characterize the dynamics.

In what follows we analyze the evolution of the financial system in two stages. First we look at the general development of the system (financial depth). Then we discuss factors that determine the structure of a *developed* financial system if it were to result in the long run.

5.1 Financial Development

The degree of credit rationing among potential capitalists determines the depth of a financial system. Hence the simplest measure of financial development comes from observing the movement of f_1.[13]

As a point of reference, it will be useful to keep in mind the *ideal* financial system − one where wealth is no constraint on borrowing (for instance, if monitoring were costless in which case there would also be no distinction between monitored and non-monitored lending). Since individuals are *ex ante* identical, this means workers and entrepreneurs must earn the same expected income in equilibrium. It is easy to show that this implies an allocation of people between wage employment and entrepreneurship, say $\breve{f}_1$, which is lower than $\tilde{f}_1$. Starting from an initial $f_{10} > \tilde{f}_1$, the economy instantaneously jumps to this $\breve{f}_1$. But, as we show below, in a world with credit frictions, equilibrium financial depth $(1 - f_1^*)$ is *always* lower than what we would observe in the ideal system. How far it departs from the ideal system is, however, what we are interested in.

We begin with Figure 2.[14] Suppose that a λ_t fraction of the f_{1t} working dynasties leave

[11] This is a feature our paper shares with Banerjee and Newman (1993) and Galor and Zeira (1993).

[12] See Aghion and Bolton (1997) and Piketty (1997) for models where returns to capital depend upon the wealth distribution.

[13] This measure corresponds closely to Levine's (2002) measures of overall financial development based on indicators of activity, size, and efficiency which are meant to "proxy for the degree to which national financial systems provide financial services: assessing firms and monitoring managers, easing risk management, and mobilizing resources". Since the only service the financial system provides here is to facilitate access to external finance, our measure $1 - f_1$ is the same as his measure of provision of financial services.

[14] We assume, for now, that returns from successful investment are high enough; specifically, successful mixed-finance capitalists become wealthy enough so that their offsprings are able to borrow using market

bequests exceeding $\underline{a}$. This means offsprings of these $\lambda_t f_{1t}$ workers will be able to borrow and become capitalists. Figures 2(a)-(c) differ only in the position of the lowest regime relative to $\underline{a}$, and hence, in λ_t.

In Figure 2(a), the wealth recursion line for $[0, \underline{a})$ lies entirely above $\underline{a}$ so that $\lambda_t = 1$. This happens when the wage rate is high enough, that is, when there are relatively few workers:

$$f_1 \leq \underline{f_1} \equiv \left[1 + \frac{\underline{a}}{(1-\beta)(1-\alpha)\pi_G A\theta q}\right]^{-1}.$$

We characterize dynamics on the two-dimensional unit simplex in (f_1, f_3). Since $\sum_\ell f_\ell = 1$, this is sufficient to determine the time-path of f_{2t}. Suppressing time subscripts, transition dynamics when $f_1 \leq \underline{f_1}$ is given by the pair of differential equations

$$\begin{aligned}
\dot{f}_1 &= (1-\pi_G)(f_2+f_3) - f_1 = (1-\pi_G) - (2-\pi_G)f_1, \\
\dot{f}_3 &= \pi_G f_2 - (1-\pi_G)f_3 = \pi_G(1-f_1) - f_3.
\end{aligned}$$

The first equation follows from noting that the outflow from the stock of workers is f_1 whereas the inflow comes from the fraction $(1-\pi_G)$ of capitalists who suffer losses on their investment and lose their entire wealth. The mass of workers, f_1, decreases or increases over time according to whether f_1 exceeds $(1-\pi_G)(f_2+f_3)$. The second differential equation is obtained similarly: the stock of market-finance capitalists increases as long as mixed-finance capitalists moving up, $\pi_G f_2$, are more numerous than market-finance capitalists suffering losses, $(1-\pi_G)f_3$. Now turn to Figure 3 for the phase-plane: when $f_1 \leq \underline{f_1}$, the $\dot{f}_1 = 0$ locus is given by the equation $f_1 = (1-\pi_G)/(2-\pi_G)$ while the $\dot{f}_3 = 0$ locus is given by $f_3 = \pi_G(1-f_1)$.

Figure 2(b) looks at another possibility, where the lowest regime of the transition mapping lies entirely below $\underline{a}$. None of the working dynasties leave bequests exceeding $\underline{a}$ here which means $\lambda_t = 0$. This happens when the wage rate is low enough, that is, workers are more numerous:

$$f_1 > \overline{f_1} \equiv \left[1 + \frac{\underline{a}[1-(1-\beta)r^*]}{(1-\beta)(1-\alpha)\pi_G A\theta q}\right]^{-1}.$$

The corresponding transition dynamics is given by:

$$\begin{aligned}
\dot{f}_1 &= (1-\pi_G)(f_2+f_3) = (1-\pi_G)(1-f_1), \\
\dot{f}_3 &= \pi_G f_2 - (1-\pi_G)f_3 = \pi_G(1-f_1) - f_3.
\end{aligned}$$

finance only.

In Figures 3(a) and (b), the $\dot{f}_1 = 0$ locus is given by $f_1 = 1$ and the $\dot{f}_3 = 0$ locus by $f_3 = \pi_G(1 - f_1)$, when $f_1 > \overline{f_1}$.[15]

A third possibility arises when the wealth recursion line on $[0, \underline{a})$ lies neither fully above nor fully below $\underline{a}$. This occurs for $\underline{f_1} < f_1 < \overline{f_1}$. In Figure 2(c), working dynasties distributed on $[\widetilde{a}_t, \underline{a})$ leave bequests exceeding $\underline{a}$, those on $[0, \widetilde{a}_t)$ do not. For a scenario like this, λ_t would depend upon the exact distribution on $[0, \underline{a})$ in general. But a moment's reflection shows we do not need details about the distribution on this interval; information about f_{1t} alone is sufficient to determine the dynamics.

To see this, we establish first that λ_t is a monotonically decreasing function of f_{1t}. An increase in f_{1t} lowers the wage rate by increasing the supply of labor; this raises $\widetilde{a}_t$ and, *ceteris paribus*, lowers λ_t. Obviously how an increase in f_{1t} gets distributed on $[0, \underline{a})$ matters, which is why detailed information about G_t may be necessary. But recall that investment failure yields zero income, which means all *new* workers start out with zero wealth. The pool of workers increases when the influx of capitalists whose investments have failed exceeds the outflow of workers who have accumulated wealth beyond $\underline{a}$. This means an increase in f_{1t} results in a bulging of the distribution at zero; hence such an increase further reduces λ_t.[16]

In addition, λ_t is a continuous function of f_{1t}. The continuous time demographic structure and a continuous initial distribution imply that changes in G_t and f_{1t} (and hence in $\widetilde{a}_t$) occur in a continuous fashion. Thus λ_t, defined by $1 - G_t(\widetilde{a}_t)/f_{1t}$, also moves continuously with f_{1t}.

We can therefore specify the dynamics corresponding to Figure 2(c) by the differential equations:

$$\begin{aligned} \dot{f}_1 &= (1 - \pi_G)(1 - f_1) - \lambda(f_1)f_1, \\ \dot{f}_3 &= \pi_G(1 - f_1) - f_3. \end{aligned}$$

Appendix A.2 demonstrates the existence of the $\dot{f}_1 = 0$ locus for a continuous $\lambda(f_{1t})$. Multiple such loci are possible but, generically, there will be an odd number of these.

Figures 3(a) and (b) illustrate dynamics under one and three such loci respectively. In both cases, when $f_{10} \leq \underline{f_1}$, point D represents a locally stable stationary distribution, while point L is a locally stable stationary distribution for $f_{10} > \overline{f_1}$. Point D, in fact, represents a well-developed financial system and is given by $(f_1^*, f_2^*, f_3^*) = \left(\frac{1-\pi_G}{2-\pi_G}, \frac{1-\pi_G}{2-\pi_G}, \frac{\pi_G}{2-\pi_G}\right)$. Point L

[15] $\overline{f_1} > \underline{f_1}$ since $(1 - \beta)r^* < 1$.

[16] Note the crucial role played by the investment technology. If failure resulted in low, but positive, returns, we would need more information about the distribution to determine how λ_t responds to G_t.

likewise represents a less-developed financial system. Indeed, there we have $(f_1^{**}, f_2^{**}, f_3^{**}) = (1, 0, 0)$, that is, a complete collapse of the financial structure.

In Figure 3(a), $\widehat{f_1}$ acts as a threshold. For values of f_{10} below $\widehat{f_1}$, the economy converges to the developed financial system, whereas for values above $\widehat{f_1}$, the long-run outcome is the primitive system. For three loci (f_1^a, f_1^b, f_1^c), as in Figure 3(b), the intermediate one acts as a local attractor for $f_{10} \in (f_1^a, f_1^c)$. In addition to the developed and underdeveloped financial systems, we now have a third kind, a moderately developed financial structure, at point M.

The complete collapse of the financial system at point L is an unattractive outcome, a consequence of there being no way out of the working class when $f_1 > \overline{f_1}$. We eliminate this extreme outcome by perturbing the dynamics slightly. We do so by allowing a very small probability (ξ) of moving up from working- to middle-class, a probability that corresponds to winning a lottery or some other form of windfall gain not captured by the model.

The phase diagram for one such perturbation is Figure 3(c). When $f_1 \leq \underline{f_1}$, this perturbation does not alter the wealth dynamics since $\lambda_t = 1$. When $f_1 > \overline{f_1}$, perturbed wealth dynamics is given by

$$\begin{aligned} \dot{f_1} &= (1 - \pi_G)(f_2 + f_3) - \xi f_1 = (1 - \pi_G) - (1 + \xi - \pi_G) f_1, \\ \dot{f_3} &= \pi_G (1 - f_1) - f_3. \end{aligned}$$

The perturbed locus $\dot{f_1} = 0$ (when $f_1 > \overline{f_1}$) lies to the left of the original one while the $\dot{f_3} = 0$ locus remains unchanged. The stationary distributions are now represented by D and L', both of which are locally stable. L' still represents a very under-developed financial structure, but we now have both f_2^{**} and $f_3^{**} > 0$.

Thus developed, underdeveloped and even moderately-developed financial systems may emerge depending upon the initial measure of credit-constrained individuals, f_{10}. High values of f_{10} are particularly inimical to financial development. For low to moderate values, a developed financial system results in the long-run, but even here, the degree to which it develops may depend upon initial conditions.

Recall that $f_{10} \equiv G_0(\underline{a})$. Clearly, f_{10} depends on the initial distribution (G_0) and on the factors that determine $\underline{a}$ – investment size (q) and institutional parameters (v and γ).[17] We consider these determinants one by one.

[17] An inspection of (12) shows that $\underline{a}$ depends positively upon investment size, q, and on the agency problems parameters v and γ.

Inequality and Financial Development

Wealth distribution is a key predictor of financial development. Suppose a high value of f_{10} results from a highly unequal initial distribution. Figure 3 shows how a developed financial system (D) results for values of f_{10} less than $\underline{f_1}$, that is, for a relatively equal distribution. Under a relatively unequal distribution, individual wealth levels will be more commonly below $\underline{a}$. When $f_{10} > \overline{f_1}$, these inequities hamper development leading to a 'collapse' of credit transactions (Figures 3(a) and 3(b)), or more plausibly, an underdeveloped financial system (Figure 3(c)). For moderate degrees of inequality such that $\underline{f_1} < f_{10} < \overline{f_1}$, outcomes depend upon the existence of additional stationary distributions. Persistent financial underdevelopment results in Figure 3(a) for values of f_{10} above $\widehat{f_1}$, whereas the economy converges to a developed financial regime for values below $\widehat{f_1}$. In Figures 3(b) or 3(c), a moderately developed financial system arises when $f_1^a < f_{10} < f_1^c$. Not as many individuals become capitalists in this case as they would when $f_{10} < f_1^a$, but credit-constrained borrowers are less numerous than at L', and credit markets thicker.

