Essays in Macroeconomics and Corporate Finance

by

Jonathan E. Goldberg

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

© Jonathan E. Goldberg, MMXI. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part.

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Department of Economics
May 18, 2011

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Ricardo J. Caballero
Chairman and Ford International Professor of Economics
Thesis Supervisor

George-Marios Angeletos
Professor of Economics
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Esther Duflo
Chairman, Department Committee on Graduate Theses, and Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics
Abstract

This thesis examines questions at the intersection of macroeconomics and finance.

Chapter 1 studies the persistent effects of a decrease in firms’ ability to borrow. I develop a tractable model of deleveraging that emphasizes (i) firms as suppliers of financial assets to consumers and (ii) the ability of firms and consumers to alleviate financial frictions by accumulating wealth. In the model, a permanent decrease in the ability of firms to borrow leads to: increased capital misallocation and decreased total factor productivity (TFP); an increased wedge between the average marginal product of capital and the interest rate; and increased riskiness of consumption. An endogenous decrease in the interest rate is shown to amplify these effects by discouraging wealth accumulation. In a calibration using U.S. firm-level data, I find these amplification effects are large.

Chapter 2 studies how proprietary trading and advising are combined on Wall Street even though a firm that engages in both of these activities may be tempted to mislead its clients.

Chapter 3 studies the effects of government purchases of long-term debt. According to one interpretation, the preferred-habitat model of Vayanos and Vila (2009) implies that Federal Reserve purchases of long-term bonds generate a reduction in long-term interest rates. In this paper, I clarify this interpretation. In particular, in a Vayanos and Vila (2009) preferred-habitat model, I show that maturity-lengthening open-market operations have no effect on long-term interest rates if agents in the economy ultimately receive the profits from the government’s portfolio via lump-sum taxes or transfers. I then introduce limited participation - an assumption that some agents are restricted from trading bonds of certain or all maturities. I show that limited participation implies that open-market operations do reduce the long-term interest rate. What drives this result is limited participation, not preferred-habitat preferences. With this motivation, I develop a model, with a more reasonable form of limited participation and without preferred-habitat preferences, in which open-market operations are relevant. Finally, I use these models to discuss how arbitrageurs’ wealth covaries with technology or endowment shocks, and how this covariance is affected by open-market operations.

Thesis Supervisor: Ricardo J. Caballero
Title: Chairman and Ford International Professor of Economics

Thesis Supervisor: George-Marios Angeletos
Title: Professor of Economics
Acknowledgments

I am extremely grateful to my advisors George-Marios Angeletos, Ricardo Caballero and Guido Lorenzoni for their invaluable guidance and feedback. I also extend a special thanks to Daron Acemoglu, Bob Gibbons, Gustavo Manso, Roberto Rigobon, Jean Tirole and Iván Werning for their comments and encouragement. I have also benefitted tremendously from the advice and friendship of current and former MIT classmates including Fernando Duarte, Maya Eden, Florian Ederer, Luigi Iovino, Pablo Kurlat, Jen-Jen La’O, Marti Mestieri, Plamen Nenov, Mandy Pallais, Sahar Parsa, Michael Peters, Mar Reguant, Florian Scheuer, Jenny Simon, Alp Simsek and the entire 2006 entering class. I also thank Veronica Guerrieri, Adriano Rampini and my cousin and fellow economist Ariell Reshef.

Participants in the MIT macro and finance lunches provided very helpful comments throughout the research process. I also had the good fortune to present the research in Chapter 2 in the organizational and IO lunches.

Tim Mulcahy and Ayesha Mahmud of the National Opinion Research Center (NORC) at the University of Chicago were very helpful in facilitating access to the data used in Chapter 1.

This research was funded in part by the Ewing Marion Kauffman Foundation and the National Science Foundation Graduate Research Fellowship Program; the contents of this thesis are solely the responsibility of the author. I am thankful for the financial support provided by the Kauffman Foundation and the NSF.

This endeavor could not have been completed without the love and support of my family, for which I am forever grateful.
## Contents

1 A Model of Deleveraging  
  1.1 Introduction ................................................. 9  
  1.2 The environment .............................................. 16  
  1.3 Steady-state equilibrium characterization ..................... 21  
    1.3.1 The entrepreneur’s problem .............................. 21  
    1.3.2 The “bite” of entrepreneurs’ financial constraints ...... 23  
    1.3.3 Steady state definition ................................. 25  
  1.4 Steady state comparisons ...................................... 26  
    1.4.1 Constant productivity ................................. 27  
    1.4.2 Productivity independent across time ................. 33  
  1.5 Calibration and Numerical Results ................................ 38  
    1.5.1 Main numerical exercise: Decrease in firms’ ability to borrow . . . 42  
  1.6 Conclusion .................................................... 49  

2 Trading and advising on Wall Street  
  2.1 Introduction .................................................. 77  
    2.1.1 How trading and advising are combined on Wall Street .......... 81  
    2.1.2 Review of the literature ................................ 83  
  2.2 The model ..................................................... 85  
    2.2.1 Autarky ................................................. 86  
    2.2.2 Relational contracting ................................. 87  
  2.3 Benchmark Case: $\alpha = 1$ or $\sigma^2 = 0$ .................... 89  
    2.3.1 When clients cannot trade on the informed trader’s information . . . 94  
  2.4 The general model with $\alpha < 1$ and $\sigma_\epsilon > 0$ .......... 97
3 Preferred Habitat versus Limited Participation:

Irrelevance and Relevance Results for Open-Market Operations

3.1 Introduction ......................................................... 121

3.2 Related literature and modelling approach ................................ 124

3.3 Model with preferred-habitat investors ...................................... 129

3.4 Limited participation without preferred-habitat preferences ............. 135

3.4.1 Two-period version ................................................. 136

3.4.2 A model with wealth accumulation .................................. 137

3.4.3 Benchmark: Ricardian equivalence .................................. 138

3.4.4 Adding limited participation ....................................... 139

3.4.5 Government asset purchases, wealth accumulation, and shocks .... 143

3.5 Conclusion ........................................................... 148
Chapter 1

A Model of Deleveraging

1.1 Introduction

A decline in firms’ ability to borrow can lead to a reduction in firms’ supply of financial assets. The resulting scarcity of financial assets can amplify the distortions caused by the decline in firms’ ability to borrow. This scarcity can also adversely affect other agents, such as workers, who need financial assets.

To explore this idea formally, this paper builds a tractable dynamic model of the persistent effects of a permanent decrease in firms’ ability to borrow. In the model, firms and workers can overcome financial frictions by accumulating liquid wealth. However, the scarcity of financial assets limits the extent to which they do so. A decline in firms’ ability to borrow leads to greater scarcity of financial assets and a decrease in the interest rate. This discourages the accumulation of wealth and results in larger investment distortions for firms and increased riskiness of consumption for workers. These amplification and spillover effects underscore the importance of a general equilibrium framework for analyzing the economy’s response to deterioration in the quality of the financial system.

In the model, workers use financial assets as self-insurance against income risk. The decline in firms’ ability to borrow requires workers to hold fewer financial assets and thus workers are less able to smooth consumption.

Economists are very familiar with how shocks to firms affect consumers through general equilibrium in the labor market: if firms want to hire more labor, wages increase and workers benefit. This paper emphasizes how shocks to firms, such as a credit crunch, affect consumers through the financial market.
By understanding how firms and workers are affected by scarcity of financial assets and low interest rates, this model can shed light on how the economy responds to a shift from loose credit for firms to tight credit. Such a shift may accompany financial crises and explain the long-lasting reductions in output and leverage that typically follow financial crises.

The dynamic accumulation of liquid wealth is central to the model. In the model, there are two forms of liquid wealth: financial assets and physical capital. Workers invest only in financial assets, because they do not have access to a technology in which to invest physical capital. Firms, which are run by entrepreneurs, can invest capital productively, and hence they are natural suppliers of financial assets to the workers. However, entrepreneurs also need liquid wealth, because they face moral hazard and liquid wealth allows them to create the collateral required to take advantage of investment opportunities.

An entrepreneur can always choose to accumulate more liquid wealth, by reducing consumption and investing in capital or supplying fewer financial assets. By accumulating more liquid wealth, the entrepreneur would be less constrained in the future. Therefore, a key question for the long-run properties of the economy is the extent to which entrepreneurs choose to accumulate sufficient liquid wealth to overcome their borrowing constraints. For an individual entrepreneur, the answer depends on the equilibrium interest rate. I show that if the interest rate is equal to the rate of time preference, an entrepreneur eventually accumulates enough liquid wealth to completely overcome the borrowing constraints. In contrast, if the interest rate is less than the rate of time preference, then at least some of the time, the entrepreneur will have a binding borrowing constraint.

I also show that, due to this endogenous wealth-accumulation channel, steady-state investment distortions are decreasing in the interest rate. One measure of investment distortions that I consider is the difference between total factor productivity (TFP) and the TFP that would obtain if there were no borrowing constraints. Another measure that I study is the wedge between the marginal product of capital and the interest rate; this wedge is the finance premium firms would be willing to pay for an additional unit of borrowing. Investment distortions are decreasing in the interest rate because taking advantage of investment opportunities requires holding liquid wealth, and holding liquid wealth is more costly when the interest rate is lower.
For the same reason, the riskiness of workers’ and entrepreneurs’ consumption is decreasing in the interest rate. When the interest rate falls, workers and entrepreneurs alike must bear more consumption risk in order to reduce the costs of holding liquid wealth.

Thus, the long-run properties of the economy depend crucially on the equilibrium interest rate and any analysis of a decrease in firms’ ability to borrow should take into account the liquidity environment and how the interest rate changes. In a baseline scenario where workers are assumed to be hand-to-mouth consumers who do not save, the unique steady-state equilibrium interest rate is equal to the rate of time preference. Hence, a decrease in firms’ ability borrow has no long-run effect on the wedge between the marginal product of capital and the interest rate. In constrast, when workers are allowed to accumulate financial assets, the steady-state interest rate is below the rate of time preference. As in Aiyagari (1994) and other Bewley-like models, this is due to the precautionary motive for savings: if the interest rate were greater than or equal to the rate of time preference, then workers would seek to accumulate infinite wealth, which in turn would be inconsistent with equilibrium. Unlike in Aiygari (1994), however, the workers’ precautionary demand for financial assets in my model opens the door to an important general-equilibrium interaction between the workers and the firms: a decrease in firms’ ability to borrow leading to increased investment distortions due to a decreased interest rate.

The financial constraints on firms in this economy arise endogenously, due to the moral hazard that firms face. In particular, if an entrepreneur reneges on a promise to repay, creditors cannot seize all of the entrepreneur’s income and assets. As a result, entrepreneurs can pledge to investors only a fraction of their income and undepreciated capital.

Importantly, there are no financial constraints on firms except for those that arise endogenously due to this moral hazard. Firms can issue state-contingent assets and assets of any maturity. Thus, this paper focuses on a single source of financial constraints – the moral hazard of reneging – and all of the results can be traced clearly to this moral hazard. Moreover, the model maintains its tractability even though firms can choose any maturity structure.

The financial constraints due to moral hazard limit the ability of firms both to take advantage of investment opportunities and to diversify their idiosyncratic risks. The accumulation of wealth can alleviate the bite of financial constraints along both of these dimensions.
In the model, a decline in firms’ borrowing ability leads to a decrease in firms’ supply of financial assets at a given interest rate. This occurs for two reasons. The first is a direct effect: conditional on the firms’ capital choices, the maximum amount of financial assets that firms can supply decreases. The second is an indirect effect: conditional on the interest rate, a decrease in firms’ ability to borrow implies firms need more liquid wealth to achieve a given level of investment, and because holding wealth is expensive, firms respond by choosing lower capital.

The general equilibrium amplification and spillover effects identified here will be stronger when the decrease in firms’ ability to borrow affects a large economy (like the U.S.) or an economy that is less open to international financial markets.

In a numerical analysis using firm-level data for U.S. entrepreneurs, I study a reduction in firms’ borrowing ability that is consistent with the recent post-crisis decline in real long-term interest rates. I calculate that this decrease in firms’ ability to borrow results in significant increases in investment distortions for firms and consumption riskiness for workers. General-equilibrium amplification plays a meaningful role. For instance, in general equilibrium, TFP losses – the gap between TFP and the TFP that would obtain if firms were unconstrained – increase by 29 percent. If the interest rate were held constant, as in a small open economy, TFP losses would increase only 13 percent.

A decrease in firms’ ability to borrow is one potential explanation of the persistent reductions in output and leverage associated with financial crises. Reinhart and Reinhart (2010) find that financial crises are followed by large and long-lasting declines in the ratio of domestic bank credit to GDP. For example, four Asian emerging-market countries hit by the 1997 crisis required a full decade or more before the ratio of bank credit to GDP bottomed. Across the four countries, bank credit to GDP one decade after the crisis was, on average, 46 percentage points lower than its highest level around the crisis. Rajan and Zingales (2003) analyze measures of financial development such as the ratio of bank deposits to output and the ratio of equity issuance to capital formation. By such measures, countries were less financially developed in 1980 than in 1913, before the Great Depression, and only surpassed their 1913 levels of financial development in the 1990s. IMF (2010) finds that seven years after a financial crisis, output and total factor productivity are, respectively, about 10 and 4 percentage points below levels predicted from a pre-crisis trend.
Further results. The core of my analysis focuses on the long-run response of the economy to an adverse shock to the financial constraints faced by firms. While this seems to be one of the key aspects of financial crises, there are other changes in the economy that may accompany a financial crisis. For example, the economy appears to be subject to more volatility at both the macro and the micro level. My model provides a useful framework for understanding either how these changes could emerge endogenously, or how they could be amplified and propagated in the economy because of the aforementioned general-equilibrium interaction through asset markets.

Consider an increase in the volatility of idiosyncratic labor-income risk. Storesletten, Telmer and Yaron (2004) provide evidence that this volatility increases during recessions.\(^1\) An increase in labor-income risk leads consumers to increase their savings, resulting in a decrease in the interest rate. Aggregate capital increases. These predictions are present in my model and in the model of Aiyagari (1994). However, my model features additional effects: because the lower interest rate leads entrepreneurs to supply more financial assets, there is a decrease in productive efficiency and an increase in the average wedge between the marginal product of capital and the interest rate. These effects are not present in standard macroeconomic models, including models like Aiyagari (1994) that are able to capture an increase in labor-income risk. The analysis for an increase in consumer risk aversion is qualitatively similar.

This shows how a shock to labor-income risk can have an amplified effect on consumption riskiness and can have spillover effects on firms. The model also suggests how increased labor-income risk could arise endogenously from a decrease in firms’ ability to borrow. For example, suppose that a worker’s idiosyncratic labor-income risk originates from the idiosyncratic risk that his employer faces.\(^2\) An extension of the model along these lines could then explain the increase in workers’ unemployment or labor-income risk as a direct consequence of the endogenous increase in the riskiness of firms.\(^3\)

Related literature. This paper is related to the long literature on the macroeconomic

---

\(^1\)Krueger and Perri (2006) provide evidence that this volatility has been increasing at low frequency in recent decades.

\(^2\)This could be due to search frictions in the labor market or firm-specific human-capital investments that tie the fate of a worker to the fate of his employer.

\(^3\)However, if workers’ labor-income is tied to the fate of their employer, and the employer issues state-contingent assets, as in this model, one would have to take into account the workers’ ability to reduce their idiosyncratic risk by short-selling the equity-like securities. In practice, workers do not seem to pursue this strategy (Massa and Simonov 2006), although there appear to be large gains from doing so.
effects of financial frictions that dates back to the work of Keynes and Fisher. Since then, economists have developed models that emphasize the importance of liquid wealth for overcoming borrowing constraints. Early work, such as Evans-Jovanovic (1989), featured a static model in which liquid wealth was exogenous and wealth accumulation played no role. Kiyotaki-Moore (1997) incorporated a dynamic framework, but one without wealth accumulation.

In contrast, more recent work by Holmstrom-Tirole (1998) and Albuquerque-Hopenhayn (2004) features a crucial role for wealth accumulation as a means of overcoming financial frictions. Holmstrom-Tirole (1998), like this paper, emphasize general-equilibrium limits to liquidity accumulation.

Wealth accumulation also plays a central role in recent contributions to the entrepreneurship literature in macroeconomics, such as Quadrini (2000), Angeletos-Calvet (2006), Cagetti-DiNardi (2006), Buera (2009), Banerjee-Moll (2010), Buera-Kaboski-Shin (2010) and Moll (2010). Indeed, for some of these papers, the empirical motivation is study how financial frictions shape the wealth distribution, rather than how the distribution of liquid wealth affects financial frictions.

This paper also builds on previous work on limited enforcement and endogenous solvency constraints, including Kehoe-Levine (1993), Alvarez-Jermann (2000), Lorenzoni-Walentin (2007), Lorenzoni (2008) and Rampini-Viswanathan (2010a,b). A common theme of this literature that also appears in this paper is that endogenous solvency constraints limit risk sharing and depress the interest rate. One paper in this literature, Cooley-Marimon-Quadrini (2004), studies the aggregate effects of limited enforcement; however, Cooley-Marimon-Quadrini (2004) features a constant interest rate and hence the endogenous changes in the interest rate that are crucial in my analysis are absent in their paper.

The two papers most closely related to this paper are Midrigan-Xu (2010) and Buera-Shin (2010a). Midrigan-Xu (2010)’s primary contribution is to thoughtfully parametrize the borrowing friction and other elements of their model using plant-level data about inputs and revenue for manufacturing firms in Korea and Colombia and aggregate data about leverage for manufacturing firms in these countries. As an exercise, Midrigan-Xu (2010) study a permanent decrease in Korean firms’ borrowing ability sufficient to reduce a leverage ratio for Korea to that for Colombia.

One important difference between this paper and Midrigan-Xu (2010) is that Midrigan-
Xu (2010) assume a small open economy and hence in their paper the interest rate is exogenous and unaffected by the decrease in firms’ ability to borrow. In contrast, the central focus of this paper is the general equilibrium change in the interest rate and its amplification effects. Of course, assuming that Korea is a small open economy may be a suitable modeling choice for their question, whereas this paper contemplates a shock affecting a large economy like the U.S. where general equilibrium is important.

Buera-Shin (2010a) compare the transition dynamics associated with three reform scenarios: removal of exogenous idiosyncratic distortions; removal of exogenous distortions combined with an increase in firms’ ability to borrow; and removal of idiosyncratic distortions combined with capital-account liberalization, which increases the interest rate from the general-equilibrium closed-economy rate to the world interest rate. In Buera-Shin (2010a), as in this paper, an increase in firms’ ability to borrow is associated with a higher steady-state interest rate and higher measured TFP. However, because the central exercises of their paper are meant to capture reform scenarios, the main numerical results reflect the variety of economic forces unleashed by the multi-faceted reforms.

One contribution of this paper to the existing entrepreneurship literature is to focus on the role of entrepreneurs as suppliers of financial assets to workers and the spillover effect on workers from a change in the interest rate when entrepreneurs’ ability to borrow decreases. This paper derives several theoretical results that provide a clearer understanding of several issues, such as the role of general-equilibrium changes to the interest rate. Also, the financial constraints in this paper arise endogenously from a single friction, the ability of firms to partially renege on their promises. In contrast, many papers in the entrepreneurship literature, such as Angeletos-Calvet (2006) or Cagetti-DiNardi (2006), feature multiple and ad-hoc forms of market incompleteness, including a restriction to safe debt with one-period maturity.

Much of the literature on entrepreneurship is concerned with cross-country differences in financial development. This highlights an alternative interpretation of many of my results, which is to explain how differences in financial development result in differences across countries in the levels of investment distortions and the abilities of workers to smooth consumption.

Lorenzoni and Guerrieri (2010) also study the effects of a credit crunch. In particular, Lorenzoni and Guerrieri (2010) study the response of an economy of Bewley-Aiyagari
consumers to a permanent, unexpected tightening of their borrowing constraint. Like this paper, Lorenzoni and Guerrieri (2010) emphasizes precautionary saving and the scarcity of liquid assets. One important difference between these papers is that Lorenzoni and Guerrieri (2010) focus on changes in the borrowing constraints of consumers, whereas this paper emphasizes shocks to the borrowing constraints of entrepreneurs. Although my paper also includes an analysis of shocks to Bewley-Aiyagari consumers, the focus in my paper is on the general-equilibrium interactions between firms and workers. In their model, in contrast, the production side is frictionless.

Jermann-Quadrini (2009), like this paper, emphasizes the macroeconomic effects of financial shocks. Earlier work, such as Bernanke-Gertler (1989) and Bernanke-Gertler-Gilchirst (1999), examined how financial frictions affect short-run fluctuations.

This paper is also related to Mendoza-Quadrini-Ríos-Rull (2009) and Angeletos-Panousi (2010), which attempt to explain current-account dynamics by investigating entrepreneurs’ precautionary saving.

The workers in this paper cannot insure their labor-income risk. This assumption is the focus of the income fluctuations problem in macroeconomics. This paper draws on key contributions to this literature by Bewley (1983), Clarida (1990), Aiyagari (1994) and Chamberlain-Wilson (2000). In the numerical analysis, I find a wedge between the inverse of the discount rate and the interest rate that is much larger than the wedge found in Aiyagari (1994). I show how this difference can be attributed to the inclusion of frictions for firms, which are absent in Aiyagari (1994).

1.2 The environment

There are two types of agents in the economy, entrepreneurs and workers.

There is only one good in the economy. It is produced by entrepreneurs. The entrepreneurs use the good and labor as inputs for production.

**Entrepreneurs**

In period $t$, an entrepreneur produces $F(k_{t-1}, l_t, s_t)$, where $k_{t-1}$ is capital, $l_t$ is labor, and $s_t$ is the entrepreneur’s idiosyncratic productivity.

Productivity $s_t \in S$ follows a Markov process with transition function $Q$. $S$ is discrete. The probability that an entrepreneur with productivity $s_t$ obtains a productivity $s_{t+1}$ to-
morrow is given by $Q(s_t, s_{t+1})$. I assume that $Q$ is monotone in the usual sense.\(^4\)

In period $t$, an entrepreneur has a history of idiosyncratic productivity shocks $s^t = \{s_0, s_1, ..., s_t\}$ and invests $k(s^t)$ in the production technology. In period $t + 1$, the entrepreneurs learns productivity $s_{t+1}$. The entrepreneur then hires labor $l(s^{t+1})$. The total output produced in period $t + 1$ is $F(k(s^t), l(s^{t+1}), s_{t+1})$.

I assume that $F(k, l, s)$ has the standard neoclassical properties except for constant returns to scale.\(^5\) In place of constant returns to scale, I assume that $F$ has decreasing returns to scale in capital and labor, consistent with the notion of limited “span of control” for entrepreneurs (Lucas 1978). I make the technical assumption that $F$ is strictly concave in capital and labor and that $F(., s)$ is increasing in $s$.

In period $t + 1$, the value of undepreciated capital is $(1 - \delta)k(s^t)$.

The labor market is competitive. The wage rate in period $t + 1$ is given by $\omega_{t+1}$. Since the economy features only idiosyncratic shocks, the process for wages will be deterministic.

The entrepreneur can access financial markets by trading state-contingent promises that pay out conditional on the realization of productivity. At history $s^t$, the entrepreneur sells $d(s^{t+1})$ Arrow-Debreu securities that represent a promise to pay one if state $s_{t+1}$ is realized. The unit price in period $t$ is $p(s^{t+1})$.

Although the promises are state-contingent, markets are incomplete because entrepreneurs

\(^4\)The monotonicity of $Q$ means that $\sum_{s \in S} h(s)Q(s, ., s)$ is non-decreasing in $s_{-}$ for any non-decreasing function $h : S \rightarrow \mathbb{R}$. Since $S$ is discrete, it is immediate that the Feller property also holds.

\(^5\)That is, $F_k > 0$, $F_l > 0$, $F_{kk} < 0$, and $F_{kl} < 0$, for all $k > 0$, $l > 0$ and all $s \in S$. In addition, the Inada conditions hold for capital and labor. The Inada conditions for capital are: $\lim_{k \rightarrow 0} F(k, l, s) = \infty$ for all $l > 0, s > 0$; and $\lim_{k \rightarrow \infty} F(k, l, s) = 0$ for all $l$ and all $s$. Moreover, $F(k, l, 0) = 0$.

\(^6\)For a proof, please see the appendix.
can renege on payment. In particular, if an entrepreneur reneges, the most that creditors can seize is a fraction $\theta \leq 1$ of the entrepreneur’s output net of labor costs. When an entrepreneur reneges, the unmet portion of the entrepreneur’s debt is erased. (When the entrepreneur is allowed to trade promises due in more than one period, all future promises by the entrepreneur are also erased). Given this, the entrepreneur will keep his promise to pay $d(s_{t+1})$ if and only if

$$d(s_{t+1}) \leq \theta f(k(s^t), s_{t+1})$$

(1.1)

This constraint can limit both the ability to borrow and the ability to diversify risks.

In Appendix B, I show that it is without loss of generality to restrict attention to one-period promises in this environment. Any allocation that can be implemented with any maturity structure can also be implemented with one-period promises. This generality contrasts with many papers, such as Angeletos-Calvet (2006) or Cagetti-DiNardi (2006), that exogenously impose a restriction to one-period debt.\(^7\)

By allowing state-contingent promises, I am able to focus exclusively on a single financial friction, the possibility of reneging. In reality, entrepreneurs can make state-contingent promises in a variety of ways. Through default or renegotiation, putatively non-contingent debt becomes state-contingent. Entrepreneurs can also sell equity stakes in their businesses. Once an entrepreneur is no longer the sole owner of the business, the entrepreneur may be able to separate payments to herself and to other equity holders by paying herself a salary.

It should be noted that with $\theta = 1$, the entrepreneur’s financial constraint can still bind and limit the financial decisions of the entrepreneurs. However, it will be shown later that with $\theta \geq 1$, an entrepreneur invests in capital as if there were no financial constraint.

The entrepreneur’s budget constraint at state $s^t$ is:

$$c(s^t) + k(s^t) - \sum_{s^{t+1} \mid s^t} p(s^{t+1})d(s^{t+1}) \leq f(k(s^{t-1}), s_t) - d(s^t)$$

(1.2)

The entrepreneur’s choices must also satisfy a no-Ponzi condition.

An entrepreneur's liquid wealth in period $t + 1$ is defined as $w(s^{t+1}) = f(k(s^t), s_{t+1})$ -

\(^7\)Of course, in reality, maturity structure does matter, as highlighted by Broner-Lorenzoni-Schmukler (2008) and the recent financial crisis. One reason maturity structure matters is that, when the interest rate is stochastic, non-contingent bonds of different maturities facilitate risk sharing, as emphasized by Angeletos (2002) and Buera-Nicolini (2004). These concerns are absent here, however, because the environment in this paper features state-contingent debt and a deterministic path for the interest rate.
Workers. Workers have a stochastic endowment of labor. In period $t$, a worker’s labor endowment is $z_t$ and her labor income is $\omega_t z_t$, where $\omega_t$ is the wage.

Labor productivity $z_t$ follows a Markov process with transition function $Q_Z$. I assume that $Q_Z$ is monotone and has the Feller property.