Our model thus predicts that concentrated ownership of assets like land and natural resources, which directly or indirectly determine access to investment opportunities, would hamper financial development and industrialization. Initial inequities persist when lower capital accumulation significantly depresses income for the working classes, the potential entrepreneurs of the future.

This negative association between inequality and financial development finds support with available evidence. Using data on income Gini from Deininger and Squire (1996) and indices of financial depth from Levine (2002), we obtained a negative correlation of -0.49 between inequality and financial development.[18] Indeed, some of the financially least developed countries in Levine (2002) (mostly from Latin America and Africa) are, at the same time, characterized by severe distributional problems.

This is all the more evident when we contrast middle-income countries in East Asia and Latin America. East Asia's better asset and income distribution has received considerable attention in the development literature;[19] our model relates how this difference may have

[18] For financial development, we use Levine's (2002) "Finance Aggregate" measure (Column 4, Table 2), constructed using indicators of financial activity, size and efficiency over the period 1980-95. For Levine's sample of 48 countries this index ranges from -2.2 to 1.88. For this sample of 48 countries, income Gini ranges from 24.9 to 62.5. We use income Gini for the year 1980 (as close as possible, permitted by availability) as a proxy for initial inequality. Details available upon request.

[19] Income Gini was 34.6 in East Asia and 53 in Latin America during the 1960s; land Ginis were 44.8 and 82 respectively (Deininger and Squire, 1998, Tables 1 and 2).

been vital for their financial development. Latin American nations like Argentina, Brazil, Chile, Peru and Venezuela are by and large financially underdeveloped while Hong Kong, Malaysia, South Korea, Taiwan, Thailand and Singapore's financial systems are comparable to those in Western Europe and North America.[20]

Investment Size

Consider next the effect of investment size (q). An increase in q raises the cutoff $\underline{a}$, given the wealth distribution. This could lead to financial underdevelopment if the economy is pushed over $\widehat{f_1}$ or f_1^c. An immediate implication is that poorer countries which are characterized by high inequality, such as Latin America and sub-Saharan Africa, ought to rely more on small- and medium-scale industries for their development. An emphasis on import-substituting heavy industries, for instance, would be counterproductive in the long-run.

Institutional Factors

Individuals do not differ in terms of their innate abilities in our model and access to credit markets is limited solely by informational asymmetries and costs. A relevant question is how better institutions mitigate these asymmetries and what that implies for financial structure. A simple way to interpret institutions here is through the parameters γ, v and V. These parameters affect the depth and structure of a financial system through the nature and magnitude of agency problems and costs of controlling it.

As noted earlier, the degree of credit-rationing, f_{10}, depends positively upon the institutional parameters γ and v through $\underline{a}$. When legal and financial institutions are too inefficient ($f_{10} > \widehat{f_1}$ or $f_{10} > f_1^c$), the financial system remains underdeveloped in the long run; efficient institutions lead to financial development in the long-run. This prediction is along the lines of recent studies of the 'legal-based' view of financial development in LaPorta et al. (1997, 1998), where the quality and nature of legal rules and law enforcements protecting shareholders and creditors are seen as fundamental to financial activity. Systematic empirical support for this view, covering a wide range of countries, is offered by Levine (2002).

5.2 Financial Structure

Financial structure refers to the combination of financial instruments, markets and institutions operating in an economy (Goldsmith, 1969). We consider the relative importance of

[20] See Table 2 in Levine (2002) and Table 3.12 in Demirgüc-Kunt and Levine (2001b).

market- and bank-finance, $\psi_t \equiv f_{3t}/f_{2t}$, as an index of structure in our model and study its evolution.

In the analysis so far we have implicitly allowed high investment returns to ensure that successful mixed-finance capitalists move up to the next wealth category. What if that were not the case? We illustrate such a scenario in Figure 5 which depicts wealth recursion dynamics for $f_1 \leq \underline{f_1}$ (the other two cases would parallel Figures 2(b) and (c)). It turns out that the nature of the transition dynamics does not change here although the composition of the stationary distribution does.

Transition dynamics corresponding to Figure 5 is given by:

when $f_1 \leq \underline{f_1}$,

$$\begin{aligned} \dot{f_1} &= (1-\pi_G) - (2-\pi_G)f_1, \\ \dot{f_3} &= -(1-\pi_G)f_3; \end{aligned}$$

when $f_1 > \overline{f_1}$,

$$\begin{aligned} \dot{f_1} &= (1-\pi_G)\,(1-f_1)\,, \\ \dot{f_3} &= -(1-\pi_G)f_3; \end{aligned}$$

and when $\underline{f_1} < f_1 < \overline{f_1}$,

$$\begin{aligned} \dot{f_1} &= (1-\pi_G)(1-f_1) - \lambda(f_1)f_1, \\ \dot{f_3} &= -(1-\pi_G)f_3. \end{aligned}$$

In Figure 6(a), when $f_{10} \leq \underline{f_1}$, D represents a locally stable stationary distribution with a 'developed' financial system, where $(f_1^*, f_2^*, f_3^*) = \left(\frac{1-\pi_G}{2-\pi_G}, \frac{1}{2-\pi_G}, 0\right)$. As before, when $f_{10} > \overline{f_1}$, point L represents a locally stable stationary distribution representing a 'less-developed' financial system with $(f_1^{**}, f_2^{**}, f_3^{**}) = (1, 0, 0)$, and when $f_1^a < f_{10} < f_1^c$, point M represents a moderately developed financial system. Not surprisingly, the long-run distribution has no capitalist relying purely on market finance since middle-class capitalists cannot move up.

As before, such an outcome can be avoided with perturbations that allow workers to move up to the middle-class with a small probability (ξ), and middle-class capitalists to move up with a similar probability (ε). Under perturbation, the stationary distributions in Figure 6(b) are represented by D' when $f_{10} \leq \underline{f_1}$, L' when $f_{10} > \overline{f_1}$ and M' when $f_1^a < f_{10} < f_1^c$. All three points are locally stable.

Compare now the financial structure of D (or D') in Figures 3 and 6 under different rates of return. When investment returns are low, all (or a very large proportion of) eligible

capitalists go through bank intermediation in the long-run. That is, bank-finance is relatively more important when investment returns are lower (this would be true even for a moderately developed financial system). The long-run financial structure is, hence, more market-based for a configuration like Figure 2 and more bank-oriented for a situation like Figure 5.

A market-based system (Figure 2) occurs when the height of C exceeds $\overline{a}$, that is when,

$$(1-\beta)\, r^* \left[\frac{v}{\pi_G - \pi_B}\right] \geq \frac{\pi_G V}{\pi_G - \pi_B} - (\alpha \pi_G A\theta - r^*)\,. \tag{14}$$

This is more likely to occur when V is relatively low and v relatively high, or when θ is relatively high but π_G low (holding $\pi_G\theta$ constant).[21] On the other hand a bank-based financial system (Figure 5) results when the height of H is less than $\overline{a}$. This happens if

$$(1-\beta)\left[\alpha \pi_G A\theta - (1+\gamma)r^*\right] < \left[1-(1-\beta)\frac{r^*}{\pi_G}\right]\left(\frac{\pi_G}{r^*}\right)\left[\frac{\pi_G}{\pi_G - \pi_B}V - \{\alpha A \pi_G \theta - r^*\}\right]. \tag{15}$$

This inequality is more likely when V and γ are relatively high, or, holding $\pi_G\theta$ constant, when θ is relatively low and π_G high.[22]

Thus, the financial structure of an economy is determined by the investment technology (π_G, θ) and institutional factors (γ, v and V). We consider these issues next.

Investment Risk

Although banks and individuals are risk-neutral in our model, investment risk affects financial structure through its effect on wealth dynamics. We begin by contrasting two types of investment that yield the same expected return, $\pi_G\theta$: type-I projects yield a high θ but are riskier since π_G is low, while type-II projects succeed more often but realize low θ.[23]

Suppose now that the two project types differ significantly in their riskiness so that Figure 2 depicts wealth dynamics for type-I projects while that for type-II projects is given by Figure 5. From (14) and (15) we know that Figure 2 is more likely to occur when θ is relatively high but π_G low (holding $\pi_G\theta$ constant) while Figure 5 is more likely to occur for the opposite case.

Figures 2 and 5 lead to dynamics shown by Figures 3 and 6 respectively. We draw two conclusions on the role of investment risk. First, lower π_G leads to higher f_1^* so that credit rationing is more widespread in the long-run. Secondly, when investment is less risky, all or a large proportion of eligible capitalists go through bank intermediation in the long-run. In

[21] We allow for a proportionate decrease in π_B as well.

[22] Again, we change π_B by the same proportion as π_G.

[23] We allow for a proportionate change in π_B between the two project types.

other words, bank-finance is relatively more important for safer technologies, whereas market finance gains importance for riskier ones.

This dependence of financial structure on risk is quite distinct from, but complementary to, the ones commonly analyzed in the finance literature. Specifically, since agents are risk-neutral our analysis misses the typical portfolio effect discussed in the literature.[24] At the same time it brings to the analysis the macroeconomic feedback that investment risk has on asset positions and financing dynamics, an effect entirely absent from the existing literature.

Nature of Institutions

Institutional parameters also affect long run financial structure in the model. Recall that a market-based system is more likely to occur (that is, (14) holds) when V is relatively low and v relatively high. On the other hand, a bank-based system (condition (15)) is more likely to occur when V and γ are relatively high. It is quite intuitive that a bank-based system is more likely when the residual moral hazard under bank monitoring (v) is low relative to what incentives would be in the absence of monitoring (V). This conclusion is consistent with Demirgüc-Kunt and Levine's (2001b) finding that countries with strong shareholder rights relative to creditor rights and strong accounting systems (that is, low V relative to v) tend to have more market-based systems.

It may seem surprising that higher monitoring costs, γ, lead to a more bank-oriented systems even though these costs are borne by the banking sector. This is easy to understand once we recognize that wealth and financing dynamics depend upon investment earnings. A higher cost of monitoring means that banks need to inject a larger amount of their own resources into the investment project. This forces mixed-finance capitalists to rely more heavily on expensive intermediated finance; consequently less of them are able to move up to become market-finance capitalists.

6 Further Discussion

6.1 Historical Implications

Several of the model's implications shed light on the development of financial systems during the Industrial Revolution.

[24] Risk averse agents would clearly make the analysis much less tractable. It is to be noted, though, that the literature is not unequivocal about whether banks or markets diversify risks better, suggesting only that both are important. Levine (1997) and Allen and Gale (2001) discuss these issues.

Inequality and Financial Structure

Consider first the effect of *initial* inequality on financing choices. For convenience, assume that the initial distribution is lognormal with mean μ_0 and variance σ_0^2, where $\underline{a} < \mu_0 < \overline{a}$. In Appendix A.3 we establish that an increase in σ_0^2 tends to raise f_{10}, lower f_{20} but increase f_{30}. Higher inequality leads to thinner capital markets since $1 - f_{10}$ is lower. But among those who obtain loans, there is a shift toward market finance and away from bank finance, increasing the ratio ψ_0. Historical reliance upon the two types of finance may, in other words, depend upon inequality.

This prediction seems to be corroborated by what we know about England and Germany during the industrial revolution. The Anglo-Saxon financial system, with its creditors pursuing more of a "hands-off" lending, was more market-oriented. Banks were mostly concerned with liquidity and did not engage in long-term lending so that British industries primarily depended upon internal finance and the London Stock Exchange for their financing needs (Collins 1995, Allen and Gale, 2000). German industries, in comparison, relied more on bank finance. German bankers kept a continuous watch over the development of companies and were often represented on the company boards (Allen and Gale, 2000, Baliga and Polak, 2004).

At the same time, substantial evidence suggests England had a more unequal land distribution than Germany (and France) (Clapham, 1936, Soltow, 1968). Landes (1969) also notes that a large number of British industrialists were "men of substance", having accumulated significant wealth from merchant activities. This distributional difference between the two regions could partly explain why the Anglo-Saxon and German financial systems have historically differed. Indeed, this also explains why other societies with better distributions than England, for instance France and Japan (or the newly industrializing East Asian countries), have traditionally relied more on bank-finance (see Allen and Gale, 2000). A systematic analysis covering a broad sample of countries is clearly required to establish this relationship.

There is an important point to be noted here. The initial distribution has no permanent impact on financial structure as long as countries similar in other respects are converging to the same (developed) financial system. Interestingly, Allen and Gale (2000) point out how there has been a convergence of financial systems in developed countries as traditionally bank-oriented societies such as France, Germany or Japan have moved closer toward market finance since the 1980s. One could interpret this as a convergence in industrialized country financial systems (policies, which are exogenous here, would have clearly played a role too).

Investment Size and Financial Structure

Investment size (q) has an impact on *initial* (historical) financial structure. With a higher q, fewer individuals are able to obtain loans either from markets or from banks. At the same time, the shortfall $q - a^i$ that has to be raised through external finance is higher for those who do invest. The Holmstrom-Tirole incentive structure has a straightforward implication: due to limited liability, a borrower's incentive to be diligent is weaker the less her personal stake in the project is, that is, the more she needs to borrow. The only way to attenuate this is through increased monitoring.

Higher investment requirements would hence push an economy towards bank finance. But whether or not this happens depends also on the wealth distribution. A higher q raises the importance of bank-finance under two conditions: when the initial wealth distribution among capitalists is more equitable, that is when f_{30}/f_{20} is low; and when banks are particularly effective at resolving incentive problems ($\delta_1 << \delta_2$ in (12)) so that the measure of individuals above $\underline{a}$ is sizeable.[25]

Historical evidence, once again, provides some support of this story. Similar to their financial systems, a distinction is often made between England and Germany's industrialization patterns. The British process of industrialization mainly relied upon small- and medium-scale industries, textile manufacturing being a prime example. Germany, on the other hand, largely utilized heavy manufacturing and chemical industries for its development, both requiring far greater investment than in the case of England (see Landes, 1969, and related references in Baliga and Polak, 2004). Consistent with the evidence, our model suggests that this technological difference would be reflected in a greater German reliance on intermediated finance especially if banks are efficient intermediaries and the wealth distribution is more equitable.

Using a static model of monitored and non-monitored debt, Baliga and Polak (2004) highlight the role of investment size for Western Europe's financial structure.[26] Our dynamic model suggests, though, that differences in financial structure are *neutral in the long run* with respect to investment size. As long as $f_{10} < \widehat{f}_1$ (or less than f_1^a), differences in f_{10} do not translate into differences in the stationary distribution, that is, point D. This is because q affects threshold wealth requirements ($\underline{a}$ and $\overline{a}$) as well as expected income from investment. In the short-run, a higher q could increase reliance upon bank-finance but it also enables successful capitalists to earn more. When more of the middle-class capitalists move above

[25] This follows from noting that $\partial\psi_0/\partial q < 0$ whenever $\delta_2(1 + f_{20}/f_{30})g_0(\overline{a}) > \delta_1 g_0(\underline{a})$.

[26] In the Baliga-Polak model a firm borrows using either monitored or non-monitored debt but not both.

$\overline{a}$, they do not need to be monitored so that reliance upon bank finance declines.

The general equilibrium dynamics also suggests a second important way investment size affects outcomes, by giving rise to threshold effects. For a given initial wealth distribution, the greater the capital needs of industrialization, the more likely it is that $f_{10} > \hat{f}_1$ in which case the financial system will continue to remain underdeveloped. The question of whether banks or markets are more important for firms' financing needs becomes less relevant in such a scenario.

6.2 Life-Cycle of Firms

A conventional wisdom in corporate finance visualizes a firm's life-cycle as follows: the firm relies mostly on internal assets and venture capitalists (the so-called 'angels') for investment in its early stages; as it matures, financial intermediaries start lending to the firm; finally, when the firm is mature enough, it raises funds from the market.

This is the pattern implied, for example, by Myers' (1984) pecking order theory of finance.[27] Diamond (1991) captures this feature in a dynamic model of firm financing choices: firms use expensive (and monitored) bank finance in the early stages of their life-cycle, and, as they develop better reputations, switch to cheaper forms of financing such as publicly held debt. A similar life-cycle pattern is discussed by Holmstrom (1996, p. 229) for a dynamic extension of his static model: "...firm financing will have a life cycle in which over time and assuming success, firms shift from using more information intensive to less information intensive capital".

Our analysis essentially carries out what Holmstrom suggests above. Doing so makes it clear why it is important to think about firm-financing choices in a dynamic general equilibrium framework. In particular our analysis cautions against generalizing from static partial equilibrium models if our goal is to understand the process of financial development. As we show in section 5 above, the life-cycle process anticipated by the finance literature depends sensitively on macro-fundamentals like wealth inequality, investment size and institutional determinants. These factors affect financial development and the life-cycle pattern of a typical firm works well only when the financial system fully develops over time. If initial conditions, policy choices and parameters of the economy are not appropriate for long-run financial development, firms may find it difficult to switch to less information-intensive sources of financing even with time and high levels of internal assets.

[27] For instance, Fluck (2000, p. 7) says "Myers predicts that firms will issue debt first and outside equity only later".

This also means a blind push towards one type of external finance in developing countries, as we have seen in the last two decades, is flawed without taking into account the deeper problems affecting these financial systems. Policies regarding patterns of industrialization, banking, transparency and availability of information on borrowing firms, can affect not just current firm financing choices but the ability to access credit markets over time and to invest in industries that require a heavy dose of investment or those that are particularly risky.

6.3 Policy Considerations

We draw several policy conclusions from the model. First, as is common to most models of market imperfections and non-convexities, temporary policies can have permanent effects in this economy. Consider for example a (temporary) policy of emphasizing small-scale enterprise in the early stages of development. Since investment requirements of these industries are typically low, not only can a larger fraction of entrepreneurs access credit markets, they can do so by direct borrowing instead of more expensive monitored finance. This facilitates financial depth and permits greater wealth accumulation, that is, faster convergence to a developed financial system. Over time, as typical wealth levels rise and industrialization needs change, accessing credit markets will be less of a problem even with rising investment needs.

Such permanent effect of temporary policies is especially important if we think financial systems remain underdeveloped because institutions are ineffective. Inefficient institutions are widespread and informational problems more pronounced in poorer countries presumably because better institutions are costly to implement (similar to the costs of operating markets in Greenwood and Smith, 1997). In our model, one way to get around such institutional bottlenecks is through a temporary income redistribution that relaxes credit constraints for a sizeable number of potential entrepreneurs. The benefits of such a redistribution will be persistent if it pushes f_{10} below the relevant threshold, and in fact, could be politically more palatable than permanent distributive policies such as land reforms.

Recall our brief discussion of an ideal financial system in section 5.1 above. The ideal system corresponds to the frictionless counterpart of the economy we have analyzed so far and 'maximizes' access to credit, hence, generation of income. One can then view the goal of policy as taking the frictional economy as close as possible to the ideal system. Clearly this means banks and markets should have better access to firm-specific information and that bank intermediation be made more effective. Long-run financial sector reforms, in other words, should focus on lowering v, V and γ. At the same time, since the investment

technology has a bearing on the convergence path and steady-state outcomes, policies should focus on encouraging technologies that are more productive (higher θ) and less capital-intensive (at least initially).

Another goal of policy can be to push one type of financing over another – as we have seen recently – in order to increase aggregate income. In this economy both bank-based and market-based systems generate the same GDP, but GNP differs. Specifically, Appendix A.4 shows that market-based systems generate a higher GNP. Policies to raise GNP would then involve promoting market-finance. Better disclosure and bankruptcy laws will of course help by lowering V. But banking sector reforms can help too. While lowering γ is desirable both as a temporary and permanent policy, lowering v helps only as a temporary policy. Lowering v via legal reforms like better protection of creditor rights and improving the bank's role as an enforcer and protector of these creditor rights enhances the reach of the banking sector (lowers $\underline{a}$). This is good in the short run as less individuals are credit-constrained, but not so good in the long run since bank monitoring costs constitute a resource drain. Thus, unless this policy is coupled with policies to reduce γ or V (both of which help capitalists accumulate wealth faster), the result may be an undesirable (again for generating higher GNP) bank-based system.[28]

7 Conclusion

This paper has analyzed the evolution of a financial system and identified factors determining its development and structure. We introduced monitored bank loans and non-monitored tradeable securities in a dynamic general equilibrium model and showed how the path to financial development exhibits non-ergodic behavior – underdeveloped financial systems persist in highly unequal societies and in economies with capital-intensive industries or inefficient legal and financial institutions. The model's key predictions are consistent with the historical development of financial systems during the Industrial Revolution. We also show that the

[28] The relative attractiveness of a market-based system needs to be qualified. Here we have abstracted from growth and assumed that credit-constrained borrowers become workers instead of engaging in low-productive 'traditional' entrepreneurial activities which may not benefit as much from industrialization. We address these issues in a companion piece, Chakraborty and Ray (2006), by abstracting from the distributional complications of this paper. There we show that a bank-based system has better distributional outcomes than market-based ones (though neither is necessarily better for growth). So how one views the desirability of one financial system over another depends on the relative emphasis placed on aggregate income versus wealth/income distribution.

typical life-cycle financing decisions of a firm, as envisioned by papers in corporate finance, can be hamstrung by weak macroeconomic fundamentals and policy parameters. Finally we shed light on the banks-versus-market policy debate that has influenced much of financial sector reforms around the world.

Compared to existing works on the dynamic interaction between credit markets and the wealth distribution, our goal has been to obtain a clearer understanding of what drives the development and structure of financial systems. Hence an important contribution of this paper lies in extending the literature by incorporating elements that allow policy analysis for developing countries.

We conclude by considering some extensions for future work. Throughout the paper we focused on the demand side of financial systems. We did this primarily for tractability. A natural extension would be to consider how important the supply of loanable funds is to financial structure, for example, by looking at a closed economy version of the model. A more challenging extension would be to capture the emergence of different financial institutions (institutions like banks and markets are taken as given in our story). For instance, if there are fixed costs to setting up an intermediary (as in Greenwood and Jovanovic, 1990), the extent to which banks emerge and monitor lending will depend upon, and affect, the pattern of wealth accumulation. It will be also interesting to endogenize the interest rate, allowing the wealth distribution to affect financial structure through returns to bank and market-finance.

Another institutional aspect we have ignored, and one likely to be important in development, is the quality of bank monitoring. In particular, banks do not face any incentive problems vis-a-vis depositors in our model. Extensions incorporating agency problems within the banking sector may be used to examine how the "quality" of bank-finance itself changes over time and with respect to macro-fundamentals.

References

[1] Aghion, Philippe and Patrick Bolton (1997), "A Theory of Trickle-Down Growth and Development", *Review of Economic Studies*, vol. 64, pp. 151-172.

[2] Allen, Franklin and Douglas Gale (2000), *Comparing Financial Systems*, Cambridge, MA, MIT Press.

[3] Allen, Franklin and Douglas Gale (2001), "Comparative Financial Systems: A Survey", forthcoming in Bhattacharya, Sudipto, Boot, Arnoud and Anjan Thakor (eds.) *Financial Intermediation*, Oxford University Press.

[4] Baliga, Sandeep and Ben Polak (2004), "The Emergence and Persistence of the Anglo-Saxon and German Financial Systems," *Review of Financial Studies*. Vol. 17, No. 1, pp 129-163.

[5] Banerjee, Abhijit and Andrew Newman (1993), "Occupational Choice and the Process of Development", *Journal of Political Economy*, vol. 101, pp. 274-298.