Consistent with Bewley (1983) and Aiyagari (1994), workers can only trade a safe asset; they cannot insure against labor endowment shocks using assets with a payoff linked to their labor process.\footnote{This can be endogenized as in Allen (1985) and Cole-Kocherlakota (2001), which study environments where workers’ income shocks are unobservable and they can privately store resources.}

Again, because there are no aggregate shocks, the path for the gross interest rate, $\{R_t\}_{t=0}^\infty$, will be deterministic.

The budget constraint of the workers is:

$$c(z_t^t) + \frac{1}{R_t} a(z_t^t) = a(z_t^{t-1}) + w_t z_t$$

Workers face a borrowing limit:

$$a(z_t^t) \geq a$$

As in Aiyagari (1994), this could be a natural or ad-hoc borrowing limit.

Financial market. Each firm takes a set of state-contingent positions in the financial market.

Consider a firm with history $s^t$. The firm makes promises $d(s^{t+1})$ with payout contingent on the realization of $s_{t+1}$. At $t$, it is known that the expected amount of these payouts in $t+1$ will be

$$E[d(s^{t+1})|s^t]$$

where the expectation is integrating over the realization of $s_{t+1}$.

Of course, in the economy, there are firms with different histories. Because the payouts are contingent on idiosyncratic shocks, by a law of large numbers, the total payout from the entrepreneurs is certain even before the idiosyncratic uncertainty is realized. At $t$, it is known that the total payouts by entrepreneurs in $t+1$ will be
\[ E[d(s^{t+1})] \]

where the expectation is integrating over all histories \( s^{t+1} \).

The financial market could be organized as follows: firms sell claims that are contingent on the firms’ idiosyncratic shocks and consumers buy a portfolio of claims; since the consumers will not want to be exposed to entrepreneurs’ idiosyncratic risk, they will buy a diversified pool of claims that amount to a safe asset. Alternatively, one could imagine a financial intermediary pooling claims sold by firms and selling a safe asset backed by the claims to workers.

Workers with history \( z^t \) purchase assets requiring payout \( a(z^t) \) next period. Hence, the financial assets of workers require a period \( t + 1 \) repayment equal to:

\[ E[a(z^t)] \]

Hence, market clearing for financial assets requires:

\[ E[d(s^{t+1})] = E[a(z^t)] \]

Because the entrepreneur’s shocks are idiosyncratic, a no-arbitrage condition requires that \( p(s^{t+1}) = \frac{1}{m_t} Q(s_t, s_{t+1}) \). Thus, the only interesting asset price in the model is the interest rate.

Preferences.

The preferences of entrepreneurs and workers are given by

\[ \sum_{t=0}^{\infty} \beta^t u(c_t) \]

with \( u' > 0 \) and \( u'' < 0 \), where \( c_t \) is contingent upon the history \( s^t \) for entrepreneurs and the history \( z^t \) for consumers.

The model emphasizes the creation and allocation of liquidity. Liquidity refers to instruments that enable an agent to transfer wealth to the times and occasions when it is needed. There are two forms of liquidity in the model: financial assets and physical capital. Workers can invest in financial assets, but they do not have access to a technology.
in which to productively invest physical capital. Entrepreneurs can invest in physical capital, making them natural suppliers of financial assets to the workers. Nonetheless, both entrepreneurs and workers need liquidity, to smooth consumption in the case of workers, and to take advantage of investment opportunities in the case of entrepreneurs.

### 1.3 Steady-state equilibrium characterization

In this environment, an equilibrium is a deterministic interest-rate and wage sequence $\{r_t, w_t\}_{t=0}^\infty$ and collections of plans for entrepreneurs $\{k(s^t), l(s^{t+1}), d(s^t)\}$ and for consumers $\{c(z^t), a(z^t)\}$ such that:

1. The plans $\{k(s^t), l(s^{t+1}), d(s^t)\}$ maximize the utility of each entrepreneur;
2. The plans $\{c(z^t), a(z^t)\}$ maximize the utility of households;
3. The bond market clears:
   \[
   E[d(s^{t+1})] = E[a(z^t)]
   \]
4. The labor market clears:
   \[
   E[l(s^t)] = E[z_t]
   \]

#### 1.3.1 The entrepreneur’s problem

In this section, I formulate the entrepreneur’s problem recursively and analyze the first-order conditions. I postpone until Section 4 a detailed discussion of conditions under which the value function is well defined and how the transition functions are specified.

To fix notation, let $\zeta$ be the distribution of entrepreneurs over liquid wealth $w$ and current productivity $s_-$. Denote by $\varphi$ the distribution of consumers over assets $a$ and current productivity $z$. Let $\mu = (\zeta, \varphi)$ be the distributions for both entrepreneurs and consumers. Write $R(\mu)$ for the market-clearing interest rate given distribution $\mu$.

The entrepreneur’s problem can be written recursively as:

\[
V(w, s_; \mu) = \max_{k, \{d_s\}} u(w + \frac{1}{R(\mu)} E[d_s|s_-] - k) + \beta E[V(f(k, s) - d_s, s; \mu')|s_-]
\]
subject to
\[ d_s \leq \theta f(k, s) \]

and the transition function for \( \mu \). Equivalently, the entrepreneur can choose \( w_s \) directly, allowing the problem to be written as:
\[
V(w, s_-; \mu) = \max_{k, \{w_s\}} u(w + \frac{1}{R(\mu)} E[f(k, s) - w_s|s_-] - k) + \beta E[V(w_s, s; \mu')|s_-] \tag{1.5}
\]
subject to
\[
w_s \geq (1 - \theta)f(k, s) \tag{1.6}
\]

and the transition function for \( \mu \).

Denote the Lagrange multiplier on the financial constraint (1.6) for state \( s \) by \( \phi_s \). We will say that the financial constraint for state \( s \) binds if \( \phi_s > 0 \).

The first-order condition for capital is:
\[
u'(c)(\frac{1}{R(\mu)}E[f_k(k, s)] - 1) - (1 - \theta)E[\phi_s f_k(k, s)] = 0 \tag{1.7}
\]

The first-order condition for capital immediately implies that if any of the financial constraints are binding, then capital \( k \) will be strictly less than the level of capital that maximizes expected output next period, discounted at the market interest rate, less investment. This latter capital level will be a useful benchmark throughout the paper:
\[
k^u(R, s_-) \equiv \max_k \frac{1}{R} E[f(k, s)|s_-] - k
\]

The first-order condition for wealth tomorrow in state \( s \) is:
\[-u'(c)\frac{1}{R} + \beta V(w, s; \mu') + \phi_s = 0 \tag{1.8}\]

and the envelope condition is:
\[
V_w(w, s_-; \mu) = u'(c) \tag{1.9}
\]

Consider any next-period idiosyncratic states \( s \) for which the financial constraint is not binding in the current period. Conditions (1.8) and (1.9) imply that next-period consumption will be equal across these states. Also, next-period consumption in these
states will be lower than this-period consumption if $\beta R < 1$.

Now consider any next-period idiosyncratic states $s$ for which the financial constraint is binding in the current period. Entrepreneurs would like to transfer wealth away from these next-period states, but the binding financial constraint prevents them from doing so. Next-period consumption in these states will be higher than in any states for which the constraint is not binding.

Define the rate of time preference $\rho$ by $\beta = (1 + \rho)^{-1}$. A net interest rate lower than the rate of time preference provides an incentive for the entrepreneur to have a downward-tilted consumption path. For a given investment $k$, however, the downward-tilting may not be possible; the entrepreneurs’ reneging constraint requires that she have a claim to at least $(1 - \theta)$ share of tomorrow’s output net of labor costs and undepreciated capital. This trade-off between downward-tilting the consumption path and investing enough to maximize expected discounted profits is central to the entrepreneur’s problem.

1.3.2 The “bite” of entrepreneurs’ financial constraints

Throughout the rest of the paper, whenever I refer to the marginal product of capital, I mean the expected marginal increase in output net of labor costs, $E[f_k(k(s^t), s_{t+1})|s^t]$. This differs from the realized marginal product of capital, $f_k(k(s^t), s_{t+1})$. When there are no financial constraints, it is $E[f_k(k(s^t), s_{t+1})|s^t]$ that will be equal to the interest rate and hence also equal across entrepreneurs. Even without financial constraints, the realized marginal product of capital will still differ across entrepreneurs if productivity is stochastic.

This paper emphasizes a positive analysis of the equilibrium impact of entrepreneurs’ borrowing constraints. In particular, the first-order condition for capital makes it clear that in this setting, there may be differences between the interest rate and an entrepreneur’s marginal product of capital, and between the marginal product of capital of different entrepreneurs.

A wedge between the marginal product of capital and the interest rate is a measure of the finance premium firms would be willing to pay for an additional unit of borrowing. One interesting statistic is the mean wedge across entrepreneurs. If the mean wedge is positive, then there is a kind of conditional under-investment; at the equilibrium interest rate, firms would like to borrow and invest more. It should be noted, however, that despite this conditional under-investment, aggregate capital may be greater than if there were no
frictions, since the financial frictions may depress the interest rate.

It is also interesting to study how the marginal product of capital differs across entrepreneurs. If the marginal product of capital differs across entrepreneurs, there is a loss of productive efficiency, in the sense that a social planner could achieve higher next-period output with same amount of aggregate capital by re-allocating current-period capital across entrepreneurs. One useful way to capture this loss of productive efficiency is through measured TFP. When technology is Cobb-Douglas, with \( F(k, l, s) = sk^a l^b \), it is possible to compute TFP in this economy for any allocation of capital to entrepreneurs \( k(w, s_-) \) and any distribution \( \zeta(w, s_-) \) over wealth and productivity.

**Lemma 1** Suppose technology is Cobb-Douglas. Then aggregate output is given by

\[
Y = Z K^a L^b
\]

where \( Z \) is measured TFP

\[
Z = \left[ \sum_{s_- \in S} E[s^{1-b}|s_-] \int (\frac{k(w, s_-)}{K}) \frac{a}{1-b} \zeta(w, s_-) dw \right]^{1-b} \tag{1.10}
\]

and \( K \) and \( L \) are aggregate capital and labor.

With Cobb-Douglas technology, measured TFP is a function of the shares of aggregate capital deployed by firms with given wealth and productivity. Maximizing TFP requires that the marginal product of capital, \( E[f_k(k, s)|s_-] \), be equal across firms. Hence, if the wealth distribution is not degenerate, maximizing measured TFP requires that capital not depend on liquid wealth \( w \).

**Definition 2** First-best TFP is defined as

\[
\overline{TFP} = \max_{x(w, s_-)} \left[ \sum_{s_- \in S} \int E[s^{1-b}|s_-] x(w, s_-)^a \zeta(w, s_-) dw \right]^{1-b}
\]

subject to

\[
\sum_{s_- \in S} \int x(w, s_-) \zeta(w, s_-) dw \leq 1.
\]
Denote \( \hat{q}(s) = \int \zeta(w, s)dw \). First-best TFP is given by

\[
TFP = \left[ \sum_{s_- \in S} \hat{q}(s_-) E[s^{-\alpha} \mid s_-] \left( \frac{E[s^{-\beta} \mid s_-]^{1-\alpha}}{\sum_{s_j \in S} E[s^{-\beta} \mid s_j]^{1-\alpha} \hat{q}(s_j)} \right)^{1-b} \right]^{\alpha - b} \quad (1.11)
\]

where \( \alpha = \frac{a}{1-b} \).

Note that it is possible to have a wedge between each entrepreneurs’ marginal product of capital and the interest rate and still have measured TFP equal to first-best TFP, so long as there is no dispersion in the marginal product of capital. Thus, this paper differentiates between the average wedge between the marginal product of capital and the interest rate, which can be considered a measure of conditional under-investment, and the dispersion of this wedge across entrepreneurs, which can be considered an indicator of productive inefficiency. This approach to studying TFP is similar to the analysis of TFP in Midrigan-Xu (2010) and Moll (2010).

### 1.3.3 Steady state definition

For a given interest rate and wage, the solution to the entrepreneur’s problem (1.5) is associated with optimal policies for capital and next-period state-contingent wealth, \( \{k^*(w, s_-), \{w^*_s(w, s_-) \}_{s \in S} \} \). Under certain conditions discussed below, the entrepreneur’s problem is well-defined and the optimal policies are continuous and single-valued. The single-valued optimal policies, together with the transition function \( Q \), define a new transition function \( P_E \). Given a distribution \( \zeta(w, s_-) \) over entrepreneurs’ wealth and productivity today, the transition function \( P_E \) determines the distribution tomorrow.

To be more precise, consider the measurable spaces \((W, W)\) and \((S, S)\). Define \( W = [0, \hat{w}] \) and \( W \) as the collection of all Borel sets that are subsets of \( W \). The definition of \( \hat{w} \) will be more precise below. Let \( (X, \mathcal{A}) = (W \times S, W \times S) \) be the product space. Recall that \( Q \) is the transition function on \((S, S)\). We can define the transition function \( P_E \) as

\[
P_E((w, s_-), A \times B) = \sum_{s \in B} Q(s_-, s) \mathbb{1}\{w^*_s(w, s_-) \in A\} \quad (1.12)
\]

for all \( w \in W, s_- \in S, A \in W, B \in S \).

For any probability measure \( \zeta \) on \((W \times S, W \times S)\), the operator mapping this probability
measure into next period’s probability measure is given by:

\[(T^\ast \zeta)(A) = \int P_E(x, A) \zeta(dx) \tag{1.13}\]

for all \(A \in \mathcal{W} \times \mathcal{S}\). An invariant distribution under \(P_E\) is a fixed point of (1.13).

One can analogously define a transition function \(P_C\) from the transition function for consumer productivity, \(Q_Z\), and the optimal policy of consumers, \(a^\ast(a, z)\), for given \(R\) and \(w\).

Now a steady-state equilibrium can be defined.

**Definition 3** A steady-state equilibrium consists of: (i) an equilibrium with \(R_t = R\) and \(\omega_t = \omega\) for all \(t\); (ii) a distribution \(\zeta\) over the wealth and productivity of entrepreneurs that is invariant with respect to \(P_E\); (iii) a distribution \(\varphi\) over wealth and productivity of consumers that is invariant with respect to \(P_C\).

### 1.4 Steady state comparisons

This section studies the steady-state properties of the economy, in partial and general equilibrium. The long-run effects of a decrease in firms’ ability to borrow are studied by comparing an initial steady-state with the steady-state that arises when firms’ moral hazard problem worsens.

In the remainder of this section, I focus on two special cases: (i) constant productivity, and (ii) productivity that is i.i.d. across time. For the case of constant productivity, the entrepreneurs’ steady-state marginal product of capital can be calculated in closed form for given \(R\) and \(\omega\) (Proposition 4). Thus, I can study how the wedge between the marginal product of capital and the interest rate varies with the ability to borrow \(\theta\) and the interest rate \(R\). Moreover, the general equilibrium steady-state can be partially characterized, allowing insight into the extent to which entrepreneurs escape financial frictions by accumulating liquid wealth and how this depends on workers’ demand for financial assets.

A key result is Proposition 7, which shows that, without consumer demand for financial assets, the general-equilibrium steady-state interest rate equals the rate of time preference, whereas with consumer demand for financial assets, the equilibrium interest rate must be less than the rate of time preference. It has long been recognized that an interest rate
less than entrepreneurs’ rate of time preference is important if financial frictions are to matter in the long run. Thus, some papers, such as Lorenzoni and Walentin (2007), assume a difference between the discount factors of entrepreneurs and consumers in order to achieve an interest rate less than entrepreneurs’ rate of time preference. In contrast, this paper assumes a common discount factor for entrepreneurs and consumers and arrives endogenously at such an interest rate. The key is consumers’ demand for liquidity, which lowers the interest rate.

Having an endogenous wedge between the steady-state interest rate and the entrepreneurs’ rate of time preference allows consideration of how shocks – such as a decrease in firms’ ability to borrow – affect the wedge. In contrast, in papers where the wedge is exogenous and comes from an ad-hoc assumption of entrepreneurs being less patient than other households, shocks such as a decrease in firms’ ability to borrow will not affect this wedge.

For the case of i.i.d. productivity, I will show that, in steady state, the average marginal product of capital is greater than the interest rate and the variance in the marginal product of capital is positive if and only if the interest rate is below the rate of time preference and moral hazard is sufficiently severe ($\theta < 1$).

1.4.1 Constant productivity

Productivity is constant if the productivity transition function $Q$ is the identity matrix. If $Q$ has more than one element, the economy features a non-degenerate distribution over entrepreneurial productivity, with the feature that each entrepreneur’s productivity is constant over time.

When productivity is constant, entrepreneurs’ long-run behavior can be characterized in closed form for any utility function and any production function that satisfy the basic technical assumptions described above. Also, though this proposition and some of the results that follow are partial-equilibrium results that take the interest rate as given, they can also be given a general-equilibrium interpretation, by assuming consumers are “CARA-normal.”

**Proposition 4** Suppose that productivity is constant over time. In steady state:

If consumers have constant absolute risk aversion (CARA) preferences and experience Gaussian shocks, consumers’ steady-state demand for assets is elastic at an interest rate $R$ that depends on consumers’ risk aversion and the variance of the shocks, as shown in Angeletos-Calvet (2006). With positive variance of the shocks, this interest rate satisfies $R\beta < 1$.

---

9If consumers have constant absolute risk aversion (CARA) preferences and experience Gaussian shocks, consumers’ steady-state demand for assets is elastic at an interest rate $R$ that depends on consumers’ risk aversion and the variance of the shocks, as shown in Angeletos-Calvet (2006). With positive variance of the shocks, this interest rate satisfies $R\beta < 1$.  

27
The marginal product of capital equals the harmonic mean of $R$ and $\frac{1}{\beta}$, where the weights are $\theta$ and $(1 - \theta)$, respectively. That is,

$$R \leq f_k(k, s) = \frac{1}{\theta \frac{1}{R} + (1 - \theta) \beta} \leq \frac{1}{\beta}$$

(1.14)

where the inequalities are strict and the financial constraint is binding if and only if $R\beta < 1$.

(ii) the supply of financial assets is decreasing in the interest rate.

According to this proposition, if the interest rate is less than the rate of time preference, then in steady-state, the financial constraint will bind for every entrepreneur.

This result builds on the observation that in steady state with constant-productivity entrepreneurs, each entrepreneur’s wealth and consumption must be constant over time. This intuitive observation is proved in Lemma 6 below. With this observation, Proposition 4 then follows directly from the first order conditions (1.7) and (1.8).

In particular, the first-order condition for wealth (1.8) implies that

$$\phi = u'(c) \frac{1}{R} - \beta u'(c_s)$$

(1.15)

Thus, in steady state, $\phi > 0$ and the financial constraint is binding if and only if $\beta R < 1$.

Now, suppose an entrepreneur increases capital at the margin, holding debt constant. The first-order condition for capital implies

$$-u'(c) + \beta u'(c_s)f_k(k, s) + \phi \theta f_k(k, s) = 0$$

(1.16)

Increasing capital at the margin results in a decrease in consumption today, with cost $u'(c)$, and an increase in consumption tomorrow, with benefit $\beta u'(c_s)f_k(k, s)$. Increasing capital also serves to loosen the financial constraint, with benefit $\phi \theta f_k(k, s)$. Combining (1.15) and (1.16) with the observation that consumption is constant over time in steady state, we obtain (1.14).

This result can also be understood by thinking about the dynamic accumulation of wealth. As will be shown in the proof for Lemma 6, if the interest rate is less than the rate of time preference, entrepreneurs with capital below [above] the level of capital defined in (1.14) will have higher [lower] next-period wealth and capital.

The expression for the marginal product of capital in (1.14) can be thought of as a user
cost of capital, as in Jorgenson (1963) and Rampini and Viswanathan (2010b). Because output net of labor costs and undepreciated capital are not fully pledgeable to investors, entrepreneurs need to hold liquid wealth in order to take advantage of their investment opportunity each period, and if the interest rate is less than the rate of time preference, holding liquid wealth is expensive.

The closed-form solution for steady-state marginal product of capital allows a clear understanding of how investment distortions depend on the equilibrium interest rate, as described in the following corollary.

**Corollary 5** Suppose productivity is constant. In steady state,

(i) There is a wedge between the marginal product of capital and the interest rate if and only if \( \theta < 1 \) and \( \beta R < 1 \).

(ii) If \( \theta < 1 \), the ratio of the marginal product of capital to the interest rate is decreasing in \( R \).

(iii) If \( \beta R < 1 \), the ratio of marginal product of capital to the interest rate is decreasing in \( \theta \).

(iv) There is no dispersion in the marginal product of capital.

Part (i) of the corollary shows that in the case of constant productivity, for there to be a positive average finance premium in the long-run, it is required both that moral hazard is sufficiently severe (\( \theta < 1 \)) and that the interest rate is below the rate of time preference. Part (ii) and (iii) of the corollary can be interpreted in terms of the user cost of capital. A lower interest rate implies that it is more expensive to hold liquid wealth; thus, entrepreneurs are willing to bear greater investment distortions to economize on the costs of holding liquid wealth when the interest rate is lower. Likewise, greater moral hazard, or a lower value for \( \theta \), implies that more liquid wealth is needed to undertake a given level of investment; if holding liquid wealth is expensive, greater moral hazard increases the user cost of capital.

With constant productivity, there is no dispersion in the marginal product of capital. Hence, if the production technology is Cobb-Douglas, there are no endogenous losses in measured TFP. The absence of capital misallocation and the irrelevance of the interest rate for steady-state TFP is driven by two special assumptions: (i) unchanging productivity; and (ii) the equality across entrepreneurs of the moral hazard parameter or, equivalently, the ability to pledge future income and assets. Dropping either of these assumptions will
give rise to capital misallocation and steady-state TFP losses if the interest rate is less than the rate of time preference, as shown in Proposition 8 (for the case of heterogeneity in entrepreneurs’ ability to pledge future income and assets) and Corollary 12 (for the case of productivity shocks). Moreover, with heterogeneity in \( \theta \), Proposition 8 shows that TFP is increasing in the interest rate, if the interest rate is less than the rate of time preference. In the calibration, which features productivity shocks, TFP is also increasing in the interest rate.

In the case of constant productivity, it is possible to fully characterize the invariant distributions for a given interest rate and show how the uniqueness or multiplicity of the invariant distribution for a given interest rate depend on whether the interest rate is equal to the rate of time preference. This is a helpful step in characterizing the supply of financial assets and moving toward the general equilibrium results below.

**Lemma 6** Suppose entrepreneurs’ productivity is constant.

(i) For \( R < \frac{1}{\beta} \), there is a unique invariant distribution \( \zeta(w, s, \cdot) \). In the invariant distribution, the wealth of entrepreneurs with productivity \( s \) is

\[
w = (1 - \theta) f\left( k_k(\cdot, s) \right)^{-1} \left( \frac{1}{\beta R (1 - \theta)} \right), s
\]

(ii) For \( R = \frac{1}{\beta} \), a distribution over wealth and productivity is an invariant distribution if and only if the measure of entrepreneurs with wealth greater than or equal to \((1 - \theta) f(k^w(R), s)\) is equal to one.

(iii) In any invariant distribution, each entrepreneur’s wealth is constant over time.

As shown above, if the interest rate is less than the rate of time preference, all entrepreneurs will be constrained and have positive capital and hence the net supply of financial assets by entrepreneurs will be positive. This leads to the next proposition, which illustrates the importance of considering workers’ demand for financial assets when studying how a decrease in firms’ ability to borrow affects workers and firms. I contrast general equilibrium in the model with the equilibrium that would obtain if workers were hand-to-mouth consumers. Define workers as hand-to-mouth consumers if each period they simply consume their labor income, because they are not allowed to or do not wish to save.

**Proposition 7** (i) Suppose workers are hand-to-mouth consumers. In general-equilibrium steady state, \( \beta R = 1 \) and there is no wedge between the marginal product of capital and the interest rate.
(ii) Suppose consumers can save. Then $\beta R < 1$ and there is a positive wedge between marginal product of capital and the interest rate in general-equilibrium steady-state.

The intuition for this result is straightforward. With hand-to-mouth workers, financial market equilibrium requires that net supply of financial assets by firms be zero. But we have already seen that $\beta R < 1$ implies that all firms are constrained and hence the net supply of financial assets is positive. Moreover, there cannot be a general-equilibrium steady-state with $\beta R > 1$, since firms would have perpetually increasing wealth. In contrast, with $\beta R = 1$, there are multiple invariant distributions consistent with zero net supply of financial assets by entrepreneurs. In particular, any distribution of wealth satisfying the lower-bound on wealth in part (ii) of the previous lemma and such that aggregate entrepreneurial wealth equals aggregate unconstrained entrepreneurial output is an invariant distribution with zero net supply of financial assets. Hence, in general-equilibrium steady state, $\beta R = 1$ and there is no wedge between the marginal product of capital and the interest rate.

In contrast, when workers can save, $\beta R = 1$ can no longer be an equilibrium; a well-known result about the income fluctuations problem is that, when $\beta R = 1$, workers’ financial assets will converge to infinity almost surely (e.g., Aiyagari 1994). Hence, general-equilibrium steady state will feature $\beta R < 1$ and, by the previous corollary, a wedge between the marginal product of capital and the interest rate.

Importantly, to shift the general equilibrium from $\beta R = 1$ to $\beta R < 1$, it is not necessary that workers’ demand for financial assets be such that savers’ assets converge to infinity when $\beta R = 1$. All that is required is a sufficiently large worker demand for financial assets when the interest rate equals the rate of time preference; the threshold is given by

$$\theta \sum_{s^- \in S} f(k^u(\frac{1}{\beta}, s^-), s^-) \hat{q}(s^-)$$

where $\hat{q}$ is the distribution over entrepreneurial productivity.

An implication of this proposition is that, with hand-to-mouth workers, a decrease in the ability of firms to borrow has no long-run effect on the interest rate or on the wedge between the marginal product of capital and the interest rate. Allowing workers to save opens the door to the predictions of the model that a decrease in firms’ ability to borrow leads to increased distortions in firms’ investment and a decreased interest rate.

Nonetheless, at this point, we cannot answer theoretically the question of how a decrease in firms’ ability to borrow affects firms and workers in general equilibrium. Although we have characterized firms’ supply of financial assets in closed form for any production technology and utility function satisfying the basic technical assumptions, much less can be
said about the workers’ problem at this level of generality. This is because of the well-known theoretical difficulties of the income fluctuations problem. Indeed, workers’ steady-state demand for financial assets may not be everywhere increasing in the interest rate, especially since the equilibrium wage is decreasing in the interest rate. Hence, a further exploration of general equilibrium response of the economy to a decrease in firms’ ability to borrow will be postponed until the calibrated example.\textsuperscript{10}

\textbf{Heterogeneity in the ability to pledge future cash flows}

Throughout the paper, I abstract from heterogeneity in entrepreneurs’ ability to pledge future cash flows. In particular, I assume that the moral hazard parameter $\theta$ is equal for all entrepreneurs. However, it is worth considering what happens when there is cross-sectional heterogeneity in the ability to pledge future cash flows. One reason that $\theta$ may vary across entrepreneurs is differences in asset tangibility, a topic which has recently received considerable attention in macroeconomics and finance (McGrattan and Prescott 2010, Rampini and Viswanathan 2010b).

Hence, in this section, I allow the moral hazard parameter $\theta$ to differ across entrepreneurs. Denote the set of possible values for $\theta$ by $\Theta \subset [0, 1]$, where $\Theta$ is a discrete set. Define $g(\theta, s) : \Theta \times S \rightarrow [0, 1]$ as the joint distribution function for the pledgeability parameter $\theta$ and productivity $s$.

Given the possibility of heterogeneity in entrepreneurs’ ability to pledge future income and assets, one might want to be flexible in how ability to pledge is related to productivity. Hence, I allow any joint distribution function $g(\theta, s)$.

When the distribution of $\theta$ is non-degenerate (i.e., there is positive cross-sectional variance in $\theta$), there will be capital misallocation if the interest rate is less than the rate of

\textsuperscript{10}However, if we make assumptions not about the primitive parameters of the workers’ problem but about the steady-state demand for financial assets (an equilibrium object), then more can be said about a close variant to this economy. Consider this alternate economy: the entrepreneurs’ technology does not depend on labor input and the workers’ wage is $\omega$ irrespective of the interest rate. The later would be the case if workers were themselves self-employed with productivity $\omega$ or if firms with a technology $\omega L$, where $L$ is the labor input, are added to the model. If workers’ steady-state demand for financial assets, $\int a^*(a, y; \omega, \omega) \varphi(a, y) \, da \, dy$, is increasing in the interest rate, then: (i) the wedge between entrepreneurs’ marginal product of capital and the interest rate is decreasing in $\theta$; and (ii) the interest rate and workers’ holdings of financial assets are increasing in $\theta$.

Note that if the workers have constant relative risk-aversion (CRRA) utility and their borrowing constraint is a multiple of the wage, as with the natural borrowing constraint, then workers’ steady-state demand for financial assets $\int a^*(a, y; \omega) \varphi(a, y) \, da$ is homogenous of degree one in $\omega$. Hence if $\int a^*(a, y; \omega, R) \varphi(a, y) \, da$ is increasing in the interest rate for any $\omega$, it is increasing in the interest rate for all $\omega$. 32
time preference, even though entrepreneurs’ productivity is constant. This is because, all
else equal, firms with a higher ability to pledge will invest more than firms with a lower
ability to pledge. A decrease in the interest rate biases the composition of output toward
entrepreneurs with a high ability to pledge future income and assets. Moreover, with Cobb-
Douglas technology, measured TFP is increasing in the interest rate. This is formalized in
the following proposition.

**Proposition 8** Suppose there is positive cross-sectional variance in the pledgeability pa-
rameter $\theta$. In steady state, for any distribution function $g(\theta, s)$ over pledgeability and pro-
ductivity:

(i) There is dispersion in the marginal product of capital if and only if $\beta R < 1$.

(ii) If technology is Cobb-Douglas, TFP is increasing in the interest rate for $\beta R < 1$.

The compositional effect underlying this proposition suggests potential ways to evaluate
this theory using panel data on firms’ financing and investment decisions.

The proposition also shows that the model is sufficiently tractable not only to admit a
closed-form solution for steady-state TFP, but also to permit comparative statics with re-
spect to the interest rate, even though the production side here is composed of a distribution
of entrepreneurs with financial constraints and with two dimensions of heterogeneity.

In the calibration, I abstract from heterogeneity in firms’ ability to pledge or asset
tangibility. In future work, it would be interesting to compare the share of TFP losses due
to capital misallocation accounted for by: (i) productivity shocks; and (ii) heterogeneity in
firms’ ability to pledge.

### 1.4.2 Productivity independent across time

In the previous section, it was shown that when workers can save, the general-equilibrium
steady state features a wedge between the marginal product of capital and the interest
rate. Moreover, with heterogeneity in firms’ ability to pledge future cash flows, there is
capital misallocation in steady state and total factor productivity is less than first best.
This section introduces productivity shocks for entrepreneurs. With productivity shocks
for entrepreneurs, the economy will feature capital misallocation in steady state even if all
entrepreneurs have the same ability to pledge future cash flows. Incorporating productivity
shocks for entrepreneurs allows the model to make richer predictions about how a decrease in firms’ ability to borrow affects capital misallocation, risk management and TFP.

**Preliminaries**

The property that productivity is independent across time is equivalent to $Q(s_-, s) = Q(s'_-, s)$ for all $s_-, s'_- \in S$. When productivity is independent across time, I will restrict $W$ to $W = [0, \hat{w}]$. For $\theta < 1$, I take $\hat{w} = \bar{w} \equiv (1 - \theta)f(k_u(R, .), \bar{\pi})$. For $\theta = 1$, any finite $\hat{w} > 0$ can be used. The restrictions on $W$ will be discussed throughout the text and the proofs.

**Proposition 9** Suppose productivity is independent across time. Then:

(i) The optimal policies $k^*(w)$ and $w^*(w)$ are continuous, single-valued functions;

(ii) $V$ is strictly increasing and strictly concave in $w$ and continuously differentiable at any $w \in \text{int}(W)$.

This proposition also applies to the case of constant productivity; to see this, for any productivity $s$, take $S = s$ and $Q = 1$.

For the next proposition, I will need a technical assumption. When productivity is independent across time, there is a possibility of zero productivity if

$$0 \in S \text{ and } Q(., 0) > 0$$

**Proposition 10** Suppose productivity is independent across time. Then, given $R$, there is a distribution over wealth that is invariant under $P_E$. Suppose further that there is a possibility of zero productivity. Then $P_E$ has a unique invariant distribution $\zeta^*$.

If $\beta R < 1$, uniqueness of the invariant distribution and weak convergence apply for any upper-bound on $W$ that is greater than or equal to $(1 - \theta)f(k_u(R, .), \bar{\pi})$. If $\beta R = 1$ and if the domain of $W$ is enlarged by taking $\hat{w} > \bar{w}$, then a distribution composed of a convex combination of $\zeta^*$ and any distribution over $[\bar{w}, \hat{w}]$ will also be an invariant distribution, and hence there will not be a unique invariant distribution. This is similar to the case of constant productivity, which also featured a unique invariant distribution when the interest is less than the rate of time preference, but featured a multiplicity of invariant distributions when the interest rate is equal to the rate of time preference. This multiplicity will not
affect the results below, in the sense that the results hold for any upper-bound on wealth greater than or equal to $\bar{w}$.

**Steady-state analysis**

The following partial-equilibrium result highlights that an interest rate below the rate of time preference leads entrepreneurs to have less liquid wealth than they would need to avoid financial constraints.

**Proposition 11** In steady state, the measure of firms with at least one binding financial constraint is strictly positive if and only if $\beta R < 1$.

Suppose, by contradiction, that $\beta R = 1$ and the measure of firms with a binding financial constraint is positive under some distribution $\zeta$. Then, by the first-order condition for wealth (1.6), for any next-period states for which the financial constraint is binding, next-period consumption will be greater than current consumption. For any next-period states for which the financial constraint is not binding, next-period consumption will be the same as current consumption. Hence, next-period aggregate consumption would be greater than current-period consumption. Since the consumption policy is a function of wealth and productivity, this must mean that current distribution $\zeta$ is not the invariant distribution $\zeta^*$. The intuition for the $\beta R < 1$ case is similar.

Studying wealth dynamics for a given interest rate is useful in understanding how steady-state properties vary with the interest rate. If $\beta R = 1$, any entrepreneur with $w < \bar{w}$ will have next-period wealth $w_s$, with $w \leq w_s \leq \bar{w}$; if $\theta < 1$, the first inequality is strict for $\bar{s}$. Entrepreneurs will eventually accumulate exactly enough wealth to overcome their financial constraints; for any initial distribution of wealth over $[0, \bar{w}]$, the steady-state wealth distribution is degenerate: $\Pr(w = \bar{w}) = 1$. Once an entrepreneur has wealth $(1 - \theta)f(k^u(R, .), \bar{\pi})$, the entrepreneur will thereafter have constant wealth and consumption and invest as if she were unconstrained.

In contrast, if $\beta R < 1$, an entrepreneur with wealth $\bar{w} = (1 - \theta)f(k^u(R, .), \bar{\pi})$ will have strictly lower next-period wealth.\footnote{To see this, suppose $\beta R < 1$ and consider an entrepreneur with wealth $w = (1 - \theta)f(k^u(R, .), \bar{\pi})$. Suppose that the entrepreneur’s financial constraint for state $s$ is not binding. Then wealth in state $s$ will be strictly less than $w$, since the first-order condition for wealth implies $V_\omega(w_s) = \frac{1}{\beta R} V_\omega(w)$, and $V$ is concave. Suppose instead that the entrepreneur’s financial constraint for state $s$ is binding. Then capital $k$ must be less than unconstrained capital, $k < k^u(R, .)$. Since the constraint is binding, wealth in state $s$ must be given by $(1 - \theta)f(k(R, .), s) < (1 - \theta)f(k^u(R, .), \bar{\pi}) = w$.} This de-cumulation of wealth fostered by an interest
rate less than the rate of time preference leads to binding financial constraints.

The next proposition shows how moral hazard and an interest rate below the rate of
time preference can together result in certain distortions in steady state.

**Proposition 12** Suppose that $S$ has more than one element and consider the steady state.

If $\beta R < 1$ and $\theta < 1$, then and only then the following hold:

(i) the average marginal product of capital is greater than the interest rate;

(ii) the variance in the marginal product of capital is positive;

(iii) entrepreneurs’ consumption is risky;

**Corollary 13** When technology is Cobb-Douglas, aggregate TFP is less than first-best, $TFP < \frac{Y}{K}$, if and only if $\beta R < 1$ and $\theta < 1$.

This proposition allows an analysis of a specific worsening in entrepreneurs’ moral haz-
ard, holding the interest rate constant. That is, the steady-state behavior of entrepreneurs
for $\theta = 1$ and $\theta < 1$ can be compared for a given interest rate. Although this is a very
coarse change, it provides some insight nonetheless. If the interest rate equals the rate
of time preference, then a worsening of moral hazard has no long-run effect on investment
distortions. That is, if $\beta R = 1$, there will be no finance premium, no dispersion in the
marginal product of capital, and no endogenous TFP losses, regardless of whether $\theta$ is less
than one or equal to one.

In contrast, if the interest rate is less than the rate of time preference, a decrease in $\theta$
from $\theta = 1$ to $\theta < 1$ will result in a positive average finance premium, positive variance of
marginal product, and endogenous TFP losses.

Likewise, we can consider how the long-run properties of the economy depend on the
equilibrium interest rate. With $\theta < 1$, there are investment distortions if and only if $R < \frac{1}{\beta}$.

In the numerical analysis below, we will see that the direction of these comparative stat-
ics (“lower theta, higher steady-state investment distortions,” “lower interest rate, higher
steady-state investment distortions”) hold when the space for $\theta$ and the interest rate are
not partitioned as coarsely as here.$^{12}$

An analysis of the well-studied problem of the workers shows that it is not possible to
have a steady-state equilibrium with $\beta R = 1$, because at this interest rate, the workers’

---

$^{12}$Note that when capital misallocation is a result of heterogeneity in entrepreneurs’ ability to pledge future
cash flow, we can show analytically that steady-state TFP is increasing in the interest rate (Proposition 8).
asset demand is not finite, as emphasized by Aiyagari (1994), Chamberlain and Wilson (2000), and others.

**Lemma 14 (Chamberlain-Wilson (2000), Sargent-Ljungqvist (2000))** Suppose that $\beta R = 1$ and that workers’ productivity is independent across time. Then $\lim_{t \to \infty} a(z^t) = \infty$ almost surely.

The reason why entrepreneurs’ asset position is finite when $\beta R = 1$, whereas consumers’ asset position converges to infinity, is due to entrepreneurs’ access to state-contingent promises. In contrast, consumers trade only a safe asset.

Thus, it is not possible to have a steady state equilibrium with $\theta < 1$ and without binding financial constraints.

**The individual entrepreneur**

This section studies further the problem of the individual entrepreneur in partial equilibrium. The results here provide some further intuition for understanding steady-state properties like endogenous TFP losses.

In the model, entrepreneurs invest less than or equal to $k^u(R, s)$, the amount that would be invested by an entrepreneur who faced no moral hazard. The next proposition shows that, for wealth levels such that capital is less than $k^u(R, s)$, capital is increasing in liquid wealth.

**Lemma 15** If any of the financial constraints bind, $k^*(w, s_-)$ is strictly increasing in $w$. Optimal consumption $c^*(w, s_-)$ and state-contingent next-period wealth $\{w^*_s(w, s_-)\}_{s \in S}$ are strictly increasing in $w$. $c^*$ increases with $w$ at a rate strictly less than one-for-one.

The next lemma further highlights the link between liquid wealth and whether financial constraints bind. The lemma requires defining $w_{l,s_-} = -\max_k -k + \frac{\theta}{1+r}E[f(k, s)]$. When productivity is independent across time, this value does not depend on $s_-$. This lemma is similar to a result in Rampini-Viswanathan (2010a).

**Lemma 16** For every $s_-$, there is a thresholds $w_{s_-} > w_{l,s_-}$ such that, for $w < w_{s_-}$, the financial constraints are binding tomorrow in all states $s$. 37
1.5 Calibration and Numerical Results

In this section, U.S. data on entrepreneurs’ revenues, capital and labor inputs, and financing are used to calibrate the model. I use micro data on U.S. entrepreneurs to find reasonable parameters for the production technology and productivity process of U.S. entrepreneurs. The workers’ income process is calibrated using standard parameters from Heaton and Lucas (1996). I then choose the moral-hazard parameter, $\theta$, and the discount factor $\beta$ to match an aggregate leverage measure for U.S. entrepreneurs and a target interest rate.

Using these parameters, I study the steady-state properties of the economy, such as TFP losses from misallocation. The central calibration exercise is to study the long-run effects of an increase in moral hazard, which corresponds to a decrease in $\theta$. I do so by comparing the steady-states for the calibrated value $\theta_1$ and a decreased value $\theta_2 < \theta_1$. I choose the size of the decrease in $\theta$ to match the decline in long-run real interest rates associated with the recent financial crisis.

The purpose of the exercise is to study further the properties of the model and the mechanisms discussed above, and to give a sense of the potential importance of the amplification and spillover effects previously identified.

The entrepreneurial sector in the United States is large. According to data from the 2007 Survey of Consumer Finances (SCF), 15 million U.S. households have a family member with an “active management role” in a privately-held business which the family owns in whole or in part. Based on the SCF data, these entrepreneurial firms had total sales in 2006 equal to more than 13 trillion dollars; total sales of all businesses in the US in 2006 were 28 trillion dollars. The total number of employees of these entrepreneurial firms was 103 million; according to the Current Population Survey (CPS), the number of employed U.S. civilians in 2006 was 144 million.\footnote{This estimate of the number of employees of entrepreneurial firms is biased downward because the number of employees for a given firm is top-coded at 5,000.} Using SCF data, the reported value of entrepreneurs’ stakes in their businesses was 9.7 trillion dollars; for comparison, according to Federal Reserve Flow of Funds data, the value of corporate equities held by U.S. households and non-profits was 9.6 trillion dollars.

In order to obtain parameters for the productivity process and production technology, I use data on U.S. entrepreneurs from the Kauffman Firm Survey (KFS), a longitudinal dataset of entrepreneurial firms. Virtually all macroeconomic studies of U.S. entrepreneurs
rely on the Survey of Consumer Finances, the Panel Study of Income Dynamics, the Consumer Expenditure Survey or the National Longitudinal Survey of Youth. The estimation approach I use could not be implemented with these data sets, because they lack information on the capital and labor inputs used by the firms. Moreover, among these data sets, the SCF has the best coverage of entrepreneurs, because it oversamples the wealthy, but the Survey of Consumer Finances (SCF) has only a very limited panel component, comprising the 1983 and 1989 surveys. Thus, any attempt to obtain parameters for the productivity process and production technology from the Survey of Consumer Finances (SCF) has to rely either on a two-period panel or on identification from cross-sectional observations. In contrast, I can make use of the longitudinal aspect of the KFS. The data sources are described in more detail in Appendix C.

For entrepreneurs, I assume a Cobb-Douglas production technology:

\[ F(k, l, s) = sk^a l^b \]

Here \( a + b \) is a measure of the returns to scale for entrepreneurs’ production technology, which can be interpreted as entrepreneurs’ “span of control” (Lucas 1978). Equally of interest is the elasticity with respect to capital of the entrepreneurs’ output net of labor cost function \( f \). This elasticity is given by \( \alpha = \frac{a}{1-b} \).

I estimate \( a \) and \( b \) using a procedure similar to Olley-Pakes (1996); the details are given in Appendix C. This procedure also produces estimates of idiosyncratic firm-level productivity. Using the estimated productivities, I perform the following regression

\[ \log s_{it} = \rho \log s_{i,t-1} + \varepsilon_{it} \]  

(1.17)

to obtain the process for productivity.

This approach to finding parameters for the entrepreneurs’ production technology and productivity process is similar to the one used by Moll (2010) to study the impact of financial frictions for Chilean and Colombian manufacturing firms.

Table 1 provides the results of this procedure. The elasticity with respect to labor, 0.841, is higher than the typical value of two-thirds corresponding to the historical aggregate U.S. labor share of total income. Correspondingly, the elasticity with respect to capital is lower than the typical value of one-third. This is likely related to entrepreneurial nature of these
Table 1. Parameters of the entrepreneurs’ productivity process and technology

| Parameter | Description                                         | Value   |
|-----------|-----------------------------------------------------|---------|
| $a$       | Elasticity of output with respect to capital        | 0.130   |
| $b$       | Elasticity of output with respect to labor          | 0.841   |
| $\sigma_\varepsilon$ | Standard deviation of productivity innovations | 0.110   |
| $\rho$    | Autocorrelation of productivity                     | 0.538   |

Source: Estimated using micro-data for U.S. entrepreneurs. See text or Appendix C.

firms; for example, there are virtually no mining firms or utilities in the KFS sample, and these industries tend to be very capital intensive. Many development-focused papers on entrepreneurship, including Moll (2010), focus exclusively on manufacturing industries and, not surprisingly, the elasticities with respect to capital that they estimate are quite high.

The span of control value, $a + b$, is 0.97, close to constant returns to scale. The elasticity with respect to capital of the output net of labor costs function $f$ is 0.82. This is slightly less than the elasticity calibrated for U.S. entrepreneurs by Cagetti-DiNardi (2006), 0.88, but much higher than the elasticity calibrated by Buera-Shin (2010a), 0.56. In Buera-Shin (2010a), identification comes mostly from matching the employment share of the top tenth of establishments and the earnings share of the top twentieth of the population.

For the workers’ labor income process, I use standard values estimated by Heaton and Lucas (1996), who estimate households’ income processes from the PSID. Heaton and Lucas (1996) assume that labor productivity is log-normally distributed and find that the average autocorrelation is 0.529 and that the average standard deviation of innovations to log productivity is 0.251. For the workers’ borrowing limit, I take the natural borrowing limit, which is zero, given the assumption of log-normal productivity.

I assume that the utility function $u$ takes the form

$$u(c) = \begin{cases} 
\frac{c^{1-\gamma}}{1-\gamma} & \text{if } \gamma \neq 1 \\
\log(c) & \text{if } \gamma = 1 
\end{cases}$$

where $\gamma$ is the coefficient of relative risk aversion. In the primary specification, I assume the coefficient of relative risk aversion is equal to one, a conservative assumption that also facilitates computation.

This leaves two parameters to be calibrated, $\beta$ and $\theta$. To pin down these parameters, I seek to match: the aggregate ratio of debt and equity to sales for entrepreneurial firms; and

---

14 Average here refers to the average across households.
a target interest rate. The aggregate ratio of debt and equity to sales is 0.3547. Data for
the aggregate value of debt and sales of U.S. entrepreneurial firms comes from the Internal
Revenue Service Statistics of Income. I use the SCF to estimate the aggregate value of
equity in U.S. entrepreneurial firms owned by people without an active management role
in the firms. This notion of equity is consistent with the model’s definition of financial
assets as claims on entrepreneurial firms held by outside investors. I include equity, rather
than only debt, because the model features state-contingent claims and I seek to match the
total value of investors’ claims on entrepreneurial firms.

I choose a target real yearly interest rate of 2.5 percentage points. Because $E[d]$ in my
model refers to the face-value of financial claims on the entrepreneurial firm plus interest
payment, I seek to match $E[d] = 0.3547(1 + r) = 0.3634$. I am able to match the leverage
ratio and target interest rate almost exactly, as shown in Table 2.

| Leverage ratio, $\frac{E[\text{deb}+\text{equity}]}{E[\text{sales}]}$ | Target | Model |
|-----------------------------|--------|-------|
|                             | 0.363  | 0.364 |
| Real interest rate          | 0.025  | 0.025 |

The calibrated discount rate is $\beta = 0.94$. This corresponds to a difference between the
inverse discount rate and the interest rate, $\frac{1}{\beta} - R$, equal to about four percentage points.
This is less than the difference found by Buera and Shin (2010a) for US data with a model
where the only friction is the lack of state-contingent debt, but it is far more than the
differences found in Aiyagari (1994), which featured no firm-side frictions and a constant
returns to scale production technology. The difference relative to Aiyagari (1994) will be
revisited at the end of the next section.

The calibrated moral hazard parameter is $\theta_1 = 0.62$. This corresponds to firms’ being
able to pledge to investors about two-thirds of their next-period output net of labor costs
and undepreciated capital.

Although no information about the size distribution of firms is used to calibrate the
model, the calibrated model captures the concentration of production in the largest firms.
For example, in US data, the share of total employment accounted for by the top decile

---

15Equity in the entrepreneurial firms I study can, in general, only be held by people; the equity cannot be
held by corporations, for example. See Appendix C for details.
(measured by employment) of establishments is 67 percent, according to Buera and Shin (2010a); in the calibration, it is 74 percent. We can also compare the size distribution in the model to the size distribution of entrepreneurial firms in the SCF. In the 2007 SCF, the top decile (measured by employment) of entrepreneurial firms accounted for 71 percent of total entrepreneurial employment.

Moreover, because of a homotheticity property of the entrepreneurs’ and workers’ problems, the average firm productivity here is a normalization; this is made precise in the following lemma. Thus, we can set average firm productivity to match the empirical capital-labor ratio.

**Lemma 17** Suppose there is a steady-state equilibrium with interest rate \( R \) and wage \( \omega \) and distributions over entrepreneurs \( \zeta(w, s_-) \) and over workers \( \varphi(a, z) \). Suppose technology increases by a factor of \( x \). Then there is a steady-state equilibrium with interest rate \( R \) and wage \( x^{\frac{1}{1-a}} \) and distributions over entrepreneurs \( \zeta'(w, s_-) \) and workers \( \varphi'(a, z) \), where

\[
\zeta'(w, s_-) = \zeta(x^{-\frac{1}{1-a}} w, x^{-1} s_-)
\]

and

\[
\varphi'(a, z) = \varphi(x^{-\frac{1}{1-a}} a, z)
\]

The cross-sectional distribution of the marginal product of capital is unchanged. TFP losses as a share of first-best TFP – the difference between first-best TFP and actual TFP, divided by first-best TFP – are unchanged.

Similarly, the ratio of firms to workers is also a normalization. Thus, we can set this ratio in the calibration to match average employment per firm.

### 1.5.1 Main numerical exercise: Decrease in firms’ ability to borrow

This section compares the steady-state properties of the benchmark economy to the steady-state properties of an economy where firms’ ability to borrow is lower than in the benchmark economy. That is, I compare the steady-states of two economies that have different levels of moral hazard but otherwise have identical parameters. The goal is to understand the persistent effects of a decrease in firms’ ability to borrow.
In the benchmark economy, the pledgeability parameter is $\theta_1 = 0.62$. I choose the new value of $\theta$ such that the endogenous decrease in the steady-state interest rate is consistent with the recent decline in long-term real interest rates associated with the financial crisis. This is simply for a point of reference, to anchor the change in $\theta$ in a way that can be readily understood.\(^1\)

The new value of $\theta$ is $\theta_2 = 0.49$, reflecting a 13 percentage point decrease in firm’s ability to borrow. This lower value for $\theta$ corresponds to greater moral hazard and lower ability to commit to repayment.

Table 3 compares several steady-state properties of the two economies. The first column lists the steady-state properties of the economy at the calibrated value for $\theta$. The second column ("$\theta = \theta_2$: Partial Equilibrium") lists the steady-state properties if $\theta$ decreases from $\theta_1$ to $\theta_2$, but the interest rate remains unchanged. This is the outcome that would obtain in a small open economy with a world interest rate such that, before the decrease in firms’ ability to borrow, the economy had zero current account surplus. This scenario is an analytical tool for thinking about the general equilibrium effects of a decrease in firms’ ability to borrow. The third column lists the new steady-state properties, after the decrease in $\theta$ from $\theta_1$ to $\theta_2$.

The first row shows losses in measured total factor productivity (TFP). If investment were unconstrained, the economy would obtain first-best TFP, as defined in (1.11). At the calibrated value for $\theta$, measured TFP is 2.45 percentage points less than first-best TFP. The second row shows the average wedge between the marginal product of capital and the interest rate. This is a measure of average finance premium firms would be willing to pay for an additional unit of borrowing. At the calibrated value for $\theta$, the average wedge is 1.62 percentage points. The third row shows the standard deviation of the marginal product of capital. Like measured TFP losses, this is a common measure of misallocation. At the calibrated value for $\theta$, the standard deviation of this wedge is 2.55 percentage points.

Now consider a decrease in firms’ ability to borrow. If the interest rate were held constant, TFP losses would increase by 0.33 percentage points (column "$\theta = \theta_2$: Partial Equilibrium"). However, in order to obtain steady-state equilibrium, the interest rate

---

\(^1\)For the real long-term interest rate, I use the rate on 20 year US Treasury Inflation Protected Securities (TIPS). On May 31, 2007, before the major summertime subprime downgrades, this interest rate was 2.53 percent. On October 4, 2010, as the Troubled Asset Relief Program (TARP) expired, this rate was 1.49 percent. The ten year real interest rate showed a greater decline, from 2.54 to 0.49. Which maturity one should use for this exercise is not obvious, as the paper’s model does not include term premia.
Table 3. Decrease in firms’ ability to borrow: Steady-state distortions

|                         | $\theta = \theta_1$ | $\theta = \theta_2$ | Increase |
|-------------------------|----------------------|----------------------|----------|
|                         | Partial equilibrium  | General equilibrium  |          |
| TFP losses              | 2.45                 | 2.78                 | 3.17     | 0.72     |
| E[marginal product of capital] - r | 1.62                 | 2.24                 | 2.74     | 1.12     |
| Sd[marginal product of capital] | 2.55                 | 3.18                 | 3.51     | 0.96     |
| Workers’ consumption risk, $(1 - E x)$ | 2.47                 |                      | 2.82     | 0.35     |

All values in percentage points. See text for details. The final column gives the difference between the steady-state value for $\theta = \theta_2$ and the steady-state value for $\theta = \theta_1$.

must fall, because of the decreased supply of financial assets. Hence, the actual increase in TFP losses is 0.72 percentage points (column “$\theta = \theta_2$: General Equilibrium”). Thus, with general equilibrium, TFP losses increase by 29 percent, whereas if the interest rate is unchanged, TFP losses increase only 13 percent.

The average wedge between the marginal product of capital and the interest rate increases from 1.62 percentage points to 2.74 percentage points, an increase of about 69 percent. Likewise, the standard deviation of the marginal product of capital increases from 2.55 percentage points to 3.51 percentage points, an increase of about 38 percent. Again, these increases are much larger than the increases that would obtain if the interest rate were held constant.