[6] Chakraborty, Shankha and Tridip Ray (2006), "Bank-based versus Market-based Financial Systems: A Growth-theoretic Analysis", *Journal of Monetary Economics*, vol. 53, pp. 329-350.

[7] Clapham, John Harold (1936), *The Economic Development of France and Germany, 1815-1914*, Cambridge University Press.

[8] Collins, Michael (1995), *Banks and industrial finance in Britain, 1800-1939*, Cambridge University Press.

[9] Deininger, Klaus and Lyn Squire (1996), "Measuring Income Inequality: A New Data-Base," *World Bank Economic Review*, vol. 10, no. 3 (September), pp. 565-91.

[10] Deininger, Klaus, and Lyn Squire (1998), "New Ways of Looking at Old Issues: Inequality and Growth," *Journal of Development Economics*, vol. 57, no. 2, pp. 259-287.

[11] Demirgüc-Kunt, Asli and Ross Levine eds. (2001a) *Financial Structure and Economic Growth*, MIT Press, Cambridge, MA.

[12] Demirgüc-Kunt, Asli and Ross Levine (2001b), "Bank-based and Market-based Financial Systems: Cross-country Comparisons", in Asli Demirgüc-Kunt and Ross Levine (eds.) *Financial Structure and Economic Growth*, MIT Press, Cambridge, MA.

[13] Demirgüç-Kunt, Asli and V. Maksimovic (1998), "Law, Finance, and Firm Growth", *Journal of Finance*, 53: 2107-2137.

[14] Diamond, Douglas (1984), "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, vol. 51, pp. 393-414.

[15] Diamond, Douglas (1991), "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt", *Journal of Political Economy*, vol. 99, pp. 689-721.

[16] Fluck, Zsuzsanna (2000), "Capital Structure Decisions in Small and Large Firms: A Life-cycle Theory of Financing", *mimeo*, New York University.

[17] Galor, Oded and Joseph Zeira (1993), "Income Distribution and Macroeconomics," *Review of Economic Studies*, vol. 60, pp. 35-52.

[18] Goldsmith, Raymond (1969), *Financial Structure and Development*, Yale University Press, New Haven, CT.

[19] Greenwood, Jeremy and Boyan Jovanovic (1990), "Financial Development, Growth and the Distribution of Income", *Journal of Political Economy*, vol. 98, no. 5, pp. 1076-107.

[20] Greenwood, Jeremy and Bruce Smith (1997), "Financial Markets in Development, and the Development of Financial Markets", *Journal of Economic Dynamics and Control*, vol. 21, pp. 145-181.

[21] Hellwig, Martin (1991), "Banking, Financial Intermediation and Corporate Finance", in *European Financial Integration*, (eds.) Alberto Giovannini and Colin Mayer, Cambridge University Press, U.K.

[22] Holmstrom, Bengt (1996), "Financing of Investment in Eastern Europe," *Industrial and Corporate Change*, vol. 5, pp. 205-37.

[23] Holmstrom, Bengt and Jean Tirole (1997), "Financial Intermediation, Loanable Funds, and the Real Sector," *Quarterly Journal of Economics*, vol. 112, pp. 663-91.

[24] La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny (1997), "Legal Determinants of External Finance," *Journal of Finance*, vol. 52, pp. 1131-50.

[25] La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert Vishny (1998), "Law and Finance," *Journal of Political Economy*, vol. 106, pp. 1113-55.

[26] Landes, D. (1969), *The Unbound Prometheus, Technological Change and Industrial Development in Western Europe from 1750 to the Present*, Cambridge, Cambridge University Press.

[27] Levine, Ross (1997), "Financial Development and Economic Growth: Views and Agenda", *Journal of Economic Literature*, vol. 35, pp. 688-726.

[28] Levine, Ross (2002), "Bank-based or Market-Based Financial Systems: Which is Better?", *Journal of Financial Intermediation*, vol. 11, pp. 1-30.

[29] Levine, Ross (2005), "Finance and Growth: Theory and Evidence", in Philippe Aghion and Steven Durlauf, eds. *Handbook of Economic Growth*, The Netherlands: Elsevier Science.

[30] Mookherjee, Dilip and Debraj Ray (2002), "Contractual Structure and Wealth Accumulation", *American Economic Review*, vol. 92, no. 4, pp. 818-849.

[31] Myers, Stewart C. (1984), "The Capital Structure Puzzle", *Journal of Finance*, vol. 39, no. 3, pp. 575-592.

[32] Piketty, Thomas (1997), "The Dynamics of the Wealth Distribution and the Interest Rate with Credit Rationing", *Review of Economic Studies*, vol. 64, pp. 173-189.

[33] Soltow, Lee (1968), "Long-Run Changes in British Income Inequality", *Economic History Review*, Second Series, 21, 17-29.

[34] Turnovsky, Stephen J. (2000), *Methods of Macroeconomic Dynamics*, 2nd edition, MIT Press.

Appendix

A.1. Optimal Contracts

Direct Finance

We have $x_t^C + x_t^U = \rho_t \theta q$. Capitalist-$i$'s incentive compatibility constraint (choosing the good project) is given by

$$\pi_G x_t^C \geq \pi_B x_t^C + Vq.$$

Direct lender's participation constraint is

$$\pi_G x_t^U \geq r^* \left(q - a_t^i\right).$$

Capitalist's incentive compatibility constraint implies

$$x_t^C \geq \frac{Vq}{\pi_G - \pi_B} \Rightarrow x_t^U = \rho_t \theta q - x_t^C \leq \rho_t \theta q - \frac{Vq}{\pi_G - \pi_B}.$$

Using this, the lender's participation constraint gives a threshold wealth level for access to direct finance

$$r^* \left(q - a_t^i\right) \leq \pi_G x_t^U \leq \pi_G \left[\rho_t \theta q - \frac{Vq}{\pi_G - \pi_B}\right] \Rightarrow a_t^i \geq \overline{a}_t \equiv \frac{q}{r^*} \left[\frac{\pi_G}{\pi_G - \pi_B} V - \left\{\pi_G \rho_t \theta - r^*\right\}\right].$$

Indirect Finance

Under indirect finance we have $x_t^C + x_t^U + x_t^B = \rho_t \theta q$. Here the optimal contracts need to satisfy the following three constraints:

(i) capitalist-i's incentive compatibility constraint (choosing the good project):

$$\pi_G x_t^C \geq \pi_B x_t^C + vq,$$

(ii) bank's incentive constraint (for monitoring):

$$\pi_G x_t^B - r^* \gamma q \geq \pi_B x_t^B,$$

assuming that banks discount monitoring costs at their opportunity cost, r^*,

(iii) participation constraint of the uninformed investors:

$$\pi_G x_t^U \geq r^* \left(q - l_t^i - a_t^i\right),$$

where l_t^i is the amount that the bank lends to capitalist-i.

The bank's return from lending l_t^i to capitalist-i is

$$x_t^B = r_t^L l_t^i, \tag{16}$$

where r_t^L is the (gross) loan rate charged to borrowers when projects succeed.

But the bank's incentive constraint implies

$$\pi_G x_t^B \geq \left[\frac{\pi_G}{\pi_G - \pi_B}\right] r^* \gamma q. \tag{17}$$

Since bank finance is relatively more expensive than direct finance (due to monitoring costs), borrowers accept only the minimum amount necessary so that

$$r_t^L l_t^i = x_t^B = \left[\frac{1}{\pi_G - \pi_B}\right] r^* \gamma q$$

$$\Rightarrow l_t^i(r_t^L) = \frac{r^* \gamma q}{(\pi_G - \pi_B)\, r_t^L}. \tag{18}$$

Capitalist's incentive compatibility constraint implies $x_t^C \geq vq/(\pi_G - \pi_B)$. Then

$$x_t^C + x_t^B \geq \frac{v + r^*\gamma}{\pi_G - \pi_B} q \Rightarrow x_t^U = \rho_t \theta q - \left(x_t^C + x_t^B\right) \leq \rho_t \theta q - \left(\frac{v + r^*\gamma}{\pi_G - \pi_B}\right) q.$$

Using this, the uninformed investors' participation constraint gives

$$r^* \left(q - l_t^i - a_t^i\right) \leq \pi_G x_t^U \leq \pi_G \left[\rho_t \theta - \frac{v + r^*\gamma}{\pi_G - \pi_B}\right] q.$$

It follows that only capitalists with wealth

$$a_t^i \geq \underline{a}_t \equiv q - l_t^i(r_t^L) - \frac{\pi_G}{r^*}\left[\rho_t \theta - \frac{v + \gamma r^*}{\pi_G - \pi_B}\right] q \tag{19}$$

are able to convince uninformed investors to supply enough funds for the investment project.

The Bank's Problem

The competitive banking sector is profit maximizing. The aggregate demand for bank loans is

$$L_t = \int_{i \in I_t} l_t^i(r_t^L) dG_t = \left[\frac{\gamma r^* q}{(\pi_G - \pi_B)\, r_t^L}\right] \int_{i \in I_t} dG_t,$$

where I_t denotes the subset of individuals using intermediated finance. The total monitoring cost borne by the banking sector is then

$$\gamma q \int_{i \in I_t} dG_t = \frac{(\pi_G - \pi_B)\, r_t^L L_t}{r^*}.$$

Let D_t denotes the flow of deposits into the banking sector. Expected banking profits are

$$\Pi_t^B = \pi_G r_t^L L_t - r^* D_t. \tag{20}$$

Banks face the resource constraint that total loans cannot exceed total deposits net of monitoring costs:

$$L_t \leq D_t - \gamma q \int_{i \in I_t} dG_t. \tag{21}$$

The banking sector's optimization problem in period t is to choose L_t so as to maximize Π_t^B subject to the capitalist's incentive constraint and the constraints (17) and (21).

Since bank profits are increasing in total loans, (21) holds with equality:

$$L_t = D_t - \gamma q \int_{i \in I_t} dG_t = D_t - \frac{(\pi_G - \pi_B)\, r_t^L}{r^*} L_t. \tag{22}$$

Moreover, in a competitive equilibrium, the banking sector earns zero expected profits. From (20) we then have

$$\pi_G r_t^L L_t = r^* D_t, \tag{23}$$

It follows from equations (22) and (23) that

$$L_t = \left(\frac{\pi_B}{\pi_G}\right) D_t,$$

and

$$r_t^L = \frac{r^*}{\pi_B}. \tag{24}$$

Hence, using (18), we observe that

$$l_t^i = \gamma \left(\frac{\pi_B}{\pi_G - \pi_B}\right) q. \tag{25}$$

In other words, for all $i \in I_t$, banks finance a fixed proportion of the borrower's investment, irrespective of a_t^i.

Taking into account the optimal loan size (25), the lower wealth cut-off (19) becomes

$$\underline{a}_t = \frac{q}{r^*} \left[\frac{\pi_G v}{\pi_G - \pi_B} - \{\pi_G \rho_t \theta - (1+\gamma) r^*\}\right]. \tag{26}$$

A.2. Existence of $\dot{f}_1 = 0$ locus when $\underline{f_1} < f_1 < \overline{f_1}$

In the text we have already established that λ_t is a continuous and monotonically decreasing function of f_{1t}. Now define

$$F(f_1) \equiv (1 - \pi_G)/[1 - \pi_G + \lambda(f_1)] - f_1$$

on the interval $[\underline{f_1}, \overline{f_1}]$. Since $\lambda(f_1)$ is continuous on $[\underline{f_1}, \overline{f_1}]$, F is also continuous on $[\underline{f_1}, \overline{f_1}]$. We have $F(\underline{f_1}) = (1-\pi_G)/(2-\pi_G) - \underline{f_1} < 0$ since $\underline{f_1} > (1-\pi_G)/(2-\pi_G)$ and $F(\overline{f_1}) = 1 - \overline{f_1} > 0$, since $\overline{f_1} < 1$. Hence, using the Intermediate Value Theorem, since F is continuous on $[\underline{f_1}, \overline{f_1}]$ and since $0 \in [F(\underline{f_1}), F(\overline{f_1})]$, we can find an $u \in [\underline{f_1}, \overline{f_1}]$ such that $F(u) = 0$. In other words, $F(f_1) = 0$ for at least one value of $f_1 \in [\underline{f_1}, \overline{f_1}]$. Figure 4(a) illustrates this when the line $(1-\pi_G)/[1-\pi_G+\lambda(f_1)]$ intersects f_1 once, at $\widehat{f_1}$. Multiple such intersections are also possible, but, generically, these have to be in odd numbers. Figure 4(b) depicts three intersections. Figures 3(a) and (b) illustrate dynamics under one and three such intersections respectively.

A.3. Effect of an increase in initial inequality on financing choices

Suppose that the initial distribution G_0 is lognormal with mean μ_0 and variance σ_0^2. Recall that the lognormal cumulative distribution is given by $\Phi\left[(\ln x - \mu)/\sigma\right]$, where Φ is the standard Normal cumulative distribution function. Then:

$$\frac{\partial G_0(\underline{a})}{\partial \sigma_0} = -\phi_0\left(\frac{\ln \underline{a} - \mu_0}{\sigma_0}\right)\left[\frac{\ln \underline{a} - \mu_0}{\sigma_0^2}\right], \ \frac{\partial G_0(\overline{a})}{\partial \sigma_0} = -\phi_0\left(\frac{\ln \overline{a} - \mu_0}{\sigma_0}\right)\left[\frac{\ln \overline{a} - \mu_0}{\sigma_0^2}\right].$$

When $\ln \underline{a} < \mu_0 < \ln \overline{a}$, the first derivative is positive, the second derivative negative. Moreover,

$$\frac{\partial[G_0(\overline{a}) - G_0(\underline{a})]}{\partial \sigma_0} = -\phi_0\left(\frac{\ln \overline{a} - \mu_0}{\sigma_0}\right)\left[\frac{\ln \overline{a} - \mu_0}{\sigma_0^2}\right] + \phi_0\left(\frac{\ln \underline{a} - \mu_0}{\sigma_0}\right)\left[\frac{\ln \underline{a} - \mu_0}{\sigma_0^2}\right],$$

is also negative in this case. Thus, as long as $\ln \underline{a} < \mu_0 < \ln \overline{a}$, we have

$$\frac{\partial f_{10}}{\partial \sigma_0} > 0, \frac{\partial f_{20}}{\partial \sigma_0} < 0, \ \frac{\partial f_{30}}{\partial \sigma_0} > 0.$$

A.4 GNP Calculations

GNP = GDP + Net interest income from abroad (NIA). Since GDP is the same in both systems (the number of workers and capitalists are the same under both systems and these are the only two factors of production), let us concentrate on NIA. Assume that workers supply their savings to the domestic financial sector first, and then invest any excess on the international capital market. Similarly, the domestic financial sector first relies on the domestic lonable funds market before approaching the international capital market. Since banks and entrepreneurs pay the world rate of return, r^* (in an expected sense), the loan market always clears.

Demand for loanable funds = demand from banks seeking deposits (D_t) + demand from entrepreneurs seeking direct finance (M_t).

$$\begin{aligned} D_t &= \left(\frac{\pi_G}{\pi_B}\right) L_t = \left(\frac{\pi_G}{\pi_B}\right) \int_{\underline{a}}^{\overline{a}} l_t^i dG_t \\ M_t &= \int_{\underline{a}}^{\overline{a}} \left(q - l_t^i - a_t^i\right) dG_t + \int_{\overline{a}}^{a^U} \left(q - a_t^i\right) dG_t \end{aligned}$$

Hence demand for loanable funds is

$$D_t + M_t = (1+\gamma)\, q \int_{\underline{a}}^{\overline{a}} dG_t + q \int_{\overline{a}}^{a^U} dG_t - \int_{\underline{a}}^{a^U} a_t^i dG_t.$$

Supply of loanable (S_t) funds comes from the workers and is: $S_t = \int_0^{\underline{a}} a_t^i dG_t$. Net lending abroad is

$$NLA_t = S_t - D_t - M_t = \int_0^{a^U} a_t^i dG_t - q \int_{\underline{a}}^{a^U} dG_t - \gamma q \int_{\underline{a}}^{\overline{a}} dG_t.$$

The last term in this expression makes it clear that the amount spent on monitoring bank-dependent capitalists is a drain from investible resources.

Now consider the two financial systems. Note that we have distinguished between the two systems only at the steady state and that too for a developed financial system. Under a market-based system

$$\int_0^{\underline{a}} dG_t = \frac{1-\pi_G}{2-\pi_G}, \quad \int_{\underline{a}}^{\overline{a}} dG_t = \frac{1-\pi_G}{2-\pi_G}, \text{ and } \int_{\overline{a}}^{a^U} dG_t = \frac{\pi_G}{2-\pi_G},$$

while under a bank-based system

$$\int_0^{\underline{a}} dG_t = \frac{1-\pi_G}{2-\pi_G}, \int_{\underline{a}}^{\overline{a}} dG_t = \frac{1}{2-\pi_G}, \text{ and } \int_{\overline{a}}^{a^U} dG_t = 0.$$

Clearly, the second term in the expression for NLA_t is the same under either system and the third term is higher (in absolute value) under the bank-based system (as the resource drain due to monitoring cost is higher). We will establish that the first term is higher under the market-based system.

First, in a developed financial system, from $t > 0$ onwards all workers are simply offsprings of capitalists whose investments failed, so that $a_t^i = 0$. This is also true in steady state. Hence $\int_0^{\underline{a}} a_t^i dG_t = 0$ under both systems. Secondly, for a market-based system, in steady-state bank-dependent capitalists are grandchildren of capitalists whose investments failed. Parents of these bank-dependent capitalsts were workers and left them bequests a_t^i given by the vertical

intercept of point A in Figure 2(a). Hence in the market-based system (in steady state) $\int_{\underline{a}}^{\overline{a}} a_t^i dG_t = (1-\pi_G)/(2-\pi_G) \times$ (vertical intercept of A).

Now consider a bank-based system. Since there are $(1-\pi_G)/(2-\pi_G)$ workers in steady-state, out of the $1/(2-\pi_G)$ bank-dependent capitalists exactly $(1-\pi_G)/(2-\pi_G)$ have wealth a_t^i given by the height of point A.

Next consider the remaining $\pi_G/(2-\pi_G)$ bank-dependent capitalists in a bank-based system and the $\pi_G/(2-\pi_G)$ market-dependent capitalists in a market-based system. Comparing Figures 2(a) and 5 it is clear that each of the market-dependent capitalist is wealthier than any bank-dependent capitalist – the wealth of the market-dependent capitalists is distributed along the bequest line EF in Figure 2(a) whereas that of the bank-dependent capitalists is distributed along the bequest line CH in Figure 5.

Hence a market-based system yields higher GNP.