The final row in Table 3 illustrates how firms’ supply of financial assets affects workers’ ability to smooth consumption. To measure workers’ ability to smooth consumption, I examine the share of average consumption that workers would be willing to give up to have constant consumption over time. This measure is a standard tool in public finance and has been used by Lucas (1987, 2003) in a celebrated study of the welfare costs of business cycles. For a given worker, the thought experiment is: what share $(1 - x)$ of the worker’s average consumption $c$ provided in perpetuity would generate the same level of utility as the worker’s stochastic consumption stream? To measure the workers’ ability to smooth consumption, I take an average of $(1 - x)$ across all workers. Due to a homotheticity property of the workers’ problem, this measure $x$ does not depend on the wage. The homotheticity comes from the assumption of constant relative risk aversion and the linearity of the natural borrowing limit in the wage.

In the initial steady state, workers would be willing to give up 2.47 percent of their average consumption to have smooth consumption. At the new steady state, because of
the decreased availability of financial assets, workers’ consumption is riskier. The amount of average consumption that workers would be willing to give to up to have smooth consumption increases to 2.82 percent. This increase of 35 basis points is far larger than the share of consumption that a representative consumer would be willing to give up to eliminate business cycles, which Lucas (2003) calculated to be just 5 basis points for the same coefficient of relative risk aversion used here.

The spillover effect of increased riskiness of workers’ consumption is also meaningful compared with the spillover effect through the labor market, which is a decrease in the wage. Due to the decrease in firms’ ability to borrow, the steady-state wage declines by less than one percent.

The change in investment distortions and workers’ consumption smoothing can be better understood by examining the change in certain quantities and prices.

As shown in Table 4, if the interest rate were held constant, firms’ equilibrium supply of financial assets would decrease from 0.36 to 0.14, as a share of initial steady-state output. However, this decrease in firms’ supply of financial assets is not consistent with an equilibrium. To obtain equilibrium, the interest rate falls from 2.50 percent to 1.44 percent. Relative to the small-open-economy counterfactual, firms’ supply of financial assets increases, from 0.14 to 0.27. This still represents a reduction in the amount of financial assets supplied relative to the initial steady state, but it is a smaller reduction than would have occurred if the interest rate were unchanged. Of course, the equilibrium counterpart of a reduction in the financial assets supplied by firms is a reduction in the assets held by workers.

Table 4. Decrease in firms’ ability to borrow: Steady-state quantities and prices

| Quantities          | Partial equilibrium | General equilibrium | Increase |
|---------------------|---------------------|---------------------|----------|
| $E[d]$ Financial assets | 0.36                | 0.14                | 0.27     | -0.09 |
| $E[w]$ Firms’ liquid wealth | 0.88                | 1.02                | 0.96     | 0.08  |
| $E[k]$ Capital      | 1.15                | 1.07                | 1.14     | -0.01 |

All quantities are measured as a share of steady-state average sales conditional on $\theta = \theta_0$.  

The decrease in firms’ borrowing ability results in an increase in the steady-state liquid wealth of firms, as firms require more liquid wealth to pursue investment opportunities. However, the increase in liquid wealth is less than it would have been if the interest rate were unchanged. This, again, reflects the endogenous decrease in the interest rate.

While the change in the interest rate does amplify the distortions due to a decrease in firms’ borrowing ability, it also dampens the decrease in capital.

The general-equilibrium amplification of investment distortions can be further understood by considering how distortions vary with the interest rate. In this exercise, for a given interest rate, the steady-state investment distortions are computed. For each interest rate, the wage is chosen so as to clear the labor market. The results are given in Figure 1. The steady-state distortions are decreasing in the interest rate.

When firms’ ability to borrow decreases, conditional on the interest rate, steady-state investment distortions increase, as shown in Figure 2.

**Increase in consumers’ risk aversion**

This model is also useful for thinking about an increase in workers’ risk aversion or the riskiness of workers’ labor income. One might think that workers’ risk aversion or
Figure 2. Steady-state distortions as a function of the interest rate, for $\theta \in \{\theta_1, \theta_2\}$ Note: See Figure 1 or text for details.

Table 5. Increase in consumer risk aversion: Steady-state distortions

|                  | $\gamma_1 = 1$ | $\gamma_2 = 1.5$ | Difference |
|------------------|----------------|------------------|------------|
| TFP losses       | 2.45           | 2.72             | 0.28       |
| E[marginal product of capital]-r | 1.61           | 1.96             | 0.35       |
| Sd[marginal product of capital]  | 2.55           | 2.78             | 0.23       |

the riskiness of workers’ income increase as the result of financial crises. The effect of either change is to increase workers’ demand for financial assets, leading to a decrease in the interest rate and an increase in investment distortions for firms.

I consider an increase in consumers’ risk aversion from $\gamma_1 = 1$ to $\gamma_2 = 1.5$, holding constant the moral hazard parameter at the calibrated value. As shown in Table 5, there are increases in capital misallocation and the average finance premium firms would be willing to pay for an additional unit of borrowing. These are driven by an endogenous decrease in the interest rate from 2.50 to 1.52. These increases in investment distortions are much smaller than those caused by a decrease in firms’ ability to borrow, in part because the increases here are driven purely by the decrease in the interest rate.

There is also an increase in steady-state capital. An increase in steady-state capital is the same qualitative prediction that one would obtain from a standard macroeconomic model, such as Aiyagari (1994), that one would use to study an increase in workers’ risk aversion or an increase in the riskiness of workers’ labor income. However, standard
macroeconomic models would not predict the increase in investment distortions for firms.

The decrease in the interest rate also limits workers’ ability to smooth consumption. If the interest rate were held constant, then with the increased risk aversion, workers, on average, would be willing to give up 2.84 percent of their mean consumption in order to have a flat consumption path. However, with the decrease in the interest rate, this increases to 3.19 percent.

**Low moral hazard and Aiyagari (1994)**

The calibrated wedge between the inverse of the discount rate, $\frac{1}{\beta}$, and the interest rate is 4.33 percentage points. In contrast, Aiyagari (1994) finds a much smaller wedge, and indeed the smallness of the wedge he finds is a central point of his paper. Aiyagari (1994) uses a variety of parameterizations; for the parameterization of the worker’s problem that is similar to mine, Aiyagari (1994) finds a wedge of 0.41 percentage points.

The key difference between the models in this paper and Aiyagari (1994) is that the later does not include frictions on the production side that limit firms’ ability to create financial assets. In Aiyagari (1994), each period, all claims on physical capital are liquid assets held by consumers. For any $\theta < 1$, this is ruled out in my model.

To better understand how the firm-side friction in my model contributes to the low calibrated interest rate, consider a counterfactual in which $\theta = 1$. With $\theta = 1$, the firm is never constrained in its choice of investment, as the first-order condition for capital (1.7) shows. Thus, $\theta = 1$ corresponds to a “low” level of moral hazard. (A zero level of moral hazard would have $\theta = \infty$, whereas a higher level of moral hazard would have $\theta < 1$.) Calculating this counterfactual also serves a check on the numerical code, because the investment distortions are analytically known to be zero.

With $\theta = 1$, the wedge between the inverse of the discount factor and the interest rate is 0.71 percentage points. This is much closer to Aiyagari (1994)’s calibrated value. There are two reasons why the wedge is smaller for $\theta = 1$ than for $\theta = \theta_1 < 1$. First, for any choice of capital, a higher $\theta$ allows a greater amount of financial assets to be created. Second, a higher $\theta$ results in a larger capital stock.

With $\theta = 1$, workers would be willing to forgo only 0.79 percent of their mean consumption in order to have a smooth consumption path, rather than the 2.47 percent they would be willing to forgo when $\theta = \theta_0$. As predicted, all three measures of investment distortions
(TFP losses, the average wedge between the marginal product of capital and the interest rate, and the standard deviation of the marginal product of capital) are zero.

This underlines the importance of considering frictions on the firm side when studying the capacity of workers to self-insure by accumulating wealth.

1.6 Conclusion

This paper has studied how endogenous changes to the scarcity of liquidity affect the economy’s response to a decrease in firms’ ability to borrow. In my framework, distortions due to a decrease in firms’ ability to borrow are amplified by a reduced availability of liquidity.

The framework features a tractable model of the dynamic financing and investment choices of firms. It also features an important role for workers’ demand for financial assets. Workers’ need for financial assets creates an interesting channel through which firms’ decreased supply of liquidity can affect workers. Workers’ need for financial assets also affects firms’ response to tighter borrowing constraints, by making it attractive for firms to be suppliers of liquidity. Indeed, in the baseline economy with hand-to-mouth workers and constant entrepreneurial productivity, a decrease in firms’ ability to borrow has no effect on long-run distortions. This implies that any analysis of deleveraging should take into account the liquidity environment and how deleveraging changes the equilibrium scarcity of liquidity.

A calibration using firm-level data for U.S. entrepreneurs provides evidence that the amplification and spillover effects identified in the paper are important. A decrease in firms’ ability to borrow consistent with the recent decline in the long-term real interest rate results in an increase in TFP losses from capital misallocation of about 29 percent. If the interest rate were constant, as in a small open economy, the increase would be only 13 percent.

In current research, I am exploring the transition dynamics that arise after a decrease in firms’ ability to borrow. Understanding the transition dynamics will make possible a welfare analysis, in contrast to the positive analysis in this paper. Also, once the transition dynamics are better understood, we can ask whether temporary shocks to firms’ ability to borrow can generate the persistent effects that we associate with financial crises.

I am also exploring the business-cycle properties of the model and the ability of scarcity
of assets to amplify macroeconomic volatility at business-cycle frequency. In the present model, an increase in foreigners' demand for financial assets, for example, leads to a decrease in the interest rate and hence an increase in investment distortions. In a model with aggregate productivity shocks, it would be interesting to understand how the endogeneity of the interest rate and the wealth accumulation channel might amplify aggregate productivity shocks, or whether an increase in demand for financial assets from foreigners makes the economy more vulnerable to aggregate shocks.

This paper also raises interesting questions about policy responses to deterioration in the financial system.

First, for governments with credible regalian powers of taxation, there may be scope to reduce distortions through fiscal policy, by creating financial assets backed by future taxation. Ricardian equivalence does not hold here. My model could be used to analyze the optimal issuance of government debt, given that government debt can alleviate the shortage of financial assets. Aiyagari and McGrattan (1998) perform such an analysis in a Bewley economy; in my model, such an analysis would also include the liquidity benefits to firms of additional government debt. Since the introduction of frictions on the firm side can lead to a much larger wedge between the rate of time preference and the interest rate than in a Bewley economy (see Section 5), the quantitative results of this analysis may be quite different when borrowing frictions for firms are taken into account.

Second, if the government can provide insurance, for example, to workers, this may decrease their demand for liquidity and hence reduce distortions for firms. Of course, if the government cannot provide more insurance than the agents can achieve through trading in safe debt – for example, due to the hidden storage and hidden information problems in Cole Kocherlakota (2001) – then this policy implication has little bite. However, the lack of insurance might reflect, for example, a standard trade off between insurance and incentives to avoid moral hazard. Floden and Linde (2001) perform such an analysis in a Bewley economy. In my model, in making the trade off, one should take into account the effects on entrepreneurs of the decrease in workers’ demand for liquidity that occurs if insurance is improved.

Finally, this paper points out costs of a government-imposed decrease in firms’ borrowing ability; these costs come not only from increasing firms’ investment distortions, but also in the form of a decrease in the liquidity that workers can use to hedge their idiosyncratic
shocks.
Appendix A  Proofs and additional lemmas

Lemma 18  The expected marginal product of capital, $E[f_k(k,s)|s_-]$, is strictly decreasing in $k$ if and only if $F$ is strictly concave in capital and labor.

Proof. Define $l(\omega, k)$ by:

$$F_L(k, l(\omega, k), s) = \omega$$

By the envelope theorem,

$$\frac{\partial f}{\partial k} = F_k(k, l(\omega, k))$$

Hence

$$\frac{\partial^2 f}{\partial k^2} = F_{kk}(k, l(\omega, k)) + F_{kl}(k, l(\omega, k)) \frac{\partial l(\omega, k)}{\partial k}$$

$$= F_{kk} - F_{kl} \frac{F_{lk}}{F_{ll}}$$

Thus,

$$\frac{\partial^2 f}{\partial k^2} < 0 \text{ if and only if } F_{kk}F_{ll} > F_{kl}^2.$$  

\[\square\]

Proof of Lemma 1

Suppose that the production function is Cobb-Douglas

$$F(k, l, s) = sk^a l^b$$

where the elasticities of output to capital and labor are given by $a$ and $b$, respectively. Then output net of labor costs is given by

$$f(k, s) = s^{\frac{1}{1-\alpha}} k^a \left( \frac{b}{\omega} \right)^{\frac{1}{1-\alpha}} (1 - b)$$
\[ \alpha = \frac{a}{1-b}. \]

Define an entrepreneur's liquid wealth at date \( t \)

\[ w(s^t) = f(k(s^t-1), s_t) - d(s^t) \]

and consider the distribution over wealth \( w_t \) and productivity \( s_t \) given by \( \zeta_t(w, s) \).

Aggregate output in period \( t+1 \) will be given by:

\[
Y_{t+1} = \int \sum_{s_- \in S} \sum_{s \in S} sk(w, s_-)^a l(w, s_-, s)^b Q(s_-, s) \zeta_t(w, s_-) dw
\]

\[
= \int \sum_{s_- \in S} \left( \frac{b}{\omega} \right)^{\frac{1}{1-b}} E[s^{\frac{1}{1-b}} | s_-] k(w, s_-)^a \zeta_t(w, s_-) dw \tag{1.18}
\]

\[
= A_{t+1} K_t^a L_{t+1}^b \tag{1.19}
\]

where \( K_t \) is aggregate capital used in period \( t+1 \), \( L_{t+1} \) is aggregate labor hired in \( t+1 \), and \( A_{t+1} \) is measured total factor productivity (TFP).

\[
K_t = \int \sum_{s_- \in S} k(w, s_-) \zeta_t(w, s_-) dw \tag{1.20}
\]

\[
L_{t+1} = \int \sum_{s_- \in S} \sum_{s \in S} s^{\frac{1}{1-b}} k(w, s_-)^a \left( \frac{b}{\omega} \right)^{\frac{1}{1-b}} Q(s_-, s) \zeta_t(w, s_-) dw \tag{1.21}
\]

\[
A_{t+1} = \left[ \int \sum_{s_- \in S} E[s^{\frac{1}{1-b}} | s_-] \left( \frac{k(w, s_-)}{K} \right)^a \zeta_t(w, s_-) dw \right]^{1-b} \tag{1.22}
\]

To obtain (1.19), divide (1.18) by \( L_{t+1}^b \), using the expression in (1.21) for \( L_{t+1} \). Then divide by \( K_t^a \).

**Proof of Proposition 4**

This proof relies on the result that if each entrepreneur’s productivity is constant over time, then in steady-state, each entrepreneur’s wealth and consumption is constant over time. This intuitive technical result is proved in Lemma 6.

Part (i). When productivity is constant, the first-order conditions for capital and next-
period wealth:

\[ u'(c)\left[ \frac{1}{R} f_k(k, s) - 1 \right] = \phi(1 - \theta) f_k(k, s) \]

and

\[ \phi = -\beta V_w(w, s) + u'(c) \frac{1}{R} \]

In steady-state, each entrepreneur’s wealth is constant over time, as shown in Lemma 6. Substituting \( V_w(w, s) = u'(c_s) = u'(c) \) yields the expression for \( f_k(k, s) \) in the proposition and shows that the financial constraint is binding if and only if \( \beta R < 1 \).

Part (ii). In steady state, an entrepreneur with productivity \( s \) invests

\[ k = \left[f_k(\cdot, s)\right]^{-1}(\frac{1}{\theta R + (1-\theta) \beta}). \]

If \( \beta R < 1 \), an entrepreneur with productivity \( s \) borrows \( d = (1 - \theta) f(k, s) \). If \( \beta R = 1 \), an entrepreneur with productivity \( s \) borrows \( d \leq (1 - \theta) f(k, s) \). The inequality is due to the multiplicity of invariant distributions when \( \beta R = 1 \), as discussed in Lemma 6. Thus, the supply of financial assets is increasing in the interest rate.

**Proof of Lemma 6**

This proof relies on the concavity of the value function, which is demonstrated in Proposition 8.

Re-arranging the first-order conditions, we can obtain:

\[ \frac{1}{R} - \frac{1}{f_k(k)} = (1 - \theta) \left( \frac{1}{R} - \beta \frac{V'(w_s)}{V'(w)} \right) \]  

(1.23)

where \( w \) is current-period wealth and \( w_s \) is next-period wealth.

(i) Suppose \( R < \frac{1}{\beta} \). The following is a steady-state: an entrepreneur with productivity \( s \) has wealth \( (1 - \theta) f([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R + (1-\theta) \beta}), s) \), invests capital \([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R + (1-\theta) \beta})\), and has next-period wealth equal to current wealth. This satisfies the first-order conditions and the complementary slackness condition; in particular, the constraint is binding. We can prove by contradiction that there is no other steady state. Consider a candidate invariant distribution with positive mass on wealth levels not equal to \( (1 - \theta) f([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R + (1-\theta) \beta}), s) \). Condition (1.23) implies that all entrepreneurs with wealth less than \( (1 - \theta) f([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R + (1-\theta) \beta}), s) \) will have next-period wealth greater than current wealth. Likewise, all entrepreneurs with wealth greater than \( (1 - \theta) f([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R + (1-\theta) \beta}), s) \)
will have next-period wealth less than current wealth. Hence the candidate distribution is not an invariant distribution.

(ii) Suppose \( R = \frac{1}{\beta} \). Any distribution over wealth and productivity is a steady state if the measure of entrepreneurs with wealth greater than or equal to \((1 - \theta)f(k^u(R), s)\) is equal to one. To see this, note that with wealth greater than or equal to \((1 - \theta)f(k^u(R), s)\), next-period wealth equal to current wealth and capital equal to \(k^u(R)\) satisfies the first-order conditions and the complementary slackness condition; in particular, the financial constraint is not binding. To see that there cannot be an invariant distribution with a positive measure on wealth less than \((1 - \theta)f(k^u(R), s)\), note that an entrepreneur with less than \((1 - \theta)f(k^u(R), s)\) will have capital less than \(k^u(R)\) and hence, from condition (1.7), next-period wealth that exceeds current wealth.

(iii) In the invariant distributions identified in parts (i) and (ii), next-period wealth equals current wealth.

Proof of Proposition 7

(i) Assume that workers are hand-to-mouth workers.

Suppose there were a steady-state with \( \beta R < 1 \). Then, based on the partial-equilibrium analysis of Proposition 4, the financial constraint of each entrepreneur would bind and capital would equal \( k = \left[ f_k(\cdot, s) \right]^{-1} \left( \frac{1}{\theta R + (1 - \theta)\beta} \right) \). Hence each entrepreneur would supply a positive amount of financial assets and entrepreneurs’ aggregate supply of financial assets would be strictly positive. Since workers demand for savings is zero, this cannot be a steady-state equilibrium.

If \( \beta R > 1 \), entrepreneurs’ wealth would be increasing over time. This is inconsistent with an invariant distribution for wealth and with market clearing, as it would result in an aggregate supply of financial assets that converges to negative infinity.

If \( \beta R = 1 \), there is a multiplicity of general-equilibrium steady states. Consider the condition:

\[
\sum_{s \in S} \int f(k^u(R, s), s)\zeta(w, s)dw = \sum_{s \in S} \int w\zeta(w, s)dw \tag{1.24}
\]

where \( \zeta(w, s) \) is a distribution over wealth and productivity. This condition says that the total liquid wealth of the entrepreneurs equals the total output net of labor costs when
investment is unconstrained. With hand-to-mouth workers, this condition is equivalent to market-clearing in the financial market. Consider also the condition:

\[ Pr[ζ \left[w \geq (1 - \theta)f(k^u(R), s)\right]] = 1 \quad (1.25) \]

where the probability distribution over wealth and productivity is given by \(ζ\). This condition says that the wealth of each entrepreneur is sufficiently high such that the choice of investment is unconstrained. For \(R = \frac{1}{\beta}\), this condition is required for a steady state, as stated in Lemma 6.

Any distribution \(ζ(w, s)\) that satisfies (1.24) and (1.25) is a general equilibrium steady state. Moreover, a comparison of the two conditions shows that there is an infinity of distributions \(ζ\) that satisfy both.

(ii) Assume that workers can save.

There cannot be a general equilibrium steady state with \(βR = 1\), because with \(βR = 1\), workers’ financial assets converge to infinity almost surely, as shown in Lemma 13. In contrast, for any invariant distribution \(ζ^*\) for \(βR \leq 1\), firms’ capital is finite and hence firms’ supply of financial assets must be finite.

There is at least one general equilibrium steady state with \(βR < 1\). To see this, first define the workers’ steady state asset demand as a function of the interest rate:

\[ A(R) = \int \int a^*(a, y; R)ϕ^*(a, y; R) \, da \, dy \]

Note that the policy function \(a^*\) and the distribution \(ϕ^*\) depend on the interest rate and the wage, which is a continuous function of the interest rate. A standard result is that, for a given interest rate, there is a unique invariant distribution \(ϕ^*(a, y; R)\). From Bewley (1984) and Clarida (1990), we have the following results:

\[ \lim_{R \uparrow \frac{1}{\beta}} A(R) = \infty \]

\[ \lim_{R \downarrow 0} A(R) = a \]
From Aiyagari (1994), and by application of a standard theorem in Stokey-Lucas-Prescott (1989), we know that $A(R)$ is continuous in $R$. We also know that firms’ supply of financial assets

$$D(R) = \theta \sum_{s \in S} \int f([f_k(\cdot, s)]^{-1}(\frac{1}{\theta R} + (1 - \theta)\beta), s)\zeta^*(w, s; R)dw$$

is strictly positive, finite and continuous in $R$ for any $R \in (0, \frac{1}{\beta})$. Moreover,

$$\lim_{R \downarrow 0} D(R) = \infty$$

Hence there is at least one general-equilibrium steady state and any general-equilibrium steady state features $\beta R < 1$.

**Proof of Proposition 8**

With constant productivity, the marginal product of capital of a firm with productivity $s$ and moral hazard $\theta$ is given by (1.14). Thus, it is immediate that part (i) of the proposition holds.

For part (ii), we can combine the assumption of Cobb-Douglas technology with (1.14) to solve for capital $k(R, \theta, s)$ as a function of $s$ and $\theta$, and substitute into (1.10) to find a closed-form expression for TFP.

In particular, TFP equals $X^{1-b}$, where $X$ is given by

$$X = \frac{1}{K^\alpha} \sum_s s^{\frac{1}{1-b}} \sum_\theta k(R, \theta, s)^\alpha g(\theta, s)d\theta$$

where aggregate capital $K$ is given by

$$K = \sum_s \sum_\theta k(R, \theta, s)g(\theta, s)d\theta$$

and $g(\theta, s)$ is the distribution function over $\theta$ and $s$.

Firm capital $k(R, \theta, s)$ is given by:

$$k = (\alpha s^{\frac{1}{1-b}})^{-\frac{1}{\alpha}} \left(\frac{1}{R\theta} + \beta(1 - \theta)\right)^{\frac{1}{1-\alpha}}$$

(1.28)
Substitute (1.28) into (1.27) and (1.26) and take the derivative of (1.26) with respect to \( R \). Suppose that \( \beta R < 1 \).

Let \( S = \{s_1, s_2, \ldots, s_{\#S}\} \). Let \( \Theta = \{\theta_1, \theta_2, \ldots, \theta_{\#\Theta}\} \).

After some algebra, one finds that

\[
\frac{dX}{dR} > 0
\]

if and only if

\[
\sum_{i=1}^{\#S} \sum_{j=1}^{\#\Theta} \sum_{m=1}^{\#S} \sum_{n=1}^{\#\Theta} [z_i y_j^0 g(\theta_j, z_i) z_m y_n^0 \theta_n g(\theta_n, z_m)] > \sum_{i=1}^{\#S} \sum_{j=1}^{\#\Theta} \sum_{m=1}^{\#S} \sum_{n=1}^{\#\Theta} [z_i z_m y_j^{2\alpha - 1} \theta_j g(\theta_j, z_i) y_n g(\theta_n, z_m)]
\]

(1.29)

where, to economize on notation,

\[
z_i = s_i^{1 - \alpha}
\]

(1.30)

\[
y_j = \left( \frac{1}{R} \theta_j + \beta(1 - \theta_j) \right)^{1 - \alpha}
\]

(1.31)

There are \((\#S)^2(\#\Theta)^2\) terms on the left-hand side and an equal number on the right-hand side of (1.29). Of these, \((\#S)^2(\#\Theta)\) terms on each side feature \( j = n \), and all of these terms cancel. In particular, if \( j = n \), the left-hand term \((i,j,m,n)\) equals

\[
\theta_j z_i z_m y_j^{2\alpha} g(\theta_j, z_i) g(\theta_j, z_m)
\]

which is equal to the right-hand term \((i,j,m,n)\).

This leaves \((\#S)^2((\#\Theta)^2 - (\#\Theta))\) terms with \( j \neq n \). Pair terms \((i,j,m,n)\) and \((m,n,i,j)\). I will show that the sum of each left-hand-side pair is strictly greater than the sum of each right-hand-side pair. That is, I will show that

\[
z_i z_m \theta_n y_j^0 y_n^0 g(\theta_j, z_i) g(\theta_n, z_m) + z_m z_i \theta_j y_n^0 y_j^0 g(\theta_n, z_m) g(\theta_j, z_i) > z_i z_m \theta_j y_j^{2\alpha - 1} y_n g(\theta_j, z_i) g(\theta_n, z_m) + z_m z_i \theta_n y_n^{2\alpha - 1} y_j g(\theta_n, z_m) g(\theta_j, z_i)
\]
With some algebra, this inequality reduces to

\[
\theta_n \left( \frac{\frac{1}{R} \theta_n + \beta (1 - \theta_n)}{\frac{1}{R} \theta_n + \beta (1 - \theta_n)} \right) - \theta_j \left( \frac{\frac{1}{R} \theta_n + \beta (1 - \theta_n)}{\frac{1}{R} \theta_j + \beta (1 - \theta_j)} \right) > \theta_j \left( \frac{\frac{1}{R} \theta_n + \beta (1 - \theta_n)}{\frac{1}{R} \theta_j + \beta (1 - \theta_j)} \right)
\]

Suppose, without loss of generality, that \( \theta_n > \theta_j \), remembering that we are only considering terms with \( j \neq n \). The assumption \( \theta_n > \theta_j \) implies that (1.32) holds.

**Proof of Proposition 9**

Because the utility function may be unbounded, I restrict the domain of \( V \) to be \( W \times S \), where \( W = [0, \hat{w}] \). If \( \hat{w} \) is chosen to be sufficiently large, this restriction of the domain will not affect the analysis.\(^{18}\) In order to prove the proposition, I will verify the conditions required in Exercise 9.7 of Stokey-Lucas-Prescott (1989), hereafter SLP. To parallel SLP, the notation is:

\[
V(w, s-) = \max_{\{k, \{d_s\} \in \Gamma(w, s-) \}} G(w, k, \{d_s\}, s-) + \beta E[V(\psi(w, k, \{d_s\}, s), s)]
\]

First, some important preliminaries.