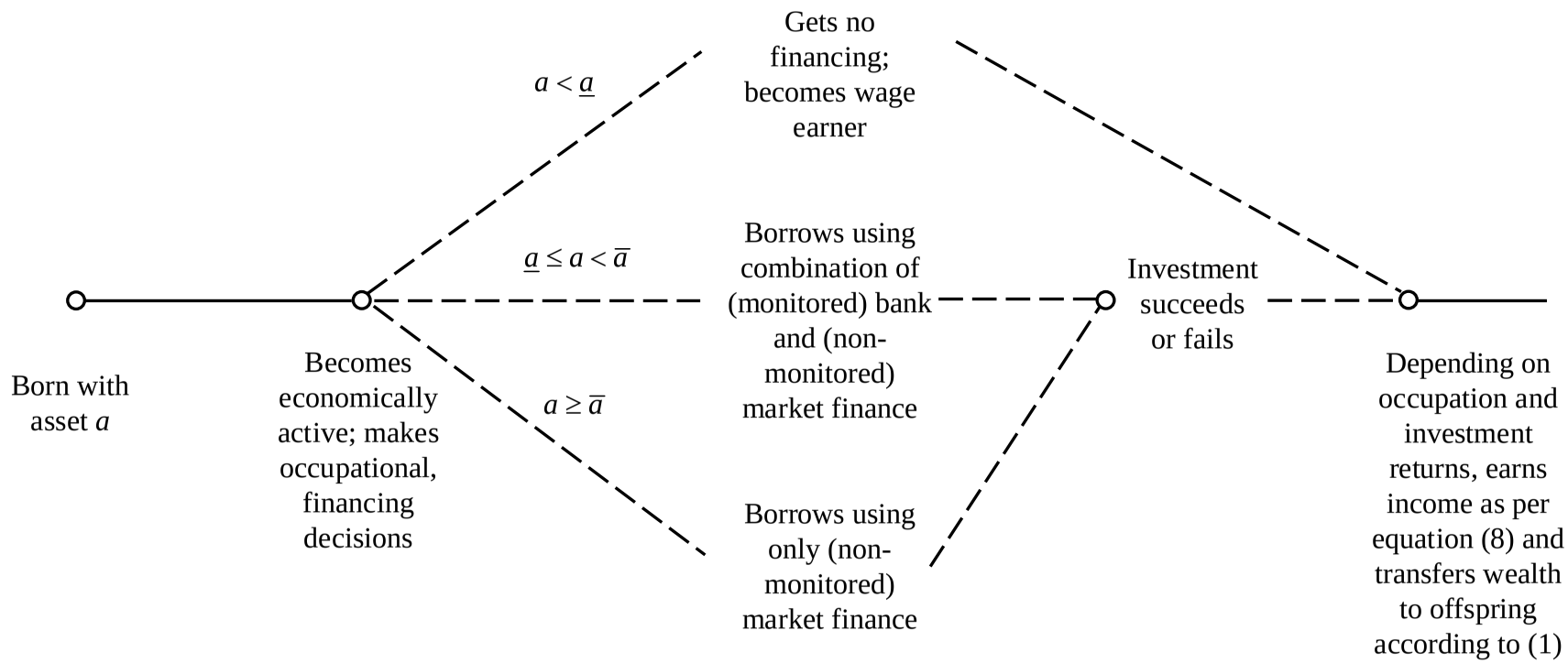

Figure 1: Timeline of Decisions for a Typical Agent

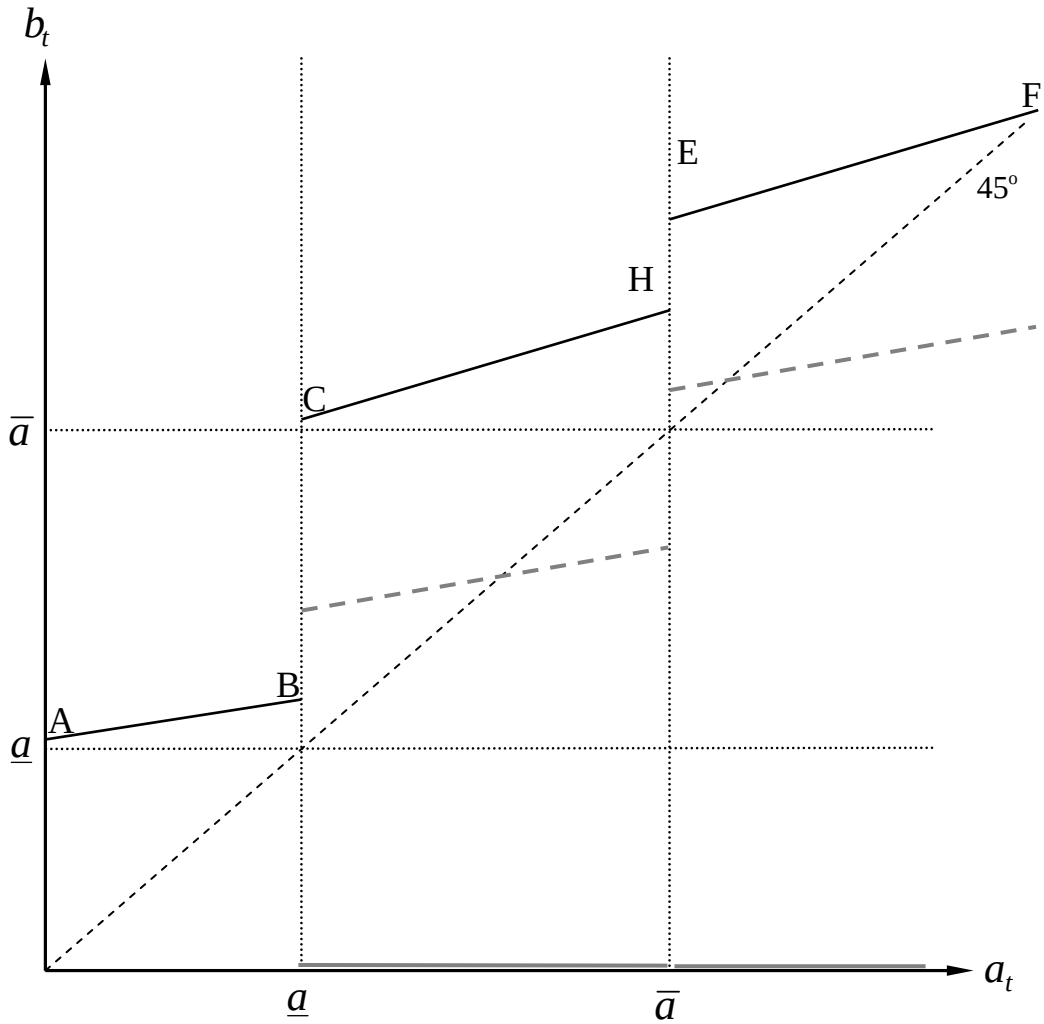

Figure 2(a): Recursion Dynamics of Wealth Accumulation when $f_1 \leq \underline{f_1}$

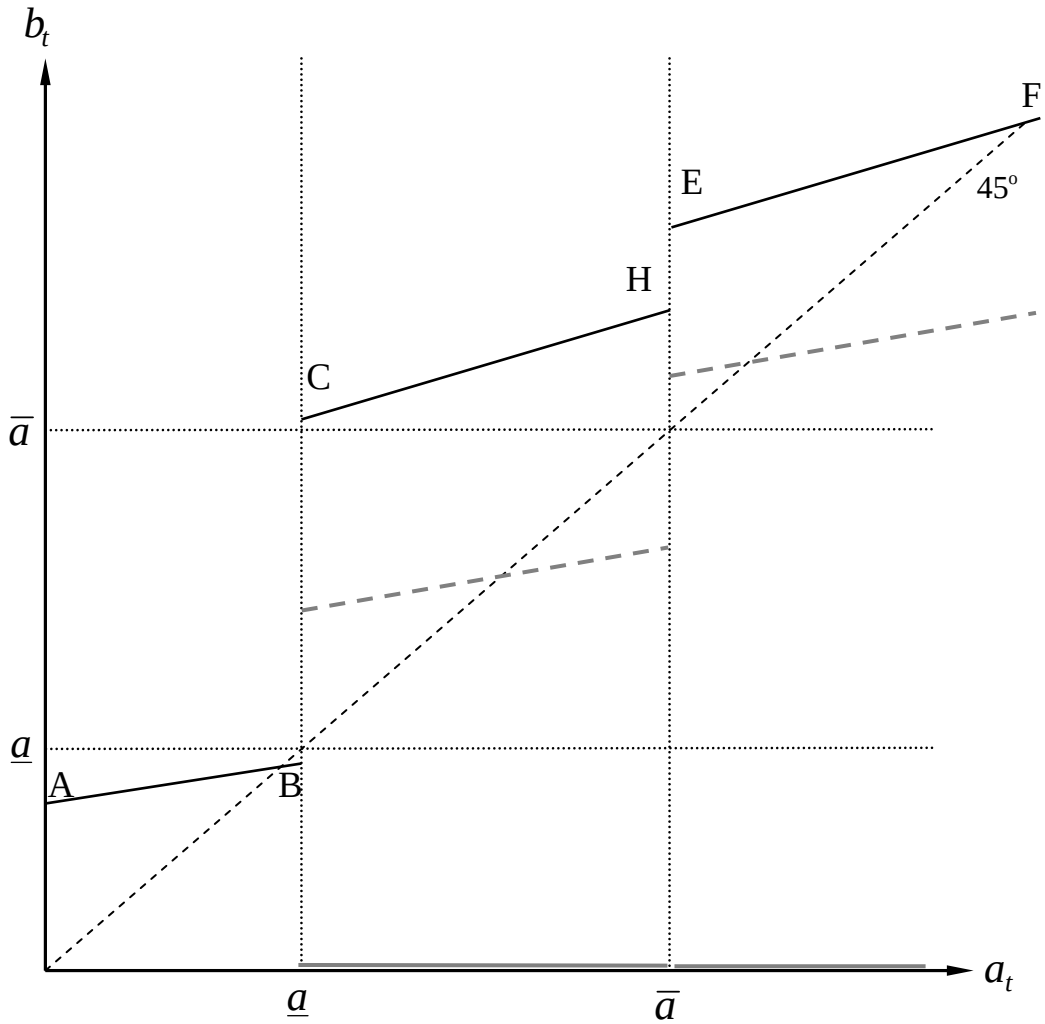

Figure 2(b): Recursion Dynamics of Wealth Accumulation when $f_1 > \overline{f}_1$

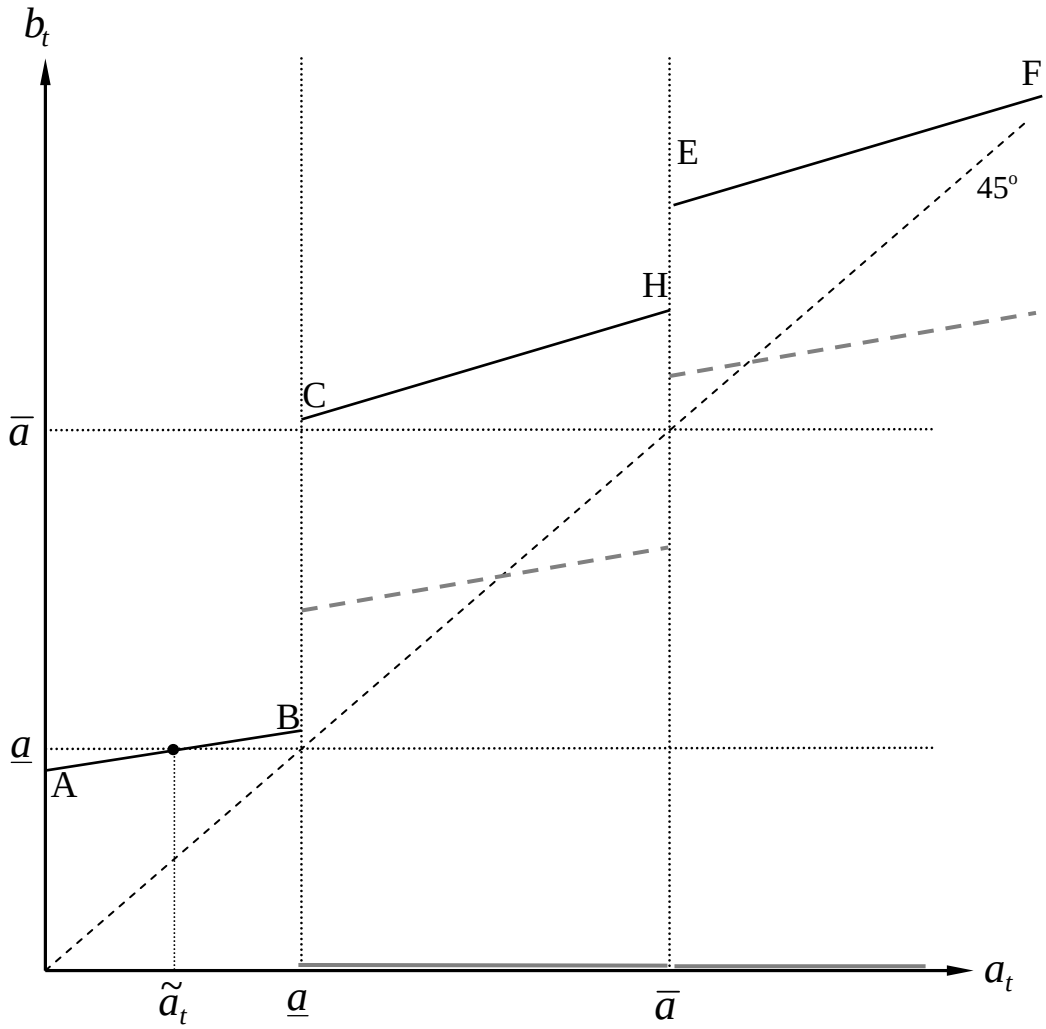

Figure 2(c): Recursion Dynamics of Wealth Accumulation when $\underline{f}_1 < f_1 < \overline{f}_1$