. \( W \) is a convex Borel set in \( \mathbb{R} \), with its Borel subsets \( W \).
. \( S \) is a countable set and \( S \) is the \( \sigma \)-algebra containing all subsets of \( S \).
. Define \( \Gamma : W \times S- \rightarrow K \times D^S \) by

\[
\Gamma(w, s-) = \{k, \{d_s\} : d_s \leq \theta f(k, s) \text{ and } w + \frac{1}{1 + r} E[d_s] - k \geq \varepsilon \text{ and } f(k, s) - d_s \leq \hat{w} \}
\]

The second and third conditions in the definition of \( \Gamma \) are not part of the economic problem; they are technical conditions useful for the proof. Because they never bind, it is acceptable to neglect them in the main text.

To obtain the proposition, I need to verify that the following conditions hold:

Condition 1: The constraint set \( \Gamma \) is non-empty, compact-valued and continuous. To see that it is non-empty valued, note that given \( s_- \)

\[
k_{\min} = \max_k -k + \frac{1}{R} \theta E[f(k, s)|s_-]
\]

\(^{18}\)For \( \theta < 1 \), we can use any \( \hat{w} \geq \bar{w} \equiv (1 - \theta) f(k^u(R, \cdot), \bar{s}) \). For \( \theta = 1 \), any \( \hat{w} > 0 \) will suffice.
and

\[ d_s = \theta f(k_{\text{min}}, s) \]

are always feasible. \( \Gamma \) is compact valued because it is closed and bounded.

**Condition 2: The payoff function \( G \) is bounded, continuous and strictly increasing in \( w \).**

Let \( A \) be the graph of \( \Gamma \). Define \( G : A \to \mathbb{R} \) by

\[ G(w, k, \{d_s\}, s_-) = u(w + \frac{1}{1 + r} E[d_s|s_-] - k) \]

\( G \) is bounded and continuous, and \( \beta \in (0, 1) \). \( G \) is bounded below by \( u(\varepsilon) \). \( G \) is bounded above by

\[ \max_{s_-} u(\bar{w} + \frac{\theta}{R} E[f(k_{\text{min}}(s_-), s)|s_-] - k_{\text{min}}(s_-)) \]

For each \((k, \{d_s\}, s_-) \in R_+ \times R^S \times S\), we have \( u(w + \frac{1}{1 + r} E[d_s|s_-] - k) \) strictly increasing in \( w \), by the assumption that \( u(c) \) is strictly increasing.

**Condition 3: The constraint set is increasing in \( w \).** For each \( s_- \in S \), \( \Gamma(\cdot, s_-) : W \to K \times \mathbb{R}^S \) is increasing in the sense that \( w \leq w' \) implies \( \Gamma(w, s) \subseteq \Gamma(w', s) \). The first and last sets of restrictions in the definition of \( \Gamma \) do not involve \( w \); a higher \( w \) makes it easier to satisfy the second set of restrictions.

**Condition 4: Concavity restriction on the payoff function.** For each \( s_- \in S \), \( G : A_{s_-} \to \mathbb{R} \) satisfies the following concavity restriction:

\[
\begin{align*}
\varphi w + (1 - \varphi)w' + \frac{1}{R} E[\varphi d_s + (1 - \varphi)d'_s|s_-] - (\varphi k + (1 - \varphi)k') \\
\geq \varphi u(w + \frac{1}{1 + r} E[d_s|s_-] - k) + (1 - \varphi)u(w' + \frac{1}{1 + r} E[d'_s|s_-] - k')
\end{align*}
\]

for all \( \varphi \in (0, 1) \) and \((w, k, \{d_s\}), (w', k', \{d'_s\}) \in A_{s_-} \). This is due to the concavity of \( u \).

**Condition 5: No “increasing returns” in \( \Gamma \).** For all \( s_- \in S \) and all \( w, w' \in W \), we have \( \{k, \{d_s\}\} \in \Gamma(w, s_-) \) and \( \{k', \{d'_s\}\} \in \Gamma(w', s_-) \) implies \( \varphi \{k, \{d_s\}\} + (1 - \varphi)\{k', \{d'_s\}\} \in \Gamma(\varphi w + \varphi w', s_-) \) for all \( \varphi \in (0, 1) \). The first and last set of constraints satisfy this condition.
because they are linear. The second set of constraints satisfies this condition by weak concavity of \( f \).

**Condition 6: Differentiability of the payoff function.** For given \( s_- \in S \),

\[
G(w,k,\{d_s\},s_-) = u(w + \frac{1}{1+r}E[d_s|s_-] - k)
\]

is continuously differentiable in \((w,k,\{d_s\})\) on the interior of \( A_{s_-} \). This is immediate from the assumption that \( u \) is twice differentiable.

**Condition 7: Restrictions on the mapping to next-period state variables.** Define \( \psi(w,k,\{d_s\},s) = f(k,s) - d_s \).

- \( \psi \) is non-decreasing in \( w \).
- For each \( s \), \( \psi(.,s) \) is weakly concave, because \( f \) is concave in \( k \).
- \( \psi(.,s) \) does not depend on \( w \).

**Proof of Proposition 10**

This proof is based on theorem 12.12 of SLP (1989). I need to verify that the transition function \( P_E \) defined by (1.12) is monotone, has the Feller property and satisfies a mixing condition.

- **Feller property.** \( P \) has the Feller property if \( \int h(w,s)P((w,s_-),(w,s)) \) is bounded and continuous in \( w \) for any function \( h \) that is bounded and continuous in \( w \). SLP theorem 9.14 implies that \( P \) has the Feller property if \( g^w(w,s_-) \) is continuous in \( w \), as verified in Proposition 8.

- **Monotonicity.** The monotonicity condition requires that \( \int h(w,s)P((w,s_-),(w,s)) \) be increasing in \( w \) if \( h \) is bounded, continuous, and increasing in \( w \). The monotonicity of \( w_s \) with respect to \( w \) shown in Lemma 14, together with SLP exercise 12.11, imply monotonicity.

- **Mixing condition.** I must show that there exists \( w^* \in [0,\bar{w}] \), \( \xi > 0 \), and \( N \geq 1 \) such that \( P^N(0,[w^*,\bar{w}]) \geq \xi \) and \( P^N(\bar{w},[0,w^*]) \geq \xi \). Take \( w^* = f(k_{\min},\min_s(s \in S : s \neq 0)) \). For arbitrary \( w \), consider the optimal policy for \( w_0 \), next-period wealth if realized productivity is zero. If \( \phi_0 \) is not binding, then \( V'(w_0) = u'(c_0) = \frac{1}{\beta R}u'(c) = \frac{1}{\beta R}V'(w) \). If \( \phi_0 \) is binding, then \( w_0 = 0 \). Thus, \( P^N(\bar{w},[0,w^*]) \geq \xi \equiv Q(0)^N \) for \( N \) equal to the smallest integer larger
than \( \frac{\log[u'(c^*(w^*))]}{\log(\frac{1}{\beta R})} \).

**Proof of Proposition 11**

First, I show that \( \beta R = 1 \) implies the measure of constrained firms is zero.

Suppose \( \beta R = 1 \). Suppose that the distribution today \( \zeta_t \) is an invariant distribution: \( \zeta_t = T^* \zeta_t \). Suppose further, by contradiction, that the mass of entrepreneurs with at least one financial constraint is binding has positive measure. For these entrepreneurs, consumption tomorrow will be strictly higher than today, as indicated by the first-order condition (1.8). For all other entrepreneurs, consumption tomorrow will be the same as consumption today. Hence, aggregate consumption tomorrow is higher than aggregate consumption today:

\[
\sum s_- \int c^*(w, s_-) \zeta_t(dw, s_-) < \sum s_- \int c^*(w, s_-) \zeta_{t+1}(dw, s_-)
\]

Thus, \( \zeta_t \neq \zeta_{t+1} \) and hence \( \zeta_t \) is not an invariant distribution.

Second, I show that \( \beta R < 1 \) implies the measure of constrained firms is positive.

This proof with very similar to the previous analysis, with appropriate modifications.

Suppose \( \beta R < 1 \). Suppose that the distribution today \( \zeta_t \) is an invariant distribution: \( \zeta_t = T^* \zeta_t \). Suppose further, by contradiction, that the mass of entrepreneurs with at least one binding financial constraint has zero measure. Hence, using the first order condition (1.8), aggregate consumption tomorrow is lower than it is today:

\[
\sum s_- \int c^*(w, s_-) \zeta_t(w, s_-)dw > \sum s_- \int c^*(w, s_-) \zeta_{t+1}(w, s_-)dw
\]

Thus, \( \zeta_t \neq \zeta_{t+1} \) and hence \( \zeta_t \) is not an invariant distribution.

From Proposition 9, we know that an invariant distribution exists. Combining these two negative results with the existence of an invariant distribution delivers Proposition 10.

**Proof of Proposition 12**

Part (i). In steady state, there is a positive measure of firms with at least one binding financial constraint if and only if \( \beta R < 1 \), according to Proposition 10. The first-order
condition for capital (1.7) implies that firms have a marginal product of capital greater than the interest rate if and only if the firm has binding financial constraints and \( \theta < 1 \).

Part (i) holds whether or not \( S \) has more than one element.

Parts (ii). Suppose \( \beta R < 1 \). Any firm with a binding constraint and \( \theta < 1 \) will have next-period wealth that depends on realized productivity. Hence, the wealth distribution will not be degenerate. The measure of firms with at least one binding financial constraint is either: less than one; or one. If it is less than one, then immediately we have that there is a positive measure of firms with marginal product of capital equal to the interest rate, and a positive measure of firms with marginal product of capital greater than the interest rate. If the measure of constrained firms is one, then the variance of marginal product of capital is positive, since capital is increasing in wealth if a constraint is binding, according to Lemma 14, and the wealth distribution is non-degenerate.

If \( \beta R = 1 \), then in steady-state, the financial constraints are not binding. The first-order condition for capital therefore implies that the marginal product of capital equals the interest rate. If \( \theta = 1 \), then the marginal product of capital equals the interest rate, regardless of whether financial constraints bind.

Part (iii). Suppose \( \beta R < 1 \). Any firm with a binding constraint and \( \theta < 1 \) will have next-period wealth, and hence next-period consumption, that depends on realized productivity. If \( \beta R = 1 \), in steady state, next-period consumption is the same as today’s, regardless of realized productivity.

**Proof of Lemma 15**

Denote the optimal policy for consumption by:

\[
e^*(w, s_-; \mu) = w + \frac{1}{1 + r(\mu)} E[f(k^*, s) - w^*_s] - k^*
\]

That \( e^* \) is increasing in \( w \) is immediate from the envelope condition, \( V'(w, s_-) = u'(e^*) \), and the concavity of \( V \) and \( u \).

Consider a given \((w, s_-)\). Consider any states of the world for which the financial constraint is not binding. The first-order condition for wealth in state \( s \) when the financial
constraint in state $s$ is not binding is
\[ u'(c) \frac{1}{\beta R} = V_w(w_s, s; \mu') \]

Since $c$ is increasing in $w$, this first-order condition implies that $w_s$ must be increasing in $w$ for any $s$ for which the financial constraint is not binding.

In the states for which the financial constraint is binding, $w_s = (1 - \theta) f(k, s)$. Hence if we can show that $k$ is increasing in $w$ whenever the financial constraint is binding for at least one state of the world, we will have shown that $\{w_s\}$ is increasing in $w$.

Combining the first-order conditions for $w_s$ and $k$, we obtain the following, which holds almost everywhere:
\[
\frac{1}{R(\mu)} E[f_k(k, s)] - 1 - \frac{1}{R(\mu)} \sum_{s \in S_r} (1 - \theta) q_s f_k(k, s) = -(1 - \theta) \beta \sum_{s \in S_r} q_s \frac{V_w((1 - \theta)f(k, s), s; \mu')}{u'(c)} f_k(k, s)
\]

The left-hand-side in decreasing in $k$, due to the concavity of $f$ in $k$. The right hand side is increasing in $k$, due to the concavity of $f$ in $k$ and $V$ in $w$. Hence, because $c$ is increasing in $w$, we must have that $k$ is increasing in $w$ whenever any of the financial constraints bind (that is, whenever $S_r$ is not the empty set).

Finally, to see that $c^*$ is increasing less than one-for-one with $w$, consider $w$ and $w'$, with $w' > w$, with optimal policies $(k, \{w_s\})$ and $(k', \{w'_s\})$. Suppose that $c' \geq c + (w' - w)$, or, equivalently, \( \frac{1}{1 + r(\mu)} E[f(k, s)] - w_s - k \leq \frac{1}{1 + r(\mu)} E[f(k', s)] - w'_s - k' \). Then $(k', \{w'_s\})$ would have been a better choice than $(k, \{w_s\})$ when wealth equals $w$; it is feasible, produces higher consumption than $(k, \{w_s\})$, and, since $w^*_s$ is increasing in $w$, it must offer higher utility tomorrow.

**Proof of Lemma 16**

For any $(w, s_-)$, it must be case that $w_s \geq (1 - \theta)f(k_{\text{min}}, s) > 0$ for all $s$. The first inequality follows from the fact that the firm will never choose a $k$ below $k_{\text{min}}$ and from the financial constraint for $s$; the second follows from the assumptions on $f$. Denote the maximum possible marginal utility of consumption tomorrow in state $s$ by $x_s = u'(c^*((1 - \theta)f(k_{\text{min}}, s), s)) = V_w((1 - \theta)f(k_{\text{min}}, s), s) < \infty$. This is the maximum possible marginal
utility tomorrow in state s because $V(.,s)$ is concave.

Note that for any $(w,s_-), c$ is bounded above by $w - w_{l,s_-}$.

Consider some $s_-$. For every $s$, we can pick an $\varepsilon_{s,s_-} > 0$ such that, if $w \leq w_{l,s_-} + \varepsilon_{s,s_-}$, we have

$$\frac{1}{\beta R} u'(c) > \frac{1}{\beta R} u'(\varepsilon_{s,s_-}) > x_s \geq V_w(w_s, s)$$

Suppose that constraint $s$ is not binding tomorrow. Then $V_w(w_s, s)$ must equal $\frac{1}{\beta R} u'(c)$, as shown by (1.7). This is a contradiction.

Since the set of $s$ is discrete, we can define

$$w_{s_-} = \min_s (\varepsilon_{s,s_-}) > 0$$
Appendix B: Multi-period promises

The following proposition makes precise the previous assertion that it is without loss of generality to restrict entrepreneurs to trading one-period promises.

Consider an environment parameterized by \( n \geq 1 \) in which entrepreneurs at history \( s^t \) can trade state-contingent promises that pay out at history \( s^{t+i} \), for \( i = \{1, ..., n\} \). Denote by \( x(s^t, s^{t+i}) \) the amount of payment that the entrepreneur at history \( s^t \) promises to pay at history \( s^{t+i} \). The entrepreneur’s budget constraint is

\[
c(s^t) + k(s^t) \leq f(k(s^{t-1}), s^t) - x(s^{t-1}, s^t) + \sum_{i=1}^{n} \sum_{s^t \rightarrow s^{t+i}} \frac{1}{R^t} E_t [x(s^t, s^{t+i}) - x(s^{t-1}, s^{t+i})] \tag{1.33}
\]

where \( x(s^{t-1}, s^{t+n}) = 0 \) since at \( t-1 \), the entrepreneur cannot trade promises to pay in period \( t + n \).

If the entrepreneur reneges at \( s^t \), this eliminates both his debt that is immediately due \( x(s^{t-1}, s^t) \) and his promises to repay in future periods, \( \{x(s^{t-1}, s^{t+i})\}_{i=1}^{n-1} \). As above, when the entrepreneur reneges, his creditors can seize only a fraction \( \theta \) of his current period output, \( f(k, s) \). Hence, given the entrepreneur’s budget constraint, the entrepreneur will not renege at \( s^t \) if and only if

\[
\sum_{i=0}^{n-1} \sum_{s^t \rightarrow s^{t+i}} \frac{1}{R^t} x(s^{t-1}, s^{t+i}) \leq (1 - \theta) f(k(s^t), s^t) \tag{1.34}
\]

The entrepreneur now chooses a collection \( \{k(s^t), c(s^t), \{x(s^t, s^{t+i})\}_{i=1}^{n}\} \) to maximize utility (1.4) subject to a no-Ponzi condition and a collection of budget constraints (1.33) and financial constraints (1.34) for each \( s^t \).

**Lemma 19** Consider a solution to the entrepreneur’s problem with \( n \) securities given by \( \{k(s^t), c(s^t), \{x(s^t, s^{t+i})\}_{i=1}^{n}\} \). There exists a policy of one-period positions \( d(s^{t+1}) \) promised at \( s^t \) to pay out at \( s^{t+1} \)

\[
d(s^{t+1}) = \sum_{i=1}^{n} \frac{1}{R^{t+1}} E_{t+1} [x(s^{t+i}, s^t)|s^{t+1}]
\]

66
that, together with the same policies for capital and consumption \( \{k(s^t), c(s^t)\} \), satisfy the budget (1.2) and financial (1.1) constraints of the problem with one-period promises and achieve the same utility.

**Proof.** This comes from comparing the expression for \( d(s^{t+1}) \) in the lemma to condition (1.34). ■
Appendix C  Numerical approach

The parameters of the Cobb-Douglas production technology and the productivity process were estimated using firm-level data for U.S. entrepreneurs from the Kauffman Firm Survey. The estimator is a modified Olley-Pakes (1996) estimator. This estimator addresses the simultaneity bias by semi-parametrically controlling for unobserved productivity.

Firms from finance and insurance (NAICS code 52) and real estate rental and leasing (NAICS code 53) were excluded from the analysis. Also, attention was restricted to sole proprietorships and Subchapter S corporations, for which aggregate leverage ratios could be obtained.

The results of the estimation are given in Table 1. Bootstrapped standard errors will be included pending data-disclosure approval.

The aggregate leverage ratio is a measure of aggregate debt and equity to aggregate sales. Data on debt and sales is obtained from the Internal Revenue Service (IRS) Statistics of Income (SOI), which provides data for sole proprietorships and S corporations disaggregated by NAICS code. I use the average ratio of debt to sales for 1998-2003, the period for which data is available.

By equity, this paper means equity held by investors, rather than the entrepreneur. For sole proprietorships, all the equity is held by the entrepreneur. For S corporations, there can be up to 100 equity holders. Data on equity in S corporations held by non-entrepreneurs is obtained from the Survey of Consumer Finances (SCF). I use the standard definition of non-entrepreneur, namely, people owning a stake in S corporations but reporting that they have no active management role in the firm. The SCF has weights that allow aggregate calculations; see Curtin, Juster, and Morgan (1989) and Cagetti-DiNardi (2006) for details. Corresponding to the IRS data availability, I used data from the 1998, 2001 and 2004 SCF.

*Computing the stationary equilibrium.* The central exercise is to calculate the steady-state equilibrium for given fundamental parameters. To do so, the following procedure is used:

1. Guess a value for the interest rate R.
2. Solve the firm’s problem, conditional on R and a guess for the wage \( \omega \). Since the model features state contingent financing, the firm’s problem requires each firm to choose capital and a schedule of next-period repayments. Solving for a schedule of repayments
rather than an uncontingent repayment increases the number of choice variables, but it does not increase the number of state variables required. Because the value function is concave in wealth and because the set of productivities $S$ is discrete, the value function can be approximated by a collection of one-dimensional Chebyshev polynomials, which are computationally efficient.

At each value-function iteration, the new value function is calculated based on the previous guess and the firm’s optimization. This new value function is calculated for a grid of points, where spacing of the grid points is (roughly) exponential, as opposed to the traditional Chebyshev spacing. This spacing of grid points allows a high density of points at the wealth levels of interest while avoiding unintentional restrictions on the state space for wealth. The firm’s choice of capital and next-period wealth are not restricted to lie on a grid. The new guess for the value function is then determined by Chebyshev interpolation.

**Step 3.** Simulate 12,000 firms for $T$ periods, where $T$ is taken to be 500 periods, to find the steady-state distribution over wealth and productivity.

**Step 4.** Solve for wage rate that would clear the labor market. This is straightforward, since the firm’s problem is homothetic in the following sense: a change in the wage from $\omega$ to $z\omega$ will result in an increase in steady-state capital from $k$ to $z^{-\frac{1}{1-\delta}} \frac{1}{1-\alpha} k$. This property comes from the assumption of constant relative risk aversion and simplifies the numerical task considerably.

**Step 5.** Solve the worker’s problem and determine workers’ demand for financial assets.

**Step 6.** Check financial market clearing. If workers’ demand is greater [less] than firms’ supply of financial assets, choose a new guess for $R$ that is lower [higher] than the previous guess.

**Step 7.** Repeat steps 2-6 until the financial market clears.

Like the firm’s problem, the worker’s problem has a useful homotheticity property. In particular, asset demand at a given interest rate is linear in the wage. This is due to the assumption of constant relative risk aversion; this assumption is made in the calibration exercise. The homotheticity also depends on the borrowing constraint being zero or linear in the wage, a property satisfied by the natural borrowing constraint. The homotheticity of the workers’ problem means that, once the workers’ asset demand is calculated for a given interest rate, it can be calculated immediately for any wage required for the various counter-factual exercises.
How to deal with non-entrepreneurial firms has been a persistent problem in this literature. One approach has been to include non-entrepreneurial firms as a constant-returns-to-scale sector that does not face any financial constraints (Quadrini 2000, Cagetti and DiNardi 2006). One problem with this approach is that even large, public firms face financial frictions, if not exactly of the type or severity faced by entrepreneurs. A second approach has been to ignore broad swaths of the economy, even in general equilibrium exercises. For example, in a general-equilibrium calibration meant to quantify TFP losses from financial frictions, Moll (2010) combines aggregate current-account and output data with estimates of productivity-process and production-function parameters for manufacturing plants, ignoring non-manufacturing firms; Moll’s calibration also features hand-to-mouth consumers (or, equivalently for steady-state analysis of workers’ demand for financial assets, workers without labor-productivity and unemployment risk and with an ad-hoc zero borrowing limit) and thus misses the important role of the savings of non-entrepreneurs.

In this part of the appendix, I incorporate non-entrepreneurial firms as a Lucas tree. This creates a gap between the equilibrium supply of financial assets by entrepreneurs and the equilibrium demand for financial assets by workers. In the calibration, I require that the value of the Lucas tree equal an estimate of the historical value of U.S. corporate equities and bonds held by households as a share of GDP, or 0.58. It should be noted that this value likely overstates the importance of non-entrepreneurial firms, since some entrepreneurial firms are incorporated, using Subchapter S, Chapter C or other corporate forms. (For example, Koch Industries, Inc., is a privately-held conglomerate run by Chairman and CEO Charles G. Koch, who has a major ownership stake in the firm. Koch Industries, Inc. has about 70,000 employees according to its website and is estimated by Forbes to have annual revenue of 100 billion dollars.)

With the non-entrepreneurial Lucas tree, the calibrated value for the discount factor $\beta$ is 0.96, and the calibrated gap between the interest rate and the rate of time preference is thus smaller. Likewise, the calibrated value for firms’ ability to borrow is $\theta_{1a} = 0.88$, which is greater than in the analysis without the non-entrepreneurial Lucas tree. As a result, the calibrated steady-state TFP losses from entrepreneurial firms are significantly smaller, as shown in Table 6 below. Here again the steady-state distortions are studied before and after a decrease in firms’ ability to borrow. Qualitatively, we see that a decrease in firms’ ability to borrow still leads to a decrease in the interest rate and an amplification of the increase
in investment distortions. However, with a corporate Lucas tree, the general-equilibrium amplification effect is smaller; the decrease in TFP in general equilibrium is only 14 percent larger than it would be if the interest rate were constant, as in a small open economy.

Table 6. Decrease in firms’ ability to borrow: Steady-state distortions

|                       | $\theta = \theta_1$ | $\theta = \theta_2$ | Increase |
|-----------------------|----------------------|----------------------|----------|
| TFP losses            | 0.50                 | 2.19                 | 2.42     | 1.91     |
| E[marginal product of capital] - $r$ | 0.17                 | 0.83                 | 1.14     | 0.96     |
| Sd[marginal product of capital]     | 0.66                 | 1.83                 | 2.06     | 1.40     |
| Workers' consumption risk, $(1 - E_x)$ | 1.40                 | .                    | 1.74     | 0.33     |

Notes: Please see Table 3 or main text.

A full analysis should take into account differences in the productivity process, production function and borrowing capacity of different sectors of the economy, and of the types of financial frictions present in each. With this approach, one could study the TFP losses due to heterogeneity in firms’ ability to borrow, which are decreasing in the interest rate in the case of constant-productivity entrepreneurs. This would enable a comparison of the two sources of TFP losses in the model – differences in firms’ ability to borrow, and productivity shocks that endogenously result in some firms having high productivity and low wealth – and their relative contribution to steady-state distortions and general-equilibrium amplification.
Bibliography

[1] Aiyagari, S. Rao, 1994. “Uninsured Idiosyncratic Risk and Aggregate Saving,” Quarterly Journal of Economics, Vol. 109, No. 3, pp. 659-684.

[2] Albuquerque, Rui and Hugo A. Hopenhayn, 2004. “Optimal Lending Contracts and Firm Dynamics.” Review of Economic Studies, Vol. 71, pp. 285–315.

[3] Alvarez, Fernando and Urban J. Jermann, 2000. “Efficiency, Equilibrium, and Asset Pricing with Risk of Default,” Econometrica, Vol. 68, No. 4, pp. 1468-0262.

[4] Angeletos, George-Marios and Laurent-Emmanuel Calvet, 2006, “Idiosyncratic Production Risk, Growth, and the Business Cycle,” Journal of Monetary Economics, Vol. 53, No. 6, pp. 1095-1115.

[5] Angeletos, George-Marios and Vasia Panousi, 2010. “Financial Integration, Entrepreneurial Risk and Global Dynamics,” mimeo.

[6] Banerjee, Abhijit and Benjamin Moll, 2010. “Why does misallocation persist?,” American Economic Journal: Macroeconomics, forthcoming.

[7] Bernanke, Ben, Mark Gertler and Simon Gilchrist, 1999. “The financial accelerator in a quantitative business cycle framework,” in John B. Taylor and Michael Woodford, eds., Handbook of Macroeconomics, Vol. 1, Ch. 21, pp. 1341-1393.

[8] Bernanke, Ben and Mark Gertler, 1989. “Agency Costs, Net Worth, and Business Fluctuations,” The American Economic Review, Vol. 79, No. 1, pp. 14-31.

[9] Buera, Francisco and Yongseok Shin, 2010a. “Financial Frictions and the Persistence of History: A Quantitative Exploration,” mimeo.
[10] Buera, Francisco and Yongseok Shin, 2010b. “Productivity Growth and Capital Flows: The Dynamics of Reforms,” mimeo.

[11] Cagetti, Marco and Mariacristina De Nardi, 2006. “Entrepreneurship, Frictions, and Wealth,” Journal of Political Economy, 2006, vol. 114, no. 5, pp. 835-870.

[12] Clarida, Richard H., 1990, “International Lending and Borrowing in a Stochastic, Stationary Equilibrium,” International Economic Review, vol. 31, no. 3, pp. 543-558.

[13] Cooley, Thomas, Ramon Marimon, and Vincenzo Quadrini, 2004. “Aggregate Consequences of Limited Contract Enforceability,” The Journal of Political Economy, vol. 112, no. 4, pp. 817-847.

[14] Curtin, Richard T., F. Thomas Juster, and James N. Morgan, 1989. “Survey Estimates of Wealth: An Assessment of Quality,” in Lipsey, Robert E. and Helen Stone Tice, eds., The Measurement of Saving, Investment and Wealth, University of Chicago Press.

[15] Evans, David S. & Jovanovic, Boyan, 1989. “An estimated model of entrepreneurial choice under liquidity constraints,” Journal of Political Economy, Vol. 97, No. 4, pp. 808–27.

[16] Farhi, Emmanuel and Jean Tirole, 2010. “Bubbly Liquidity,” mimeo.

[17] Heaton, John and Deborah J. Lucas, 1996. “Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing,” Journal of Political Economy, Vol. 104, No. 3, pp. 443-487.

[18] Holmström, Bengt and Jean Tirole, 1998. “Private and Public Supply of Liquidity,” The Journal of Political Economy, Vol. 106, No. 1, pp. 1-40.

[19] Jorgenson, Dale W., 1963. “Capital Theory and Investment Behavior,” The American Economic Review, Papers and Proceedings, Vol. 53, No. 2, pp. 247-259.

[20] Kehoe, Timothy J. and David K. Levine, 1993. “Debt-Constrained Asset Markets,” The Review of Economic Studies, Vol. 60, No. 4, pp. 865-888.

[21] Krueger, Dirk and Fabrizio Perri, 2006. “Does Income Inequality Lead to Consumption Inequality? Evidence and Theory,” Review of Economic Studies, Vol. 73, No. 1, pp. 163–193.
[22] Lorenzoni, Guido, 2008. “Inefficient Credit Booms,” Review of Economic Studies, Vol. 75, No. 3, pp. 809-833.

[23] Lorenzoni, Guido and Veronica Guerrieri, 2010. “Credit Crises, Precautionary Savings and the Liquidity Trap,” mimeo.

[24] Lorenzoni, Guido and Karl Walentin, 2007. “Financial Frictions, Investment and Tobin’s q,” mimeo.

[25] Lucas, Robert E., 1978. “On the Size Distribution of Business Firms,” The Bell Journal of Economics, Vol. 9, No. 2, pp. 508-523.

[26] Lucas, Robert E., 1987. Models of Business Cycles. New York: Blackwell.

[27] Lucas, Robert E., 2003. “Macroeconomic Priorities,” mimeo., University of Chicago.

[28] Massa, Massimo and Andrei Simonov, 2006. “Hedging, Familiarity and Portfolio Choice,” The Review of Financial Studies, Vol. 19, No. 2, pp. 633-685.

[29] McGrattan, Ellen R. and Edward C. Prescott, 2010. “Unmeasured Investment and the Puzzling U.S. Boom in the 1990s,” Federal Reserve Bank of Minneapolis, Research Department Staff Report 369.

[30] Mendoza, Enrique G., Vincenzo Quadrini, and José-Víctor Ríos-Rull, 2009. “Financial Integration, Financial Development, and Global Imbalances,” Journal of Political Economy, Vol. 117, No. 3, pp. 371-416.

[31] Midrigan, Virgiliu and Daniel Yi Xu, 2010. “Finance and Misallocation: Evidence from Plant-Level Data,” NBER Working Paper 15647.

[32] Moll, Benjamin, 2010. “Productivity Losses from Financial Frictions: Can Self-Financing Undo Capital Misallocation?,” mimeo.

[33] Olley, G. Steven and Ariel Pakes, 1996. “The Dynamics of Productivity in the Telecommunications Equipment Industry,” Econometrica, Vol. 64, No. 6, pp. 1263-1297.

[34] Quadrini, Vincenzo, 2000. “Entrepreneurship, Saving and Social Mobility,” Review of Economic Dynamics, Vol. 3, No. 1, pp. 1-40.
[35] Rajan, Raghuram G. and Luigi Zingales, 2003. “The great reversals: the politics of financial development in the twentieth century,” Journal of Financial Economics, Vol. 69, No. 1, Tuck Symposium on Corporate Governance, pp. 5-50.

[36] Rampini, Adriano A. and S. Viswanathan, 2010a. “Collateral, Risk Management, and the Distribution of Debt Capacity,” Journal of Finance, Vol. 65, pp. 2293-2322.

[37] Rampini, Adriano A. and S. Viswanathan, 2010b. “Collateral and capital structure,” mimeo., Duke University.

[38] Reinhart, Carmen M. and Vincent R. Reinhart, “After the fall,” NBER Working Paper 16334.

[39] Storesletten, Kjetil, Chris I. Telmer and Amir Yaron, 2004. “Cyclical Dynamics in Idiosyncratic Labor Market Risk,” The Journal of Political Economy, Vol. 112, No. 3, pp. 695-717.
Chapter 2

Trading and advising on Wall Street

2.1 Introduction

On Wall Street, it is typical for a firm to trade with its own capital and to provide advice to clients. Combining these activities generates a conflict of interest: the firm may be tempted to mislead its clients and trade against them with the firm’s own capital. Indeed, Wall Street firms have recently been criticized for reported instances of trading contrary to their advice to clients.

The prevalence of firms that combine advising and proprietary trading raises questions about how and why these activities coexist within a firm, despite the possibility of malfeasance that comes from combining them. To explore these questions, this paper builds a repeated-game model with imperfect information in which firms choose whether to pursue proprietary trading, advising, or a combination of the two. Profiting from information by advising incurs lower capital costs than proprietary trading, but not all the firm’s information can be used by clients, and clients will only pay for what (in expectation) they will use. In choosing whether to engage in proprietary trading, advising, or both, a firm takes into account that a larger-proprietary trading capacity endogenously reduces the number of clients that the firm can credibly advise.

Understanding how and why firms combine proprietary trading and advising is an essential step toward evaluating recent policies meant to separate the two activities. The
recently enacted “Volcker rule” prohibits banks from engaging in proprietary trading and from making certain investments in hedge funds and private equity funds.\textsuperscript{1} Although a primary motivation for the Volcker rule was to reduce the potential for trading losses that would require a bailout using public funds, another primary motivation was to address conflicts of interest.\textsuperscript{2}

On Wall Street, it is typical for a firm to combine proprietary trading with an advisory brokerage business in which the fee paid by clients does not vary with the eventual accuracy of the advice. Although so-called Chinese walls are said to separate these activities, Wall Street firms typically combine advising and proprietary trading even at very fine levels of their organizational charts.\textsuperscript{3}

Taking this institutional structure as given, I model an informed trader that trades for its own account and sells information to clients. The model emphasizes that the informed trader is tempted to mislead its clients and trade against them. That is, when the information is positive, the informed trader can buy and tell the uninformed traders to sell, garnering a lower price for the informed trader.

I show that the amount of clients that the informed trader can credibly inform decreases in the proprietary-trading capacity of the informed trader. This is because the profits from lying are increasing in the volume the clients trade based on the informed trader’s report. Thus, given a certain punishment for lying, in order for the informed trader to prefer to tell the truth, the volume that clients trade based on the information must be less than a cutoff. And since the profits from lying are increasing in the capacity of the informed player to trade, this cutoff will be decreasing in the capacity of the informed player to trade. Thus, the informed trader can be “too big to believed.”

The limit on the informed trader’s proprietary-trading position might be due to: finite assets combined with a positive-wealth constraint; risk management; or other regulations or capacity constraints.

Of course, if the punishment for lying is sufficiently large, the clients need not worry that

\textsuperscript{1}This rule is contained in Section 619 of the Dodd-Frank Wall Street Reform and Consumer Protection Act, which was enacted on July 21, 2010.

\textsuperscript{2}Paul Volcker, after whom the rule was named, testified about “the strong conflicts of interest inherent in the participation of commercial banking organizations in proprietary or private investment activity.” Statement of Paul A. Volcker before the Committee on Banking, Housing and Urban Affairs, United States Senate, February 2, 2010. See Section 2 for more details.

\textsuperscript{3}For example, an individual trader sometimes trades both using the firm’s capital and for separate accounts owned by clients who have delegated capital to the trader. See Section 1.1 for more institutional detail.
the informed trader will lie. This paper features a repeated game in which the punishment for lying is limited because the informed trader and the clients cannot formally contract on the accuracy of the informed trader's information. Any performance bonus or penalty must be self-enforceable. Thus, the penalty for lying is limited to forgoing the profits from future cooperation.

In the model, it cannot be determined with certainty whether the informed trader lied and thus the informed trader sometimes is punished even when truthful (and sometimes goes unpunished when deviating from the equilibrium path by lying). This makes it harder to provide incentives to be truthful. I provide a closed-form solution to this imperfect-monitoring game and show that the optimal punishment rule is a cutoff rule, where the cutoff price is a function of the accuracy of the informed trader's advice. When the informed trader's signal is imperfect, an increase in noise trading results in a decrease in the amount traded by clients on the advice of the informed trader, resulting in a higher expected spread between the security's price and fundamental.

I show how the choice of business plan for the informed trader – whether the informed trader seeks to profit from its information through proprietary trading, advising or a mix of both – depends on the endogenous tension between proprietary trading and advising. In the model, clients cannot trade some securities, because regulations or their investment mandates prohibit it; for example, mutual funds and insurance companies face restrictions on what types of assets they can invest in. When an informed trader has information about a security that the clients cannot trade, the clients cannot profit from the information. When choosing its position limit, the informed trader faces the following tradeoff: (i) in periods of autarky (i.e., when the clients cannot trade on the information), profits are weakly increasing in the position limit; (ii) in periods when the clients can trade on the information, monopoly profits can be achieved with a low, or even zero, position limit and increasing the position limit could actually decrease profits. The tradeoff is made taking into account the cost of the position limit and, naturally, the probability that the clients can trade on the information. In trading off these two concerns, the informed trader may choose a position limit such that when clients can trade on its information, the informed trader could make more money if it had a lower position limit.

A firm's choice of its position limit is like a choice of a business plan and variation in the probability that clients can trade on its information determines whether: (i) the firm
goes it alone by trading only for its own book, giving up altogether on client fees; (ii) the
firm does not engage in any trading for its own book, earning all its income from selling
information; or (iii) the firm both trades for its own book and sells information. Firms that
trade exclusively for their own book resemble hedge funds. Firms that hold no positions
themselves and earn all their income from client fees resemble independent research firms
or mutual fund companies. Firms that do both resemble broker-dealer investment banks.

In the model, the parameter that determines a firm’s equilibrium choice of business plan is
the probability that the clients can trade on the informed trader’s information. Firms that
have information only about securities that clients can trade earn all their income from fees.
Firms that have information only about securities that clients cannot trade earn all their
income from proprietary trading. Nonetheless, and surprisingly, the equilibrium position
limit (i.e., business plan) is neither monotonic nor continuous in the probability that clients
can trade the securities about which the informed trader has information.

The key mechanism of the model is that a firm, in choosing its position limit, takes into
account that a larger position limit reduces its ability to advise clients. With uncertainty
about whether the informed trader lied, a larger proprietary-trading capacity makes it
easier for the clients to determine whether the informed trader lied; however, even with
this effect, the amount of clients that can be credibly informed still decreases with the
proprietary-trading capacity of the informed trader.

Studying proprietary trading and advising together has implications for how a security’s
liquidity affects the informativeness of its price. With perfect monitoring, it is easier to
communicate about more liquid securities, because the price impact of misled clients is
smaller for more liquid securities. With imperfect monitoring, the effect of illiquidity on
communication is a priori ambiguous. While the trading profits from lying are increasing
in a security’s illiquidity, increasing illiquidity also makes it easier to detect lying in an
environment with ex-post uncertainty about whether the informed trader lied. The reason
is that price is a better signal of the informed trader’s position when a security is more
illiquid.

The model also makes predictions about how informational efficiency is affected by
the trading capacities of informed traders, a topic that has recently seen renewed research
interest (Kyle and Xiong 2001, Gromb and Vayanos 2002, He and Krishnamurthy 2008,
Brunnermeier-Pedersen 2008, Mitchell-Pedersen-Pulvino 2007). In the model, an exogenous

80
decrease in the firm’s proprietary trading capacity (due to a crisis or an unrelated loss in another part of the firm’s business, for example) results in an increase in the spread between price and fundamental for assets that cannot be traded by clients. This result is standard and, in my model, very mechanical. What is new, however, is that the decrease in trading capacity results in a decrease in this spread for assets that can be traded by clients, if the firm’s trading capacity was already below a certain threshold. This is because a lower trading capacity allows the firm to credibly inform a greater amount of clients and, in the model, the resulting increase in trading volume is greater than the firm’s decrease in trading capacity.

The approach in this paper highlights the importance of studying proprietary trading and advising together. It also responds to a question raised by models of limited arbitrage, such as Shleifer-Vishny (1997), in which informed traders do not close the gap between prices and fundamentals because they have limited funds. One might ask whether investors in such a setting could profit more and further reduce the gap by selling their information to others. The answer provided in my model is that selling information is possible, but total trading by the informed trader and its clients on the informed trader’s information may still be less than trading by an unconstrained informed trader, since the possibility of lying limits communication.

In the remainder of this section I examine how trading and advising are actually combined on Wall Street and explain how my approach fits with the previous literature. Section 2 presents the model and Section 3 solves the model in the benchmark case in which clients can determine at the end of each period whether they have been misinformed. Section 4 solves the model in which there is uncertainty about whether the informed trader lied. Section 5 allows the informed trader to choose its proprietary-trading capacity before the trading game begins. Section 6 concludes.

2.1.1 How trading and advising are combined on Wall Street

On Wall Street, sophisticated financial players engage in both proprietary trading and selling advice, an observation that is a starting point of the model.

Concerns about the incentives of advisors to lie to clients and trade against them are central to the business plans of agency brokerages and independent research providers that do no proprietary trading and trumpet this point in their marketing. Argus Research, an
independent research firm, notes on its website that it does not “make markets in stocks or manage money....Because we’re not a broker/dealer, we don’t have ‘inventory’ of a stock that we have to move with a tainted recommendation.” Another example is Investment Technology Group, Inc. Barron’s said of ITG in 2006: “Because ITG doesn’t trade its own capital, it finds fans among clients who, despite all assurances of propriety and Chinese Walls, are wary of brokerages that bet vats of house money.” Broker-dealer and brokerage are, in this context, synonyms for investment banks. Recently, investors in a fund managed by legendary hedge-fund manager James Simons raised questions about why their fund was performing poorly at the same time that a fund held largely by Simons and his traders saw big gains.4

Many have expressed skepticism about the efficacy of Chinese Walls. Volcker testified that although not all conflicts of interest can be expunged from banking, “[I am not] so naïve as to think that, even with the best efforts of boards and management, so-called Chinese Walls can remain impermeable against the pressures to seek maximum profit and personal remuneration.”

There are several ways in which financial players such as investment banks mix advising and proprietary trading. Research analysts provide fundamentals-driven recommendations; sales and trading desks provide short-term recommendations and advice on order execution. At the same time, firms engage in trading for their own books via market-making and proprietary trading. A trader can trade for its firm’s own book and, separately, for discretionary accounts belonging to clients.5

The strategy of lying to clients in order to trade against them is clearly illegal. In particular, investors and the government can pursue private, administrative, and criminal causes of action and statutory penalties under Rule 10b-5, “Employment of Manipulative and Deceptive Devices,” promulgated under Section 10 of the Securities Exchange Act of 1934. Section 17 of the 1933 Securities Act may also apply in prosecuting lying to clients in order to trade against them. Moreover, broker/dealers face further regulations through the Financial Regulatory Authority (FINRA) and the National Association of Securities Dealers (NASD). Despite these laws, it appears that the “firewalls” separating advising

---

4 “Simons Questioned by Investors,” Wall Street Journal, May 15, 2009.
5 An example of this is mentioned in a Wall Street Journal article about trader Boaz Weinstein: “By early 2008, Mr. Weinstein was at the top of his game. He, along with a colleague in London, was overseeing global credit trading for all of Deutsche Bank...And his control also extended to the bank’s trading for customers.” Wall Street Journal, “Deutsche Bank Fallen Trader Left Behind $1.8 Billion Hole,” Feb. 6, 2009.
and trading for one’s book are not as impenetrable as those that separate advising and investment banking.⁶

Broker-dealer investment banks are keenly aware that their various businesses present potential conflicts of interest and that reputation can help overcome the incentive problems such conflicts present, but that reputation is an imperfect solution and one with its own risks. In its 10-K, a major investment bank notes: “Our reputation is one of our most important assets. As we have expanded the scope of our businesses and our client base, we increasingly have to address potential conflicts of interest....Appropriately identifying and dealing with conflicts of interest is complex and difficult, and our reputation could be damaged...if we fail, or appear to fail, to identify and deal appropriately with conflicts of interest.” The bank points out that “appearing” to fail is a problem; this is a distinct possibility in Section 4, where ex-post uncertainty about whether the informed trader lied results in punishments even when the informed trader is always truthful. In Section 5, this paper gives a sense of why investment banks mix proprietary trading and advising customers, despite the difficulties of dealing with conflict of interest.

2.1.2 Review of the literature

Market manipulation

The incentive of informed traders to dupe others in order to achieve a lower price when buying (and a higher price when selling) is well known. Benabou-Laroque (1992) cite an infamous example involving Nathan Rothschild.⁷

⁶The 2003 global research analyst settlement, for example, requires firms to “separate research and investment banking, including physical separation, completely separate reporting lines, separate legal and compliance staffs, and separate budgeting processes.” (Testimony of SEC chairman W. H. Donaldson, May 7, 2003, before the Senate Committee on Banking, Housing and Urban Affairs.) In contrast, prior to January 2009, the Financial Regulatory Authority (FINRA) ban on proprietary trading based on advance knowledge of recommendations to be made by the same firm’s research department applied only to exchange-listed securities or derivatives thereof, thereby excluding many debt instruments. Moreover, prior to January 2009, it was recommended – but not required – that firms “establish policies and procedures...to isolate specific information within research and other relevant departments so as to prevent the trading department from utilizing advance knowledge of the issuance of research reports.” FINRA Rule 5280 extends the ban on trading on advance knowledge to all securities and eliminates the option to establish policies and procedures to prevent such trading.

The independent research providers that the major firms have contracted for their clients as part of the 2003 global settlement are not required to be research-only shops; some, like Thomas White International, both manage money and provide research.

⁷From Benabou-Laroque (1992): “During the battle of Waterloo (1815) the banker Nathan Rothschild, who was known to have superior information from the continent due to a system of carrier pigeons, walked around the city looking dejected, spreading the news that the battle was going badly, and had his agents openly sell British government securities. Meanwhile, he was secretly buying much larger quantities of these securities, taking advantage of the depressed price and of his actual knowledge of an impending victory.”
Admati and Pfleiderer (1986) study an informed seller of information, but specifically prohibit it from trading, remarking, “If the seller traded, his incentives to reveal truthfully the information he promised would be severely distorted.”

Benabou-Laroque (1992) study an informed journalist who announces a signal. There is uncertainty about the journalist’s type; with some probability, the journalist always reveals his signal honestly. The opportunistic journalist exploits this uncertainty and in equilibrium the opportunistic type always lies with positive probability. The Benabou-Laroque setup does not readily admit the type of comparative statics analysis pursued in my paper, because the journalist is unconstrained in the size of his position, which is limited only by risk aversion.

Van Bommel (2003) provides a model of rumormongers closely related to Crawford and Sobel (1982). An informed trader who faces a position limit knows a security’s payoff, which comes from a continuous distribution. The trader takes a position and then spreads a binary rumor (“buy, I bought”, “sell, I sold”). Others then trade on the rumor, moving the price. When the payoff is small in absolute value, the price overshoots, and the informed trader profits by trading against the overshooting. When the informed trader is allowed to lie, rumormongering is not possible in one-shot game, but is possible in a repeated game for a sufficiently high discount factor. Thus, the model explains how privately communicated rumors, such as Internet rumors, can contain information.

**Limits to arbitrage**

Brunnermeier-Pedersen (2008) and Kondo-Papanikolaou (2005) also study limited arbitrage when an informed trader faces position limits. Brunnermeier-Pedersen (2008) examines the link between market liquidity and funding liquidity.

This paper is closely related to Kondo-Papanikolaou (2005). In Kondo-Papanikolaou (2005), an informed trader faces a position limit. The informed trader can share its information with a bank in order to increase its position limit, but the bank will then have the incentive to trade on the informed trader’s information for its own account, reducing the informed trader’s profit. Kondo-Papanikolaou study efficient subgame perfect equilibria in this setting.

There are two key differences between this paper and Kondo-Papanikolaou (2005). First, what limits communication in Kondo-Papanikolaou is the possibility of front-running
by the uninformed. In contrast, in this paper, what limits communication is the possibility that the informed can mislead the uninformed. Also, Kondo-Papanikolaou addresses only settings with perfect information about deviations, whereas this paper provides a closed form solution for the game with imperfect information.

**Market microstructure with an informed trader**

Kyle (1985) and Kyle (1989) study how an informed trader optimally exploits its information by revealing it slowly over time. The informed trader trades only for its own account; it cannot attempt to sell its information. In my model, the trading of a given asset is one-shot, rather than repeated over time, so the informed trader does not have an incentive to reveal his information slowly.

**Optimal oligopoly and relational contracting**

In my model, the informed trader and its clients cannot contract formally on the accuracy of the informed trader’s advice. Instead, they make payments to each other and any performance-contingent fee must be self-enforcing. This draws on the literature on relational contracting, pioneered by Baker-Gibbons-Murphy (1994), Baker-Gibbons-Murphy (2002) and Levin (2003).

The search for the optimal subgame perfect equilibrium in my model bears some similarities with papers on optimal oligopoly, especially Rotemberg-Saloner (1986), in which collusive quantities are set such that neither party wants to deviate in equilibrium.

### 2.2 The model

Time $t \in \{0, 1, \ldots\}$ is infinite and discrete. There is an informed trader, also called the bank. At $t = 0, 1, \ldots$ the informed trader engages in a repeated game with its clients.

Each period, $t = 0, 1, \ldots$, a new asset is traded. At the end of the period, the asset pays a random dividend $\theta \in \{-\beta, \beta\}$ and then becomes worthless. The timing of the stage game is:

1. The informed trader proposes that the clients pay a total fee of $w$
2. The informed trader receives a signal $s \in \{\beta, -\beta\}$ about $\theta$ that is correct with probability $\alpha > \frac{1}{2}$
3. The informed trader makes a report $\hat{\theta}$ to the clients
4. The informed trader submits an order $d^b$, $|d^b| \leq x$, and the clients submit orders totalling $d^c$, $|d^c| \leq y$

5. Markets clear

6. The asset pays out $\theta$

7. The informed trader chooses a voluntary payment $w^- \geq 0$ to the clients

The net supply of the asset is zero. There is a third group of investors that submits demand $d^l = -\frac{p - \varepsilon}{\sigma}$ with $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Thus, markets clear at price $p = \sigma(d^b + d^c) + \varepsilon$. Investors with downward-sloping demand curve are standard in the literature (Shleifer-Vishny (1997), Xiong (2001), Brunnermeier-Pedersen (2005)). One can think of them as investors needing liquidity or unsophisticated investors. The $\varepsilon$ term can be due to noise traders or stochastic liquidity needs. Sigma is a measure of the illiquidity of the market when one considers liquidity to be price insensitivity to a trader taking a large position.

The informed trader and the clients are risk neutral with discount factor $\delta < 1$.

The payoff for the bank is $w + d^b(\theta - p(d^b, d^c)) - w^-$. The total payoff for the clients is $-w + d^c(\theta - p(d^b, d^c)) + w^-$. The bank’s offer at Step 1 is a take-it-or-leave-it offer. The clients respond immediately.

The clients are atomistic and do not take into account the effect of their trades on the price of the asset.

The orders are submitted simultaneously in Step 4. The informed trader faces a constraint on the position he can take, $|d^b| \leq x$. The total client position is similarly limited, $|d^c| \leq y$. $y$ is a choice variable of the bank, corresponding to its mass of clients. One can assume $y$ is chosen before period $t = 0$ begins. Trading games with position limits and market orders are also found in Brunnermeier-Pedersen (2005), Van Bommel (2003) and Kondo-Papanikolaou (2005).

The informed trader provides the same report $\hat{\theta}$ to all of its clients.

### 2.2.1 Autarky

Suppose the informed trader has no clients. This is autarky for the informed trader. Suppose further that $x = +\infty$. That is, the informed trader faces no position limit. The
informed trader solves

\[
d^m(s) = \arg \max E[d_b(\theta - \sigma d^m - \varepsilon)|s]
\]
\[
\Rightarrow d^m(s) = \frac{E[\theta|s]}{2\sigma}.
\]

This position \(d^m\) taken by the unconstrained informed trader in autarky is called the monopoly position, because the problem (2.1) resembles the choice of a monopolist facing a downward-sloping demand curve. Call \(E[d^m(s)(\theta - \sigma d^m(s) - \varepsilon)|s]\) the monopoly profit.

Now suppose that \(x\) is finite. When the informed trader’s signal is \(s = \beta\), the trader’s optimal order is \(d^b = \min\{\frac{E[\theta|s=\beta]}{2\sigma}, x\}\). The \(-\beta\) case is symmetric, with \(d^b = -\min\{\frac{E[\theta|s=\beta]}{2\sigma}, x\}\). When \(x \leq \frac{E[\theta|s=\beta]}{2\sigma}\), the bank’s position limit is binding. In this case, the bank could benefit from “selling” his information to the clients.

### 2.2.2 Relational contracting

A relational contract is a complete plan for the relationship.

Let the public history \(h^t = (y, w_0, p_0, \theta_0, \widehat{\theta}_0, w_{t-1}, p_{t-1}, \theta_{t-1}, \widehat{\theta}_{t-1}, w_{t-1})\).

Let the history of the informed trader’s signals be \(s^t = (s_0, ..., s_{t-1})\). For each date \(t\) and history, a relational contract is:

1. a bank proposal \(w : h^t \rightarrow \mathbb{R}^+\);
2. whether the clients accept or decline the proposal \(A : h^t, w \rightarrow [0, 1]\);
3. a reporting strategy for the bank \(\widehat{\theta} : h^t, w, A, s \rightarrow \{-\beta, \beta\}\);
4. orders \(d^b : h^t, w, A, s, \widehat{\theta} \rightarrow [-x, x]\) and \(d^c : h^t, w, A, \widehat{\theta} \rightarrow [-y, y]\);
5. a voluntary payment \(w^- : h^t, w, A, \widehat{\theta}, \theta, p \rightarrow \mathbb{R}^+\).

Client optimization requires the clients to accept the contract if and only if the net expected payoff (trading profits less fees) is greater than or equal to their outside option. Their outside option is the profits they could earn without receiving a signal from the informed trader, which turn out to be zero. Since clients do not take into account the effect of their trading on the price, the clients buy \(y\) units [sell \(y\) units] if the expected trading profits conditional on its information is positive [negative].

A relational contract is considered self-enforcing if it is a PBE. This paper will find the PBE with the highest expected payoff to the informed trader. It will always be the case that this PBE involves the informed trader truthfully reporting its signal to a (weakly)
positive client mass $y$ along the equilibrium path. Thus, along the equilibrium path, if the informed trader’s clients have positive mass, the informed trader must find it unattractive to mislead her clients and trade against them by, for example, buying for her own account while telling the clients to sell.