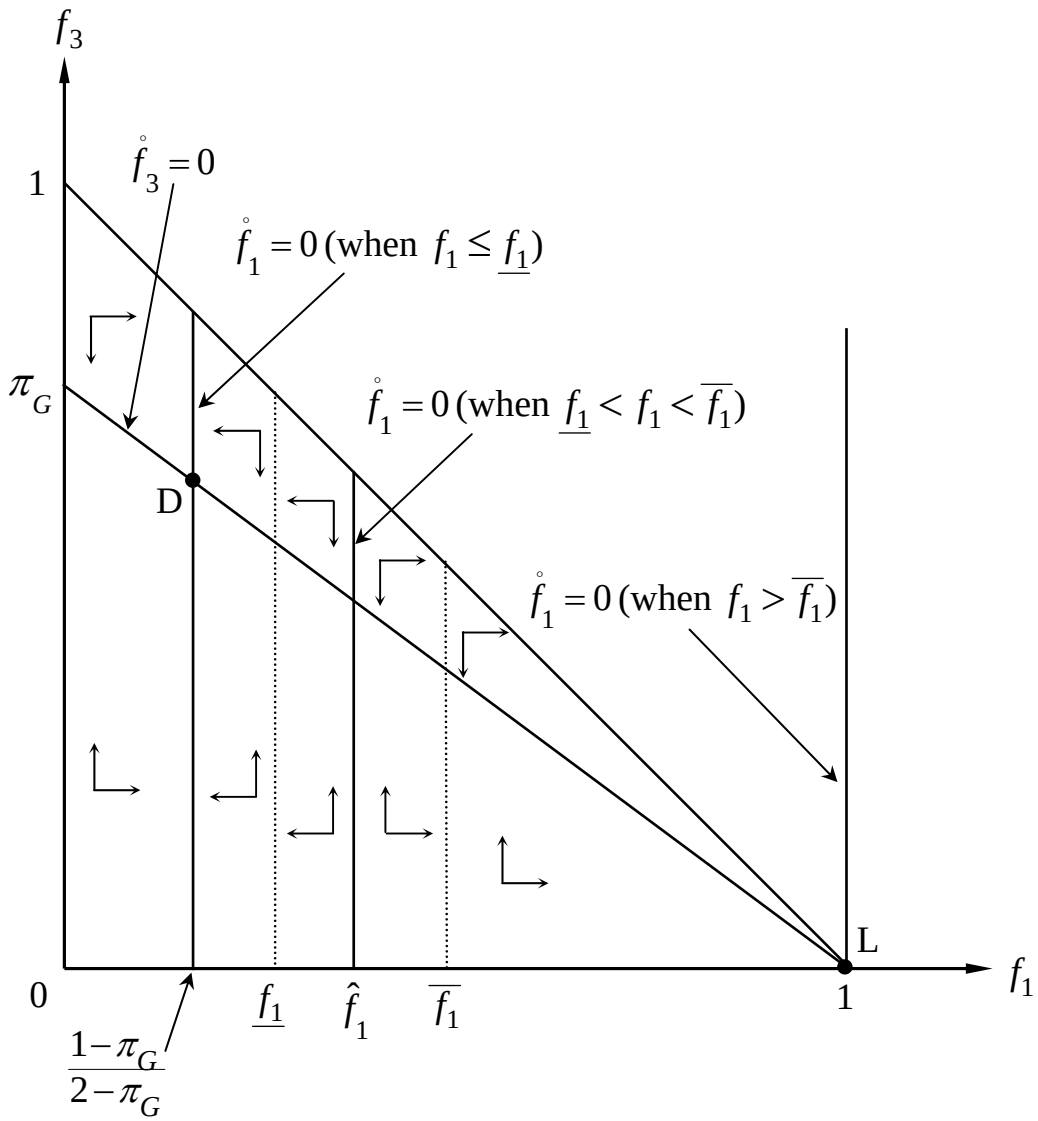

Figure 3(a): Phase Diagram for a single Threshold

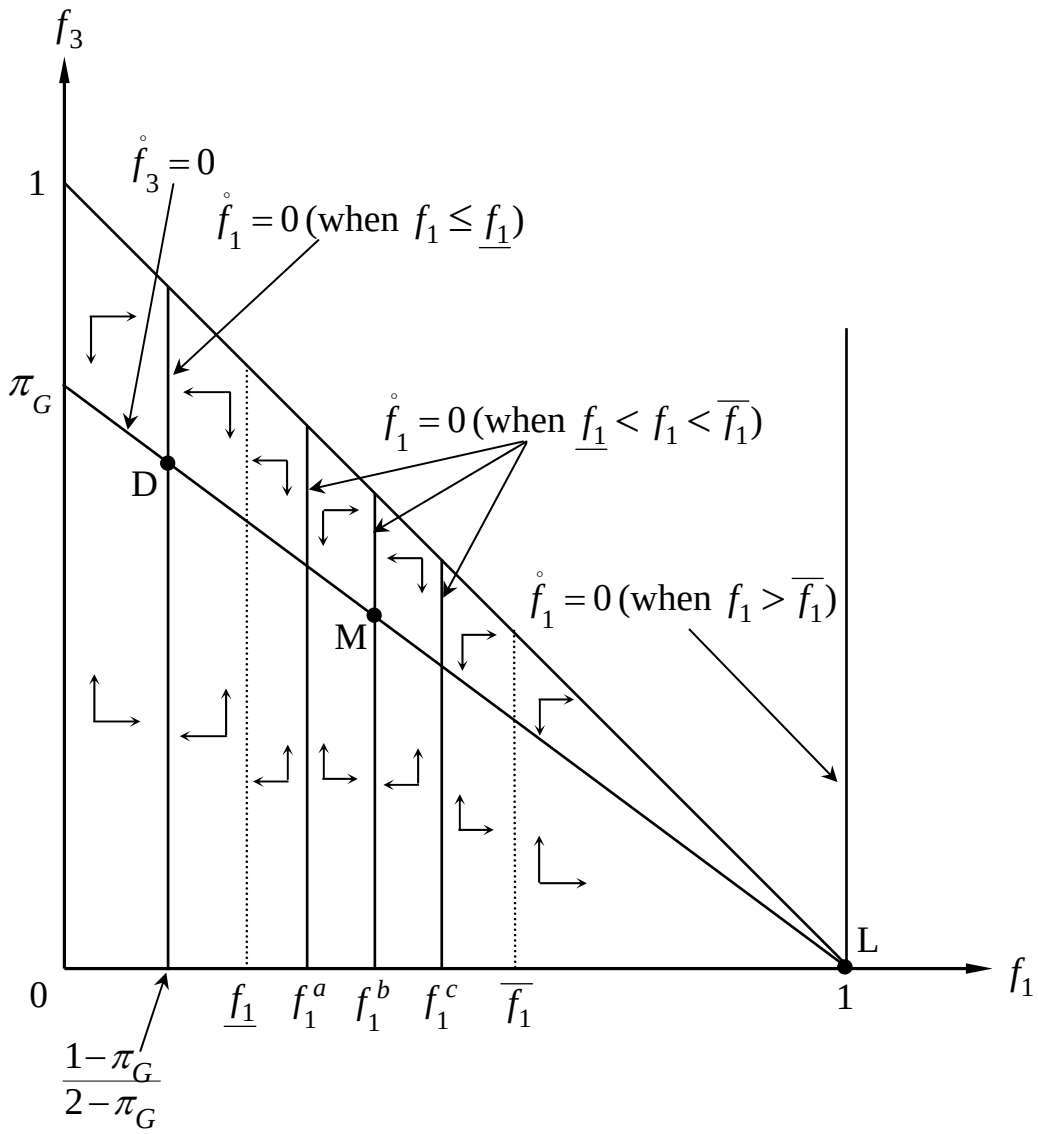

Figure 3(b): Phase Diagram for two Thresholds

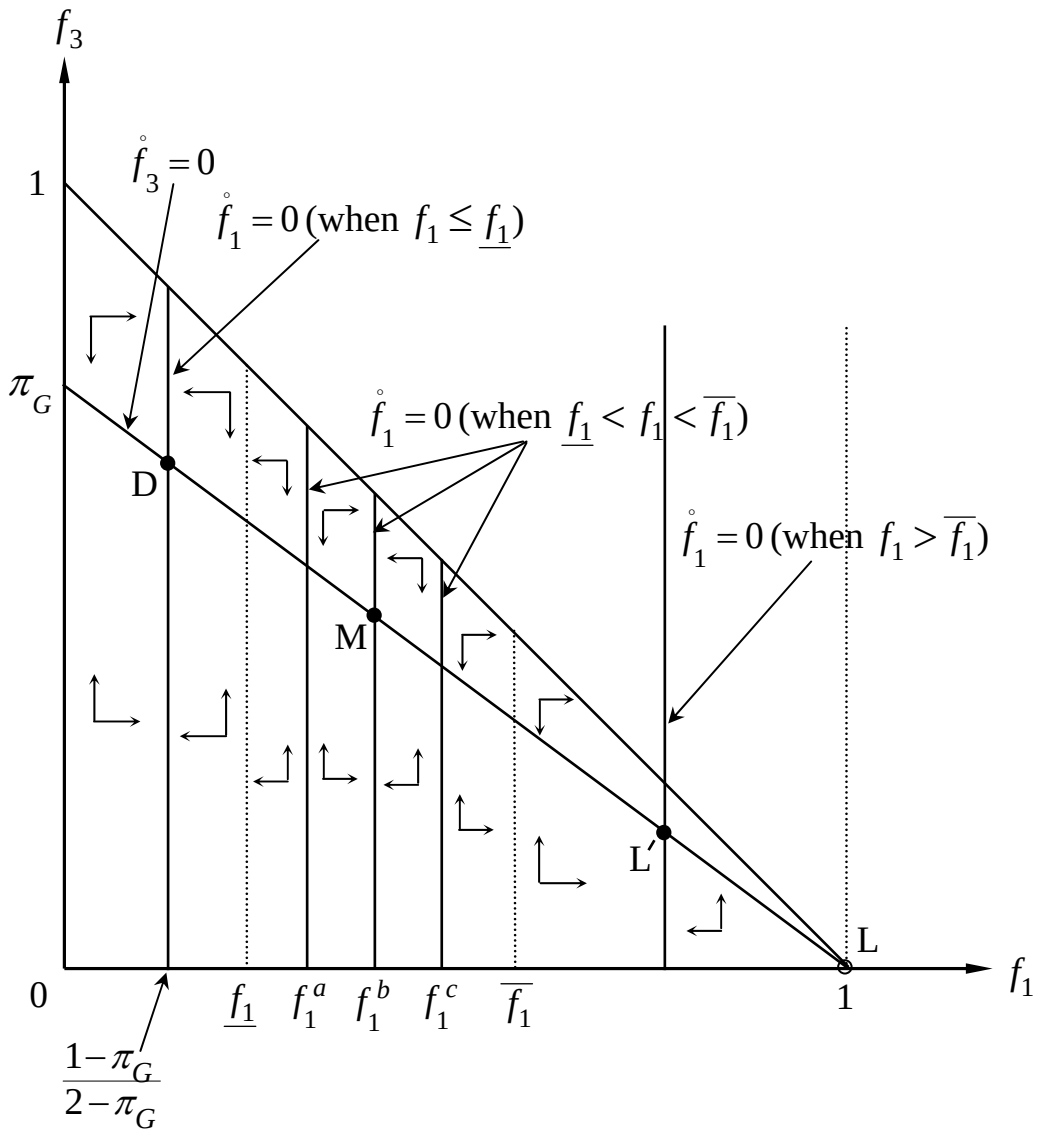

Figure 3(c): Phase Diagram with Perturbed Wealth Dynamics

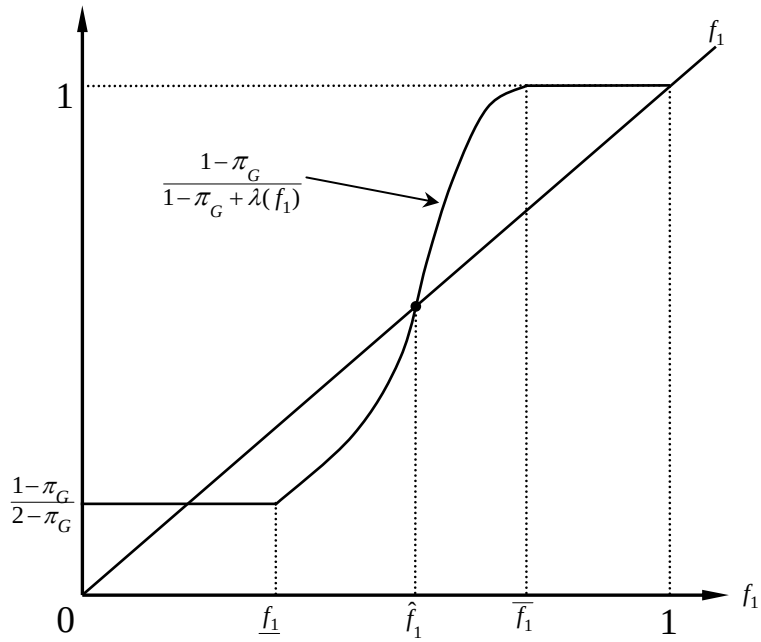

Figure 4(a)

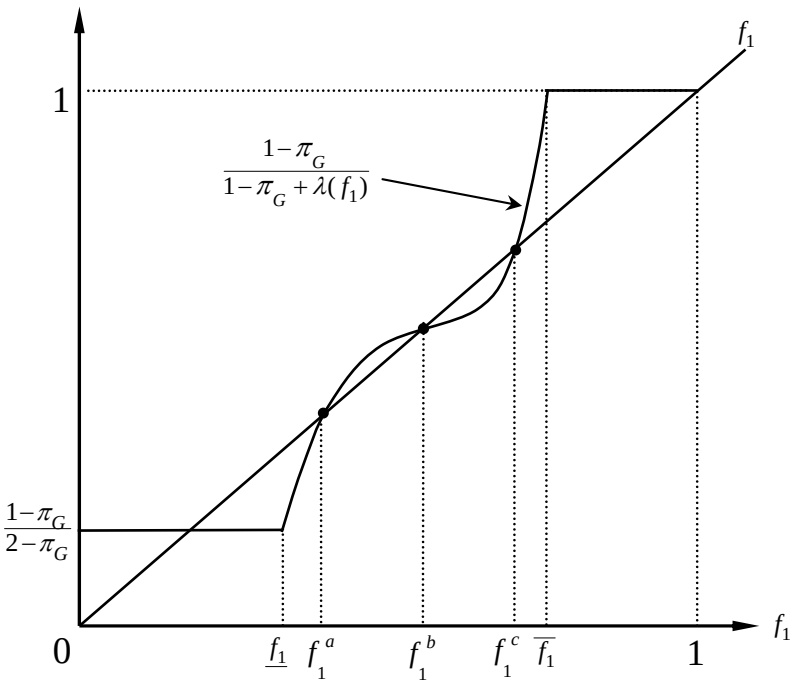

Figure 4(b)