When $\alpha = 1$ or $\sigma^2_\varepsilon = 0$, any profitable deviation (i.e., when the signal is positive, buy and tell the clients to sell) is perfectly detected after the market-clearing price $p$ and the fundamental $\theta$ are revealed. If $\alpha = 1$, and the report of the signal differs from the realized fundamental, then the informed trader lied about its signal. If $\sigma^2_\varepsilon = 0$, then the clients can infer the informed trader’s order from the price and thus determine if the informed trader’s report and order do not correspond. Thus, I first analyze equilibrium in a benchmark setting in which either $\alpha = 1$ or $\sigma^2_\varepsilon = 0$ hold. I then relax the assumptions of the model by allowing $\alpha < 1$ and $\sigma^2_\varepsilon > 0$. In this general case, the informed trader will be punished along the equilibrium path, complicating the analysis.

Along the equilibrium path, the clients simply follow the reports of the informed trader. Thus, this could also be considered a model of delegation, in which the client authorizes the informed trader to trade up to a certain number of shares and the client can withdraw this delegation each period. On Wall Street, these are called discretionary accounts.

**Remark.** The fee $w$ corresponds to a commission or an advisory fee. The voluntary payment $w^-$ reflects that a bank and its clients are allowed to exchange money (unlike oligopolists, for example). Allowing the voluntary payment means the continuation game beginning at every node along the equilibrium path is efficient, in the sense that trading profits are not foregone in order to punish the informed trader. Rather, punishments that deter the informed trader from lying can take the form of a cash transfer from the informed trader to the clients; the informed trader is willing to make this cash transfer because of the threat of foregone profits. Instead of destroying surplus, surplus is transferred, as in Levin (2003). This simplifies the analysis in the general case when $\alpha < 1$ and $\sigma^2_\varepsilon > 0$. In Section 4, I derive a condition on the parameters such that $w^-$ is less than $w$. When this condition holds, the $w^-$ payment can be considered a discount for the clients on next period’s fee; no money needs to change hands. When $\alpha = 1$ or $\sigma^2_\varepsilon = 0$, allowing for $w^-$ is irrelevant. In this case, the voluntary payment could be disallowed (i.e., eliminate Step 5) without changing the equilibrium prices, quantities or profits; punishments would be achieved by reverting to autarky. Payments to clients who have followed bad advice do happen on Wall Street, but
it is hard to disentangle reputational and legal concerns for such payments.\(^8\)

**Remark.** It is essential for the model that clients move the price through their trading. The 10-K of a major investment bank notes that its Trading and Principal Investments unit serves “a diverse group of corporations, … financial institutions, investment funds, governments and individuals.” The model envisions these types of clients (i.e., pension funds) as sufficiently massive to move prices.

2.3 **Benchmark Case: \(\alpha = 1\) or \(\sigma^2_\varepsilon = 0\)**

The stationary PBE will feature policy functions \(d^b(\theta, \hat{\theta})\) and \(d^c(\hat{\theta})\). I will look for an equilibrium in symmetric policy functions, in the following sense: \(d^b(\beta, \beta) = -d^b(-\beta, -\beta)\) and \(d^c(\beta) = -d^c(-\beta)\). Given the symmetry of the problem, this is without loss of generality.

In an abuse of notation, denote by \(d^b\) the position \(d^b(\beta, \beta)\) taken by an informed agent who receives a signal \(s = \beta\) and honestly reports it. Likewise, denote: \(d^b(-\beta, -\beta)\) by \(-d^b\); \(d^c(\beta)\) by \(d^c\); and \(d^c(-\beta)\) by \(-d^c\).

Exploiting the symmetry of the problem, the informed trader’s problem becomes:

\[
\max_{d^b, d^c, w, w^-} w + E[\Pi(d^b, d^c)|s = \beta]
\]

subject to

\[
\max_{|d^c| \leq x} E[\Pi(d^b, -d^c)|s = \beta] - w^- \leq E[\Pi(d^b, d^c)|s = \beta] \tag{2.2}
\]

\[-w^- + \frac{\delta}{1 - \delta} (w + E[\Pi(d^b, d^c)|s = \beta]) \geq \frac{\delta}{1 - \delta} \max_{|d^b| \leq x} E[\Pi(d^b, 0)|s = \beta] \tag{2.3}
\]

\[w \leq E[\Pi(d^c, d^b)|s = \beta] \tag{2.4}\]

\[|d^b| \leq x \tag{2.5}\]

\[d^b \in \arg\max_{|d^b| \leq x} E[\Pi(d^b, d^c)|s = \beta] \tag{2.6}\]

\(^8\)For example, in the Madoff case, several investment houses that recommended Madoff to clients provided some re-imbursement to those clients for losses. The National Bank of Kuwait, one of the largest banks in the Arab world, repaid principal and fictional gains, while Spanish banking giant Santander and Union Bancaire Privée (UBP), the second-biggest investor in hedge funds, offered partial compensation to clients. Smaller institutions, such as Celfin Capital (Chile) and Safra Group (Brazil), also are said to be paying compensation. Santander’s CFO “said the bank and its shareholders would be ‘better off’ in the long term paying compensation because they would be ‘keeping clients with the bank.’” (New York Times, January 28, 2009).
where

\[ \Pi(q_1, q_2) = q_1(\theta - p(q_1, q_2)) \]
\[ p(q_1, q_2) = \sigma(q_1 + q_2) + \varepsilon \]

Constraint (2.2) is the IC constraint; the informed trader must prefer to tell the truth and receive expected trading profits \( E[\Pi(d^b, d^c)|s = \beta] \), rather than lie and receive expected trading profits \( \max_{|d^b_{\text{lie}}| \leq x} E[\Pi^b(d^b_{\text{lie}}, -d^c)|s = \beta] \) less the penalty \( w^- \). When the informed trader lies and pays a penalty, the informed trader and the clients continue cooperating in the next period, so continuation values do not enter this constraint. Constraint (2.3) is the self-enforcement constraint; when the informed trader is supposed to pay a penalty \( w^- \), it must be in his interest to do so in order to continue in the cooperative regime. Constraint (2.4) is the individual rationality constraint for the clients; the fee they are willing to pay is bounded above by their expected trading profit. Constraint (2.5) is the informed trader’s position limit. Constraint (2.6) requires that the informed trader’s order after truthfully reporting his signal maximizes his trading profit in that period. The only role of Constraint (2.6) is to pin down a unique equilibrium when the informer trader’s position is such that his total profit from trading and fees equals the monopoly profit. This point will be discussed further below. Constraint (2.6) also highlights that the central difficulty here is creating incentives for truthful communication, rather than creating incentives for the traders to collude on quantities to preserve profit margins.

**Lemma 20** The informed trader’s problem can be simplified to:

\[
\max_{d^b, d^c} E[\Pi(d^b + d^c, 0)|s = \beta] 
\] (2.7)

subject to

\[
\Pi_{\text{tie}} + \frac{\delta}{1 - \delta} \Pi_{\text{aut}} \leq E[\Pi(d^b, d^c) + \frac{\delta}{1 - \delta} \Pi(d^b + d^c, 0)|s = \beta] 
\] (2.8)

\[ |d^b| \leq x \]
\[ d^b \in \arg \max_{|d| \leq x} \Pi(d, d^c). \]
where

\[
\Pi_{\text{tie}} = \max_{|d_{\text{tie}}| \leq x} E[\Pi(d_{\text{tie}}, -d^c)|s = \beta] \\
= E[\Pi(x, -d^c)|s = \beta] \text{ for } x \leq d^m(\beta), d^c \geq 0
\]

and

\[
\Pi_{\text{aut}} = \max_{|d_{\text{aut}}| \leq x} E[\Pi(d_{\text{aut}}, 0)|s = \beta] \\
= E[\Pi(x, 0)|s = \beta] \text{ for } x \leq d^m(\beta)
\]

**Proof.** The IR constraint (2.4) must be binding. Suppose it were not. Then increasing \(w\) would increase the maximand, relax constraint (2.3) and leave the other constraints unaffected. Thus we substitute \(w = \Pi(d^c, d^b)\) in the problem. Constraint (2.3) becomes:

\[-w^- + \frac{\delta}{1 - \delta} E[\Pi(d^b + d^c, 0)|s = \beta] \geq \frac{\delta}{1 - \delta} \Pi_{\text{aut}}\]

This constraint together with (2.2) imply (2.8). Conversely, if \((d^b, d^c)\) satisfy (2.8), this constraint and (2.2) are satisfied by setting \(w^- = \frac{\delta}{1 - \delta}(\Pi(d^b + d^c, 0) - \Pi_{\text{aut}})\).  

If \((d^b, d^c)\) satisfy (2.8), then \((d^b, d^c)\) can be implemented in a simple way that leaves no rents to the clients: the clients pay \(E[\Pi(d^c, d^b)|s = \beta]\) up-front and the bank pays a penalty \(w^- = \frac{\delta}{1 - \delta}(\Pi(d^b + d^c, 0) - \Pi_{\text{aut}})\) if it is caught lying. Indeed, in the benchmark case, in which there is no ex-post uncertainty about whether the bank lied, the problem with the possibility of paying a penalty is equivalent to the problem without this possibility, in which lying is punished by a permanent reversion to autarky.

**Proposition 21** The unique solution to the informed trader’s problem is:

\[
d^b = \begin{cases} 
  x, & \text{if } x < \frac{E[\theta|s=\beta]}{2\sigma} \\
  \frac{E[\theta|s=\beta]}{2\sigma}, & \text{if } x \geq \frac{E[\theta|s=\beta]}{2\sigma}
\end{cases}
\]

\[
d^c = \begin{cases} 
  \frac{E[\theta|s=\beta]}{2\sigma} - x, & \text{if } x < \frac{E[\theta|s=\beta]}{2\sigma} \frac{\delta}{2 - \delta} \\
  \max\left\{\frac{E[\theta|s=\beta]}{\sigma} - 2\frac{x}{\delta}, 0\right\}, & \text{if } x \geq \frac{E[\theta|s=\beta]}{2\sigma} \frac{\delta}{2 - \delta}
\end{cases}
\]

The amount of clients that the informed trader can credibly inform, \(\max\left\{\frac{E[\theta|s=\beta]}{\sigma} - 2\frac{x}{\delta}, 0\right\}\), is decreasing in \(x\) for \(x \leq \frac{E[\theta|s=\beta]}{2\sigma} \delta\) and is zero for \(x\) greater than this threshold.
Total trading by the informed trader and its clients, $d^b + d^c$, is non-monotonic in the informed trader’s position limit $x$. For $x \in (E[\theta|s=\beta] \frac{\delta}{2-\delta} \frac{\delta}{2-\delta}, \frac{E[\theta|s=\beta]}{2\sigma})$, total trading is less than the monopoly position; that is, $d^b + d^c < d^b_m$. For $x \in (\frac{E[\theta|s=\beta]}{2\sigma} \frac{\delta}{2-\delta}, \delta \frac{E[\theta|s=\beta]}{2\sigma})$, total trading and total informed-trader profits are decreasing in $x$; (ii) the gap between the expected fundamental and the price is increasing in $x$.

**Proof.** Consider the problem when $x < E[\theta|s=\beta] \frac{\delta}{2-\delta}$. Ignore constraint (2.6) momentarily.

If $d^b + d^c < E[\theta|s=\beta] \frac{\delta}{2-\delta}$, it must be that $d^b = x$. Otherwise, increasing $d^b$ would increase the maximand and slacken constraint (2.8).

It cannot be that $d^b + d^c > E[\theta|s=\beta] \frac{\delta}{2-\delta}$. If this were true, we could decrease $d^c$, which would increase the maximand and slacken constraint (2.8).

Suppose $d^b + d^c = E[\theta|s=\beta] \frac{\delta}{2-\delta}$. If $d^b < x$, then the tuple $(\tilde{d}^b, \tilde{d}^c) = (x, E[\theta|s=\beta] \frac{\delta}{2-\delta} - x)$ is also a solution with the same payoff.

Thus, in seeking the maximum feasible payoff, it is without loss of generality to set $d^b = x$. Constraint (2.8) becomes

$$d^c \leq \frac{E[\theta|s=\beta]}{\sigma \delta} - \frac{2x}{\delta}$$

The solution follows immediately. Constraint (2.6) is satisfied if one of the following two conditions holds: (i) $d^b = E[\theta|s=\beta] \frac{\delta}{2-\delta} - d^c$; or (ii) $E[\theta|s=\beta] - \sigma(2x - d^c) > 0$ and $d^b = x$. Constraint (2.6) is satisfied at the proposed solution. For $x \geq E[\theta|s=\beta] \frac{\delta}{2-\delta}$, the proposed solution is the unique solution to the problem without (2.6). For $x < E[\theta|s=\beta] \frac{\delta}{2-\delta}$, there are other possible solutions to the problem without (2.6), but these are ruled out by (2.6).

When $x \geq E[\theta|s=\beta] \frac{\delta}{2\sigma}$, the proposed solution achieves the unconstrained upper bound and is unique. $

**Corollary 22 (Folk Theorem)** For any $x$, there exists a $\delta < 1$ such that, for $\delta > \delta$, the monopoly profit can be achieved.

Figure 1 shows the equilibrium total demand when $\alpha = 1$ and $s = \beta$ as a function of the informed trader’s position limit. For low levels of $x$, $x < \frac{\beta}{2\sigma} \frac{\delta}{2-\delta}$, condition (2.8) is slack, and so $d^b + d^c$ can be set to the full monopoly level. I call this the full cooperation region. For intermediate levels of $x$, with $x \in (\frac{\beta}{2\sigma} \frac{\delta}{2-\delta}, \frac{\beta}{2\sigma} \frac{2\delta}{2-\delta})$, the amount of cooperation is
positive, but constrained. For these intermediate values of $x$, the informed trader cannot set $d^b + d^c$ to the full monopoly level without violating condition (2.8). Thus, in this region, the informed trader proposes a value of $d^c$ greater than zero but lower than $d^m(\beta) - x$ and the informed trader’s total profit from trading and fees is less than the monopoly profit. I call this the constrained cooperation region. Finally, for $x \in (\frac{\beta}{2\sigma}, \frac{\beta}{2\sigma})$, no cooperation is possible; the informed trader is forced into autarky, even though in autarky it is unable to fully exploit its information due to its position constraint. I call this the constrained autarky region. For $x \geq \frac{\beta}{2\sigma}$, the bank can fully exploit its information in autarky and communication is irrelevant. I call this the unconstrained autarky region.

The expected price equals total demand $d^b + d^c$ multiplicatively scaled by $\sigma$. Also, total informed trader profits are increasing in the total quantity traded when the total quantity traded is less than the monopoly position. Thus, Figure 1 is indicative of how prices and profits vary with the informed trader’s position limit.

The role of constraint (2.6) can be seen in Figure 1. In the full cooperation region, without constraint (2.6), there would be a multiplicity of equilibria which all featured total trading equal to the monopoly position, but which divided the trading among the informed trader and the clients differently. In all of these equilibria, total trading and total informed trader profits are the same. Constraint (2.6) seems a natural way of eliminating this multiplicity. Constraint (2.6) plays no role in the constrained cooperation or autarky regions.

The existence of the constrained cooperation region and the constrained autarky region are related to the following observation: the maximum client position consistent with the informed trader telling the truth declines more than one-for-one with the informed trader’s position limit. One intuition for this observation is that, in the model, autarky is less threatening to an informed trader with a higher position limit. However, dependence of the penalty for lying on the position limit is not needed to achieve price being non-monotonic in the position limit; even in a one-shot game with an exogenous punishment for lying, there is an interval of informed trader position limits where the price is declining in the position limit.\textsuperscript{9} Moreover, price declining in the informed trader’s position limit does not depend on the linearity of the long-term investors’ demand curve.

\textsuperscript{9}This holds so long as the exogenous punishment is not so large as to make cooperation (i.e., $d^c > 0$) impossible for all $x > 0$.  

93
Because of risk neutrality, Figure 1 is correct for arbitrary $\sigma^2 < \infty$.

The notion that profits could be declining in the informed trader’s capacity to trade is perhaps surprising. One way to make sense of it is to note that illiquidity $\sigma$ and position limit $x$ jointly determine incentives to lie and thus sustainable cooperation levels. Figure 2 divides $x\sigma$ space into four regions. In the unconstrained cooperation (UC) region, the incentive to lie is sufficiently small to allow the bank and client to achieve the unconstrained monopolist price. In the constrained cooperation (CC) region, some cooperation is possible, but the bank and the client are constrained by condition (2.8). In the constrained autarky (CA) region, no cooperation is possible and the bank has insufficient assets to fully exploit its information on its own. In the unconstrained autarky (UA) region, the bank can fully exploit its information on its own. The notion that cooperation is more difficult in illiquid markets is intuitive; in more illiquid markets, the price impact of misled clients trading in the wrong direction is greater, and thus so is the incentive to lie. The role of illiquidity, however, will be complicated in the section with ex-post uncertainty about whether in informed trader lied.

2.3.1 When clients cannot trade on the informed trader’s information

One element of realism missing thus far is the possibility that the clients are unable to trade on the informed trader’s information. This will play a central role in Section 5. There are several reasons why the clients may not be able to trade on the informed trader’s information. For one, clients may not be allowed to trade the security about which the
informed trader has information. Some investment vehicles are prevented, by law or by charter, from investing in certain classes of securities. For example, the investments of mutual funds are limited by the Investment Company Act of 1940. Insurance companies and pension funds face legal restrictions on their investments. Moreover, many investment vehicles are bound by their prospectuses not to invest, or not to invest more than a certain percentage of assets, in asset classes such as junk bonds, derivatives, micro-cap stocks, etc. Another reason clients may not be able to trade on the informed trader’s information is lack of liquidity. One way to model the fact that clients may be unable to trade on the informed trader’s information is to suppose that each period, a deterministic fraction of the clients are unable to trade the security. This would not affect the model’s solution, except to increase the total mass of clients so that each period the total mass who can trade on the security is equal to $d^c$. This, however, does not accord with the fact that some securities can be traded by a smaller share of investors (or that investors tend to face liquidity constraints at the same time other investors face them). A simple way to model this is to assume that, each period, there is a probability $\eta$ that clients can trade on the informed trader’s information. The above analysis implicitly assumed $\eta = 1$. The solution for $\eta < 1$ is straightforward.
Figure 3. Total trading as a function of the informed trader’s position limit, for \( \eta = 0.1 \) and \( \eta = 0.8 \).

**Proposition 23** The unique solution to the informed trader’s problem with \( \eta \in (0, 1] \) is:

\[
\begin{align*}
\theta^b &= \begin{cases} 
  x & \text{if } x < \frac{E[\theta|s=\beta]}{2\sigma} \\
  \frac{E[\theta|s=\beta]}{2\sigma} & \text{if } x \geq \frac{E[\theta|s=\beta]}{2\sigma}
\end{cases} \\
\theta^c &= \begin{cases} 
  \frac{E[\theta|s=\beta]}{\sigma} - x & \text{if } x < \frac{E[\theta|s=\beta]}{2\sigma} - \frac{\delta \eta}{2-2\eta + \delta \eta} \\
  \max\left\{ \frac{E[\theta|s=\beta]}{\sigma} - 2(1 - \frac{1}{\eta^2})x, 0 \right\} & \text{if } x \geq \frac{E[\theta|s=\beta]}{2\sigma} - \frac{\delta \eta}{2-2\eta + \delta \eta}
\end{cases}
\end{align*}
\]

**Proof.** Analogous to Proposition 21. ■

Substituting \( \eta = 1 \) reproduces the solution in Proposition 21. When \( \eta < 1 \), it is harder to provide incentives for truthful communication, since even if the informed trader is truthful, the informed trader will sometimes be in autarky when the clients cannot trade on the information and thus his fee income will be lower. This is illustrated in Figure 3, which shows total trading as a function of the position limit for two different values of \( \eta \).

The possibility that \( \eta < 1 \) will be returned to in Section 5, where it will play an essential role. In Section 4, it will be assumed that \( \eta = 1 \) to simplify the results, but this is not essential.
2.4 The general model with $\alpha < 1$ and $\sigma_\varepsilon > 0$

When $\alpha < 1$, the bank’s signal can differ from the realized fundamental, and thus $\theta \neq \hat{\theta}$ does not necessarily imply that the informed trader lied. And when $\sigma_\varepsilon > 0$, the clients cannot use the price to back out the informed trader’s order with certainty.

Thus, when $\alpha < 1$ and $\sigma_\varepsilon > 0$, if there is communication in equilibrium, the informed trader will sometimes be punished along the equilibrium path. In trying to implement a given $(d^b, d^c)$, the informed trader designing the contract would like to punish lying as much as possible and avoid punishing truthful reporting. This is difficult because $\theta = \beta, \hat{\theta} = -\beta, p = z$ can be the result of either: (i) truth-telling, with $(s = -\beta; \varepsilon = z + \sigma(d^b + d^c); \theta \neq s)$; or (ii) lying, with $(s = \beta; \varepsilon = z - \sigma(d^b_{tie} - d^c); \theta = s)$. I write $w^-(\theta, \hat{\theta}, p)$ to make explicit the dependence of the penalty $w^-$ on the realized $\theta$, the report $\hat{\theta}$ and the price $p$. The price serves as a noisy signal of the informed trader’s position and thus it may be optimal to condition the penalty on price.

Define

$$\bar{V} = w + \frac{1}{2} E[\Pi(d^b, d^c) - w^-(\theta, \beta, p(d^b, d^c))|s = \beta] + \frac{1}{2} E[\Pi(-d^b, -d^c) - w^-(\theta, -\beta, p(-d^b, -d^c))|s = -\beta]$$

$\bar{V}$ is the informed trader’s per-period expected total profits when truth-telling.

The incentive-compatibility constraints for $s = \beta$ and $s = -\beta$ are:

$$\max_{|d^b_{tie}| \leq x} E[\Pi(d^b_{tie}, -d^c) - w^-(\theta, -\beta, p(d^b_{tie}, -d^c))|s = \beta] \leq E[\Pi(d^b, d^c) - w^-(\theta, \beta, p(d^b, d^c))|s = \beta]$$

(2.12)

$$\max_{|d^b_{tie}| \leq x} E[\Pi(d^b_{tie}, d^c) - w^-(\theta, \beta, p(d^b_{tie}, d^c))|s = -\beta] \leq E[\Pi(-d^b, -d^c) - w^-(\theta, -\beta, p(-d^b, -d^c))|s = -\beta]$$

(2.13)

The individual-rationality constraint for the clients is:

$$\frac{1}{2} E[\Pi(d^c, d^b) + w^-(\theta, \beta, p(d^b, d^c))|s = \beta] + \frac{1}{2} E[\Pi(-d^c, -d^b) + w^-(\theta, -\beta, p(-d^b, -d^c))|s = -\beta] \leq w$$

(2.14)
The self-enforcement constraints are:

\[ 0 \leq w^-(\theta, s, p) \leq \frac{\delta}{1 - \delta} (\bar{V} - \Pi_{aut}) \quad \forall \theta, s, p \] (2.15)

The informed trader’s problem is

\[ \max_{d^b, d^c, w^-} \bar{V} \]

subject to (2.12)-(2.15), as well as the position limit constraint (2.5) and the optimal-trading constraint (2.6). The problem, as written, already takes advantage of the symmetry of the informed trader’s problem.

The penalty function \( w^- (\cdot, \hat{\theta}, \cdot) \) appears on the left-hand-side of the \( s = -\theta \) IC constraint, while also appearing on the right-hand-side of the \( s = \theta \) IC constraint, for \( \theta \in \{-\beta, \beta\} \). Thus, for example, increasing \( w^- (\beta, -\beta, p) \) makes lying less attractive to an informed trader with signal \( s = \beta \), but it makes truth-telling less attractive to an informed trader with signal \( s = -\beta \). Also, the informed trader’s optimal order when lying, \( d^b_{lie} \), now takes into account both his short-term trading profits and the probability of having to pay a penalty. When lying was perfectly detectable (Section 3), an informed trader’s optimal deviation would set \( d^b_{lie} \) purely to maximize short-term trading profits, since his lie would be detected with certainty. In an environment in which lying is not perfectly detected and in which the penalty depends on the price, an informed trader who lies might forgo some short-term trading profits in order to reduce the expected penalty he will have to pay.

As in Section 3, the client individual-rationality constraint (2.14) will bind. Thus, we can replace \( \bar{V} = E[\Pi(d^b + d^c, 0)|s = \beta] \) in the informed trader’s problem.

The following Lemma says that ex-post uncertainty makes the informed trader weakly worse off, relative to the benchmark case, even holding constant the value of the information in autarky.

**Proposition 24** (i) Define

\[ \bar{\beta} = E[\theta|s = \beta] \]

The informed trader’s expected payoff and the amount traded by the clients are weakly less than the expected payoff and amount traded in the benchmark problem without ex-post uncertainty with parameters \( \beta = \bar{\beta} \) and \( \alpha = 1 \). This statement holds strictly when \( x \) is such...
that the clients trade a positive amount, but not enough to achieve the monopoly profit, in
the benchmark problem with parameters $\beta = \tilde{\beta}$ and $\alpha = 1$.

(ii) Define

$$\tilde{\delta} = \frac{\delta(2\alpha - 1)}{1 - \delta + \delta(2\alpha - 1)}$$

The informed trader’s expected per-period payoff and the amount traded by the clients
are bounded below by the expected per-period payoff and amount traded in the problem with
no ex-post uncertainty with parameters $\delta = \tilde{\delta}$, $\beta = \tilde{\beta}$ and $\alpha = 1$.

(iii) The solution to the problem with ex-post uncertainty is full cooperation, attaining
the monopoly profit, for $x \leq x^{L}$ and constrained autarky, attaining the autarky profit, for
$x \geq \delta \frac{E[|\theta| s = \beta]}{2\sigma}$, where

$$x^{L} = \frac{E[|\theta| s = \beta]}{1 - \delta - 2\alpha - 1} + 1$$

Proof of Proposition 24. (i) The problem with ex-post uncertainty with $\alpha < 1$ and
the benchmark problem with $\beta = \tilde{\beta}$ and $\alpha = 1$ are equivalent except for the IC constraints.
Any pair $(d^{b}, d^{c})$ that satisfy the constraints of the problem with ex-post uncertainty can
be implemented in the benchmark problem by setting the punishment for lying and trading
$d^{b}_{\text{lie}}$ in the $\alpha = 1$ case equal to the expected punishment for lying and trading the amount
d^{b}_{\text{lie}}$ in the $\alpha < 1$ case (and likewise setting the punishing for truth-telling and trading $d^{b}$
in the $\alpha = 1$ case equal to the expected punishment for truth-telling and trading $d^{b}$ in the
$\alpha < 1$ case). This can be done without violating the self-enforcement constraints because
the surplus from cooperation across the two cases are equal. Put another way, loosely, the
best the informed trader can achieve via the $w^{-}(\cdot)$ mapping is to always punish lying and
never punish truth-telling over a set of price, report and outcome realizations of measure
one, where the measure is the probability measure defined by the equilibrium actions with
$d^{c} > 0$. To see that the upper-bound cannot be achieved for $x$ for which the solution $(d^{b}, d^{c})$
to the benchmark problem does not attain the monopoly profit, note that it is never possible
in the case with ex-post uncertainty to always punish lying and never punish truth-telling.
Thus, at least one of the IC constraints (2.12-2.13) will be violated for the solution $(d^{b}, d^{c})$
to the benchmark problem, for all $w^{-}(\cdot)$ functions.