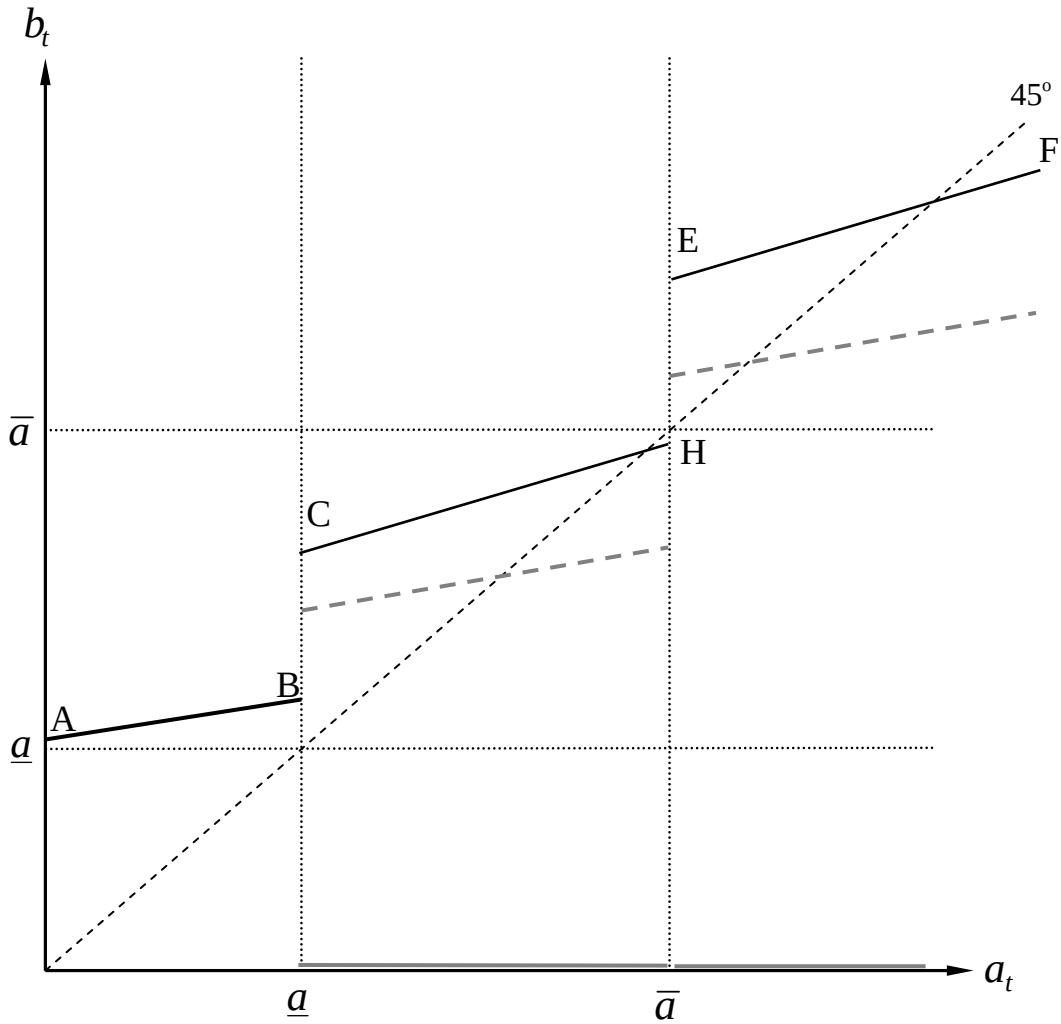

Figure 5: Wealth Dynamics with Low Return on Investment ($f_1 \leq \underline{f}_1$)

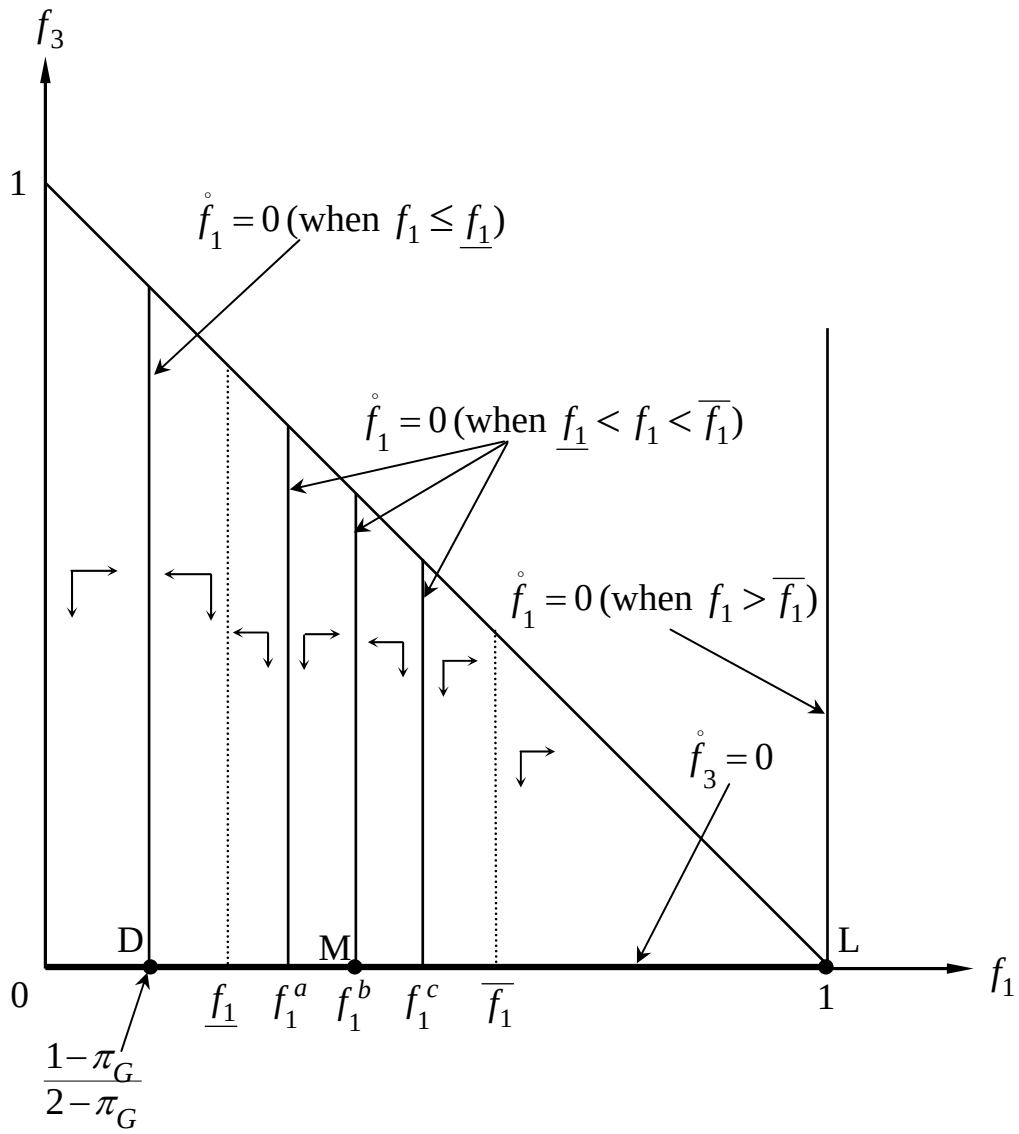

Figure 6(a): Phase Diagram (for two Thresholds) with Low Investment Return

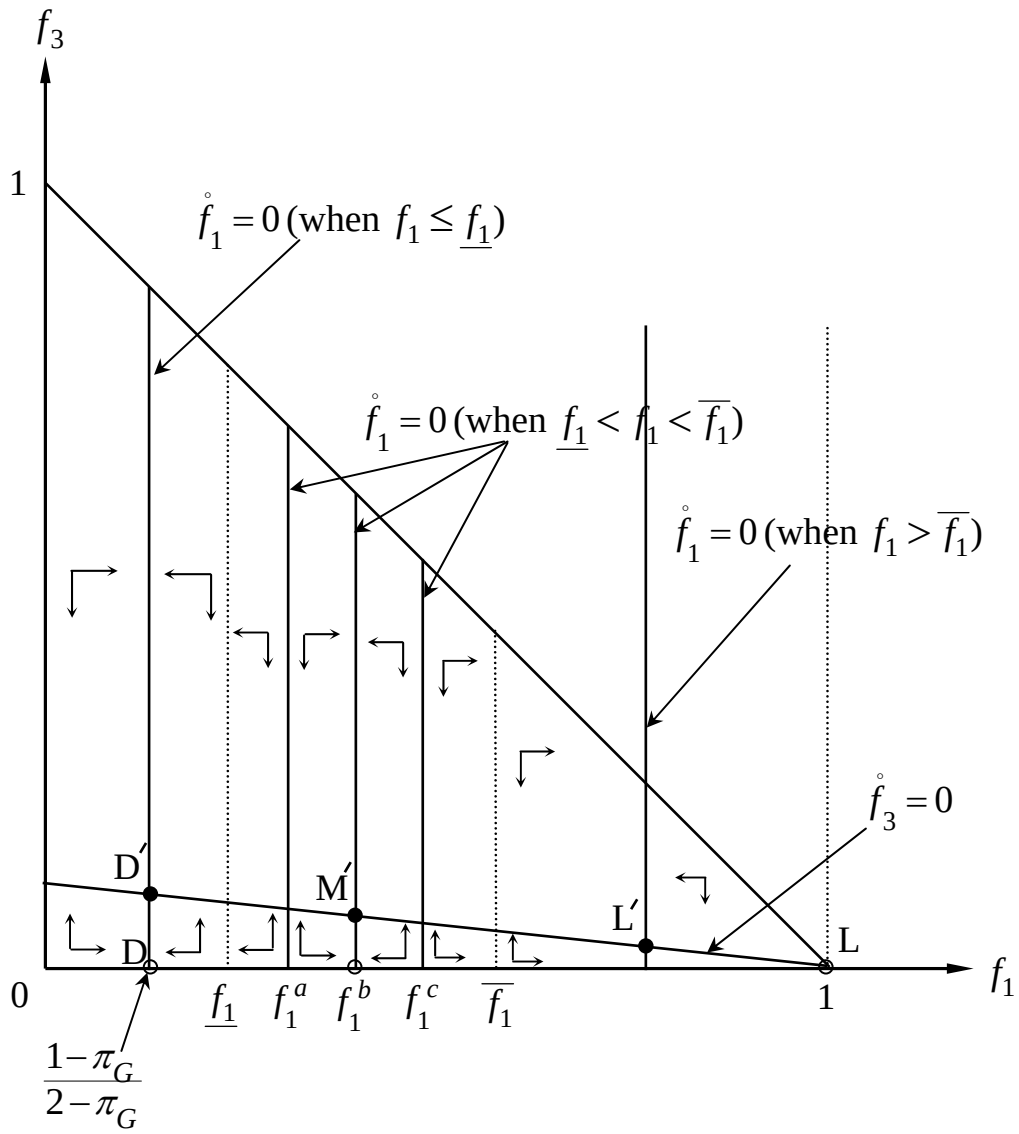

Figure 6(b): Phase Diagram for the Perturbed Wealth Dynamics with Low Investment Return