(ii) A feasible punishment scheme is to punish the maximum self-enforceable amount
whenever the outcome $\theta$ differs from the report $\hat{\theta}$. For this penalty function, the problem
with $\alpha < 1$ and the benchmark problem with $\delta = \tilde{\delta}$, $\beta = \tilde{\beta}$ and $\alpha = 1$ coincide.
(iii) An immediate consequence of (i) and (ii) and the solution to the benchmark problem given in Proposition 21.

Profits in autarky are the same when \( \alpha < 1 \) and when \( \beta = \tilde{\beta} \) and \( \alpha = 1 \). Thus, part (i) of Proposition 24 emphasizes that even holding the ”quality” of the informed trader’s information constant, the informed trader is made worse off by ex-post uncertainty. Part (ii) of the lemma defines an uncertainty-adjusted discount factor such that the benchmark problem using this discount factor (and an uncertainty-adjusted \( \tilde{\beta} \)) provides a lower-bound for the problem with ex-post uncertainty. Since a Folk Theorem held for the benchmark problem, and since \( \lim_{\delta \to 1} \tilde{\delta} = 1 \), a Folk Theorem also applies to the problem with ex-post uncertainty. Part (iii) partially characterizes the solution to the problem with ex-post uncertainty. It already shows that the informed trader, in the case of ex-post uncertainty, can be too big to be believed. For x below a cutoff, the informed trader attains the monopoly profit via unconstrained cooperation. For x above a different cutoff, no cooperation is possible and the informed trader attains only the autarky profit.

Below I will provide a closed-form solution to the problem with ex-post uncertainty under the following parametric assumption:

\[
\sigma_\varepsilon > \frac{1}{\sqrt{2\pi}} \frac{E[\theta|s = \beta]}{\delta} \left[ \frac{(1 - \delta)^2}{(1 - \delta^2)^2 + 1} \right]
\] (A)

This assumption says that the variance of the noise-trader demand is large relative to the informed trader’s information, measured by \( E[\theta|s = \beta] \).

Before proceeding with this analysis, I provide in Figure 4 a comparison of the maximum

---

10 As \( \alpha \) approaches one, this condition becomes

\[
\sigma_\varepsilon > \frac{1}{\sqrt{2\pi}} \delta E[\theta|s = \beta]
\]

As \( \alpha \) approaches \( \frac{1}{2} \) from above, this condition becomes

\[
\sigma_\varepsilon > \frac{1}{\sqrt{2\pi}} \frac{E[\theta|s = \beta]}{1 - \delta}
\].
trading that can be achieved with and without noise traders. The maximum trading position, as a function of the position limit $x$, displays the familiar V shape when noise traders are and are not present. Total trading with ex-post uncertainty is weakly below total trading in the benchmark case, and strictly so for values of $x$ for which the benchmark case features constrained cooperation.

**Proposition 25** When Condition (A) holds, the unique solution to the informed trader’s problem is given by:

$$d^b = \begin{cases} 
    x & \text{if } x < \frac{E[\theta|s=\beta]}{2\sigma} \\
    \frac{E[\theta|s=\beta]}{2\sigma} & \text{if } x \geq \frac{E[\theta|s=\beta]}{2\sigma}
\end{cases}$$

(2.16)

$$d^c = \max\{\min\left\{\frac{E[\theta|s=\beta]}{2\sigma} - x, \frac{E[\theta|s=\beta]}{\sigma} - 2x(1 + \frac{1}{q_{\text{lie}} - q_{\text{true}} \frac{1 - \delta}{\delta}})\right\}, 0\}$$

(2.17)

where $q_{\text{lie}}$ and $q_{\text{true}}$ are given by

$$q_{\text{lie}}(x) = \alpha \Phi(-\frac{\sigma_x \ln(1-\alpha)}{2\sigma x} + \frac{\sigma}{\sigma_x} x) + (1-\alpha) \Phi(-\frac{\sigma_x \ln(1-\alpha)}{2\sigma x} + \frac{\sigma}{\sigma_x} x)$$

$$q_{\text{true}}(x) = \alpha \Phi(-\frac{\sigma_x \ln(1-\alpha)}{2\sigma x} - \frac{\sigma}{\sigma_x} x) + (1-\alpha) \Phi(-\frac{\sigma_x \ln(1-\alpha)}{2\sigma x} - \frac{\sigma}{\sigma_x} x).$$
for \( x > 0 \), and where \( q^{\text{lie}} = \alpha \) and \( q^{\text{true}} = 1 - \alpha \) for \( x = 0 \).

Define the maximum self-enforceable punishment \( w^{\text{max}} \):

\[
w^{\text{max}} = \frac{\delta}{1 - \delta} (E[\Pi(d^b + d^c, 0)|s = \beta] - \Pi_{\text{aut}}).
\]  

(2.18)

For \( x \in (0, d^m(\beta)) \), the penalty function \( w^-(.) \) is given by

\[
w^-(\beta, -\beta, p) = \begin{cases} 0 & \text{if } p < p_N \\ w^{\text{max}} & \text{if } p > p_N \end{cases}
\]

(2.19)

\[
w^-(\beta, \beta, p) = \begin{cases} w^{\text{max}} & \text{if } p < -p_Y \\ 0 & \text{if } p > -p_Y \end{cases}
\]

\[
w^-(\beta, -\beta, p) = \begin{cases} w^{\text{max}} & \text{if } p < -p_N \\ 0 & \text{if } p > -p_N \end{cases}
\]

\[
w^-(-\beta, -\beta, p) = \begin{cases} w^{\text{max}} & \text{if } p < -p_N \\ 0 & \text{if } p > -p_N \end{cases}
\]

where \( p_N \) and \( p_Y \) are uniquely defined by

\[
\alpha \phi(\frac{p_N - \sigma(x - d^c)}{\sigma}) = (1 - \alpha) \phi(\frac{p_N - \sigma(-d^b - d^c)}{\sigma})
\]

(2.20)

\[
(1 - \alpha) \phi(\frac{p_Y - \sigma(x - d^c)}{\sigma}) = \alpha \phi(\frac{p_Y - \sigma(-d^b - d^c)}{\sigma}).
\]

(2.21)

For \( x = 0 \), the penalty function \( w^-(.) \) is given by

\[
w^-(\beta, -\beta, p) = w^-(\beta, \beta, p) = w^{\text{max}}
\]

\[
w^-(\beta, -\beta, p) = w^-(\beta, \beta, p) = 0
\]

When \( x \) is such that the clients trade a positive amount, but not enough to achieve the monopoly profit, the punishment function is unique in that any other punishment function that achieves the optimum must be equivalent except over a set of measure zero.

**Proof.** See Appendix. ■

\( q^i \) is the probability that the informed trader is punished after he has told the truth \((i = \text{true})\) or lied \((i = \text{lie})\). This difference \( q^{\text{lie}} - q^{\text{true}} \) will always be greater than
$2\alpha - 1$ since the punishment rule could always ignore the price and punish if and only if the informed trader’s report is correct. Indeed, when $x = 0$, this punishment rule is used and $q^{\text{lie}} - q^{\text{true}}$ equals this lower-bound. This is because the informed trader does not trade and thus there is no reason to incorporate the price, a signal of the informed trader’s trade, into the penalty function.

When $x > 0$, however, the price is a signal of the informed trader’s behavior. The optimal punishment rule incorporating the price is a cutoff function. The cutoff is determined as follows. Consider some point $(\theta, \hat{\theta}, p')$ of the domain of $w^-(\cdot)$. Take $\theta = \beta$ and $\hat{\theta} = -\beta$, for example. This tuple of outcomes, reports and prices can occur in two situations: (1) $s = -\beta$, the informed player is truthful ($\hat{\theta} = s$), the outcome is the opposite of the signal ($\theta = -s$) and $\varepsilon = p' - \sigma(-x - d\varepsilon)$; or (2) $s = \beta$, the informed player is untruthful ($\hat{\theta} = -s$), the outcome is the same as the signal ($\theta = s$) and $\varepsilon = p' - \sigma(x - d\varepsilon)$. The penalty $w^-(\beta, -\beta, p')$ appears on the left-hand-side of the $s = -\theta$ IC constraint, while also appearing on the right-hand-side of the $s = \theta$ IC constraint. If the Lagrange multipliers on these constraints are equal, as they are in equilibrium, then $w^-$ should be set equal to the maximum self-enforceable penalty if ($\theta = s$) and $\varepsilon = p' - \sigma(-x - d\varepsilon)$ is more likely than ($\theta = -s$) and $\varepsilon = p' - \sigma(x - d\varepsilon)$). This logic applies because $w^-(\cdot)$ enters the problem solely in the IC and self-enforcement constraints. The probability $\theta = s$ is $\alpha$ and the probability distribution of $\varepsilon$ is $N(0, \sigma^2\varepsilon)$. This defines $w^-(\beta, -\beta, p)$. An analogous procedure pins downs the rest of $w^-(\cdot)$. In this sense, the punishment rule given by (2.19) maximizes $q^{\text{lie}} - q^{\text{true}}$ for a given level of $x$.

**Corollary 26** The difference in the probability of punishment, $q^{\text{lie}} - q^{\text{true}}$, is continuous and increasing in the position limit $x$ and the illiquidity parameter $\sigma$ and is continuous and decreasing in the standard deviation of noise trading $\sigma_{\varepsilon}$. The elasticity of $q^{\text{lie}} - q^{\text{true}}$ with respect to the position limit $x$ is less than one.

**Proof.** Some derivatives, for $x > 0$, are given by:

\[
\frac{d(q^{\text{lie}} - q^{\text{true}})}{dx} = \varphi \frac{\sigma}{\sigma_{\varepsilon}} > 0
\]

\[
\frac{d(q^{\text{lie}} - q^{\text{true}})}{d\sigma_{\varepsilon}} = -\varphi \frac{\sigma}{\sigma_{\varepsilon}^2} < 0
\]

\[
\frac{d(q^{\text{lie}} - q^{\text{true}})}{d\sigma} = \varphi \frac{x}{\sigma_{\varepsilon}} > 0
\]
where

\[ \varphi = 2[\alpha \phi(-\frac{\sigma \epsilon \ln(\frac{1-\alpha}{\alpha})}{2\sigma x} + \frac{\sigma}{\sigma \epsilon} x) + (1 - \alpha) \phi(-\frac{\sigma \epsilon \ln(\frac{\alpha}{1-\alpha})}{2\sigma x} + \frac{\sigma}{\sigma \epsilon} x)] \in (0, \sqrt{\frac{2}{\pi}}). \]

These derivatives can be written in this simple form using the definitions (2.20) and (2.21) to note that:

\[ \alpha \phi(-\frac{\sigma \epsilon \ln(\frac{1-\alpha}{\alpha})}{2\sigma x} - \frac{\sigma}{\sigma \epsilon} x) = (1 - \alpha) \phi(-\frac{\sigma \epsilon \ln(\frac{\alpha}{1-\alpha})}{2\sigma x} + \frac{\sigma}{\sigma \epsilon} x). \]

The corollary is intuitive. When the position limit of the informed trader, and hence the trading of the informed trader, is greater, it is easier to detect his misbehavior. When illiquidity \( \sigma \) is higher, the price impact of the position taken by the informed trader is larger, making it easier to detect. Similarly, greater variance in noise trading reduces the informativeness of the price signal.

**Corollary 27** For \( d^c \in (0, \frac{E[\theta|s=\beta]}{2\sigma} - x) \), the following comparative statics hold:

\[
\begin{align*}
\frac{d(d^c)}{dx} &= -2(1 + \frac{1}{q^{lie} - q^{true}} \frac{1 - \delta}{\delta}) \\
&+ 2x \frac{1}{(q^{lie} - q^{true})^2} \frac{1 - \delta d(q^{lie} - q^{true})}{dx} < -1 \quad (2.22) \\
\frac{d(d^c)}{d\sigma} &= \frac{2x}{\sigma^2} \frac{1 - \delta d(q^{lie} - q^{true})}{d\sigma} < 0 \\
\frac{d(d^c)}{d\sigma} &= -\frac{E[\theta|s=\beta]}{\sigma^2} + \frac{2x}{(q^{lie} - q^{true})^2} \frac{1 - \delta d(q^{lie} - q^{true})}{d\sigma} \quad (2.23)
\end{align*}
\]

**Proof.** To see that \( \frac{d(d^c)}{dx} < -1 \), note that

\[
\frac{d(d^c)}{dx} < -1 \iff x \frac{d(q^{lie} - q^{true})}{dx} < \frac{1}{2} \frac{\delta}{1 - \delta} (q^{lie} - q^{true})^2 + (q^{lie} - q^{true})
\]

The later holds for all \( \delta \in (0, 1) \) if and only if

\[
x \frac{d(q^{lie} - q^{true})}{dx} < (q^{lie} - q^{true})
\]

which is equivalent to the elasticity of \( q^{lie} - q^{true} \) with respect to \( x \) being less than 1. ■

We can compare these comparative statics to those in case of no ex-post uncertainty.
For example, in the benchmark model without ex-post uncertainty, \( \frac{d(d^c)}{dx} = -\frac{2}{\delta} < -1 \). Increasing \( x \) led to a more than one-for-one decrease in the amount traded by clients, because a higher position limit increases the short-term trading gain from lying and decreases the maximum self-enforceable punishment. In the benchmark model with ex-post uncertainty, the first term of (2.22) reflects these effects. Indeed, these effects are magnified in the case with ex-post uncertainty, since \( q^{\text{true}} - q^{\text{false}} < 1 \). To offset the increased gain from lying, \( d^c \) falls to reduce the trading benefit from lying; because lying is sometimes not punished and truth-telling sometimes is punished, \( d^c \) has to fall more in the case with ex-post uncertainty. The second term of (2.22) is new. It captures that increasing \( x \) increases the difference between the probability of being punished when lying and the probability of being punished when telling the truth.

Intuitively, the amount of trading by clients decreases in the variance of noise trading, because noise trading reduces the informativeness of the price signal. In the benchmark problem with \( \alpha = 1 \), an increase in noise trading had no effect on the amount traded by clients.

In the benchmark model, trading by clients decreases in the market’s illiquidity, because the trading profits from lying are increasing in the market’s illiquidity and the maximum self-enforceable punishment is decreasing in the market’s illiquidity. For the benchmark model, \( \frac{d(d^c)}{dx} \) corresponds exactly to the first term of (2.23). The second term is new: in the model with ex-post uncertainty, greater illiquidity facilitates cooperation because it strengthens the accuracy of the price signal, as explained in Corollary (26).

If

\[
E[\theta|s = \beta] - \sigma(x + d^c) \geq \frac{1 - q^{\text{true}}(x)}{1 - \delta} (E[\theta|s = \beta] - \sigma d^c - 2\sigma x)
\]

then the maximum equilibrium penalty \( w^{\text{max}} \) is less than the fee \( w \) and thus the informed trader never need pay any compensation to the clients; rather, when the informed trader is punished, the punishment can take the form of a discount on next period’s fee.

The analysis thus far has assumed condition A holds. When condition A holds, the informed trader’s ability to "game the system" after lying by not trading as much as its position limit turns out not to affect the equilibrium level of cooperation. Condition A implies that, for \( x \in (x_L, \delta \frac{E[\theta|s = \beta]}{2\sigma}) \), after lying, the informed trader’s position limit does in fact bind in the maximizations in (2.12) and (2.13). For \( x \in (x_L, \delta \frac{E[\theta|s = \beta]}{2\sigma}) \), the informed
trader trades $d_{lie}^b = x \cdot \text{sign}(s)$ after lying. For $x < x^L$, the informed trader after lying may trade less than his position limit allows in order to reduce the likelihood of being punished, but this does not affect the equilibrium, since full cooperation is sustainable in equilibrium when the trader is punished if and only if his report differs from the outcome.

When A does not hold, the payoff using (2.16) and (2.17) provide an upper-bound for the payoff of the informed trader in the problem with ex-post uncertainty, as shown in the proof of Proposition 25. Moreover, we can calculate the solution numerically. The following proposition is useful for finding a numerical solution.

**Proposition 28** Whenever constraints (2.12) and (2.13) are binding, $d_b = x$ and the optimal $w^-(\cdot)$ function is a cutoff function characterized by two parameters $p_N$ and $p_Y$. The optimal cutoff function $w^-(\cdot)$ is given by (2.19) where

$$w_{\text{max}}^b = \frac{\delta}{1 - \delta}(V - \Pi_{\text{aut}}^b)$$

and $p_N$ and $p_Y$ are uniquely defined by:

$$\alpha \phi\left(\frac{p_N - \sigma(d_{lie}^b)}{\sigma_\varepsilon}\right) = (1 - \alpha) \phi\left(\frac{p_N - \sigma(-d_b - d_c)}{\sigma_\varepsilon}\right)$$

$$\alpha \phi\left(\frac{p_Y - \sigma(d_{lie}^b)}{\sigma_\varepsilon}\right) = (1 - \alpha) \phi\left(\frac{p_Y - \sigma(-d_b - d_c)}{\sigma_\varepsilon}\right).$$

Moreover, either both (2.12) and (2.13) bind, or neither does.

**Proof.** The proof is analogous to Step (ii) of the proof of Proposition 25, except that the envelope condition is used in taking first order conditions involving constraints (2.12) and (2.13). ■

Figure 5 provides shows equilibrium trading, calculated numerically, for a set of parameters for which Condition A does not hold. In the constrained cooperation region, total trading is a convex function of the informed trader’s position limit, whereas total trading was linear in the informed trader’s position limit in this region in the benchmark model.

### 2.5 Endogenous position limits

In the previous analysis, the position limit of the informed trader was exogenously given. This section studies the informed trader’s problem when the trader can freely choose its
Figure 5. Total trading as a function of the informed trader’s position limit

Note: The parameters used are: $\alpha = 0.51, \beta = 100, \delta = 0.9, \sigma_\varepsilon = 5, \sigma = 0.5$.

The timing of the game is as follows. At time zero, before the repeated game starts, the informed trader chooses a position limit $x$ at cost $f(x)$. The parties then play the repeated game exactly as in Section 3.1 with $x$ as the position limit. The informed trader’s payoff is:

$$\max_{x \in [0, \infty)} -f(x) + E \sum_{t=0}^{\infty} \delta^t \Pi_t(x, \eta)$$

where $\Pi_t$ is the period-$t$ profit.

The function $f(x)$ could reflect the costs of the financial, human or IT capital necessary to trade. I will study the problem under two different specifications of $f(x)$: a linear specification $f(x) = \kappa x$ that is not grounded in any specific theory of where $f(x)$ comes from; a non-linear specification that comes from requiring that the informed trader can never have negative end-of-period wealth. It turns out that the former specification, with $\kappa$ set appropriately, is a good approximation for the later. Non-negative wealth constraints are common in models of arbitrage (Vayanos and Gromb 2002). They are also a real-life constraint; a non-negative wealth constraint is like a requirement to hold more capital than value-at-risk (VaR), where the VaR confidence level is 100%.
The possibility that clients are not able to trade on the informed trader’s information, addressed in Section 3.1, plays an essential role here. Recall that the clients are able to trade on the informed trader’s information with probability $\eta \in [0, 1]$. When choosing the level of capital, the informed trader faces the following trade-off: (i) profits in periods of autarky (i.e., when the clients cannot trade on the information) are weakly increasing in the position limit; (ii) when the clients can trade on the information, monopoly profits can be achieved with a low, or even zero, position limit and increasing the position limit could actually decrease profits. The trade-off is made taking into account the cost of the position limit, $f(x)$, and, naturally, the probability that the clients can trade on the information.

For analytical clarity, I assume $\sigma_x = 0$ and return to the Section 3.1 benchmark model.

**Proposition 29** Suppose

$$f(x) = \kappa x$$

and that $\eta^* > 0$ and $\eta^{**} < 1$, where $\eta^*$ and $\eta^{**}$ are defined below.

Then

$$x = \begin{cases} \frac{E[\theta|s=\beta]-(1-\delta)\kappa}{2\sigma} & \text{if } \eta \in [0, \eta^*) \\ \frac{(1+2\frac{1-\delta}{\kappa})E[\theta|s=\beta]-\kappa(1-\delta)}{2\sigma(1+4\frac{1-\delta}{\kappa}+4(1-\delta)^2)} & \text{if } \eta \in [\eta^*, \eta^{**}] \\ \min\left\{ \frac{E[\theta|s=\beta]-(1-\delta)\kappa}{2\sigma}, 0 \right\} & \text{if } \eta \in [\eta^{**}, 1] \end{cases}$$

where

$$\eta^{**} = 2\frac{(1-\delta)E[\theta|s=\beta]+\kappa(1-\delta)^2}{2(1-\delta)E[\theta|s=\beta]-\kappa(1-\delta)\delta}$$

The optimal choice of position limit $x$ is: (i) constant over $\eta \in [0, \eta^*)$; (ii) increasing over $\eta \in [\eta^*, \eta^{**}]$; and (iii) weakly decreasing over $\eta \in [\eta^{**}, 1]$. The position limit $x$ is continuous in $\eta$ except at $\eta = \eta^*$.

For $\eta < \eta^*$, the endogenous position limit is such that, in periods when the clients can trade on the information, the informed trader is in a region of constrained autarky: he does not achieve the monopoly profit, but he cannot truthfully communicate his information to any clients because he cannot be trusted. In this region, the informed trader chooses the same high position limit regardless of $\eta$. All of the informed trader’s profits come from trading for his own book; none come from client fees.

For $\eta \in (\eta^*, \eta^{**})$, the endogenous position limit is such that, in periods when the clients can trade on the information, the informed trader is in a region of constrained cooperation;
that is, greater profits could be had by the informed trader in such periods if his position limit were lower. This suggests that the existence of such a region in Sections 3 and 4 is not irrelevant. The reason why the informed trader for \( \eta \in (\eta^*, \eta^{**}) \) does not choose a lower position limit is that the trader would be less profitable in periods when clients cannot trade on its information.

For \( \eta \in (\eta^*, \eta^{**}) \), surprisingly, capital is increasing in \( \eta \). Moreover, there is a discontinuity at \( \eta^* \). Roughly speaking, the ”business plan” of the informed trader for \( \eta \in (\eta^*, \eta^{**}) \) is to earn fairly high profits for his own book when the clients cannot trade on his information and to earn nearly the monopoly profits from fees and trading when the clients can trade on the information. That is, the informed trader chooses a position limit slightly higher beyond the first kink in the graph of profits against the position limit. If the informed trader reduced its position limit, it could achieve a slightly higher profit in periods of communication, at the expense of significantly lower profits in periods of autarky. As \( \eta \) increases, the kink dividing the regions of unconstrained and constrained cooperation shifts right, because the incentives for truthful communication are greater (see Figure 3). Thus, the informed trader’s ”business plan” calls for a higher position limit. That is, as \( \eta \) increases, the point at which the marginal loss of profit in periods of communication exactly offsets the marginal benefit gain in profit in periods without communication moves rightward. Only in the region \( \eta \in (\eta^*, \eta^{**}) \) does this marginal condition hold, since for other values of \( \eta \), the solution is a type of corner solution.

For \( \eta \in (\eta^{**}, 1] \), the choice of capital results in a position limit such that, in periods when the clients can trade on the information, the informed trader is in a region of unconstrained cooperation. Here the informed trader achieves the monopoly profit when its clients are able to trade, at the expense of earning low profits in autarky when the clients cannot trade on the information. In the region \( \eta \in (\eta^{***}, 1) \), the informed trader chooses a zero level of capital.

We can interpret variation in how informed traders profit from their information as depending on the type of securities they are informed about. Players that frequently have information about very exotic securities are unlikely find that clients want to buy their information (\( \eta \) small); these players choose to earn their profits by trading for their own books, rather than communicating their information to clients. Their proprietary desks are so large that no one would believe their advice. These players resemble hedge
Figure 6. Equilibrium position limit as a function of clients’ understanding probability $\eta$

Players with information about securities that sometimes can be traded by clients balance trading for their own book and client fees; these players resemble investment banks that combine proprietary trading with prime, institutional and high-end retail brokerage services. Some of these players are constrained in the size of their client base because clients know that the informed players could lie and trade against them. Finally, players informed about securities that can be readily traded by clients earn all of their income from selling their information and none from trading for their own books. These players resemble independent research firms. One can think of these as mutual-fund companies; in this analogy, the clients delegate the decision making to the informed trader, who has no incentive to lie because it has no proprietary trading desk separate from its funds.

Thus far, the position limit cost function $f(x)$ was not micro-founded, but rather simply assumed to be linear (with no intercept). One micro-foundation for $f(x)$ is to assume it reflects costs of capital and that the informed trader needs capital to trade for his own book because of a zero-wealth constraint. Assuming that the informed trader’s information is noisy ($\alpha < 1$), taking a position is risky. The informed trader is required to hold enough capital such that even if he turns out to be misinformed, his resulting wealth is non-negative in every period. That is, in order to achieve a position limit $x$, the informed trader must invest $K(x, \eta)$ in period zero, where $K(x, \eta)$ is given by

$$K(x, \eta) = x \ast (\beta + p(x, \eta))$$  \hfill (2.25)
where \( p(x, \eta) \) is the equilibrium price of the repeated game in Section 3.1, when the probability of a correct signal is \( \alpha \) and the probability that clients can trade on the informed trader’s information in a given period is \( \eta \). The notation in (2.25) makes explicit the dependence of the equilibrium price on the probability \( \eta \); the equilibrium price can be calculated using Proposition 23. We can take \( f(x, \eta) = K(x, \eta) \).

**Lemma 30** \( K(x, \eta) \) is given by:

\[
K(x, \eta) = \begin{cases} 
  x(\beta + \frac{E[\theta | s=\beta]}{2}) & \text{if } x < \frac{E[\theta | s=\beta]}{2\sigma} \cdot \frac{\delta \eta}{2 - 2\delta + \delta \eta} \text{ or } x > \frac{E[\theta | s=\beta]}{2\sigma} \cdot \frac{\delta \eta}{1 - \delta + \delta \eta} \\
  x(\beta + E[\theta | s=\beta] - (1 + \frac{1 - \delta}{\eta \delta})\sigma x) & \text{if } x \in \left[ \frac{E[\theta | s=\beta]}{2\sigma} \cdot \frac{\delta \eta}{2 - 2\delta + \delta \eta}, \frac{E[\theta | s=\beta]}{2\sigma} \cdot \frac{\delta \eta}{1 - \delta + \delta \eta} \right] \\
  x(\beta + \sigma x) & \text{otherwise}
\end{cases}
\]

\( K(x, \eta) \) is continuous and, for \( \alpha < 1 \), strictly increasing in \( x \).

**Proof.** Follows from Proposition 23. ■

When the maximization in (2.24) is solved using \( f(x, \eta) = K(x, \eta) \), the solution is qualitatively similar to the solution when \( f(x, \eta) = \kappa x \) is used, for \( \kappa \approx \beta + \frac{E[\theta | s=\beta]}{2} \), except that there is an additional discontinuity at \( \eta = \eta^{**} \).

### 2.6 Conclusion

This paper argues that two activities that Wall Street firms typically pursue simultaneously – proprietary trading and advising – create a conflict of interest with the result that the extent of proprietary trading limits the size of the firm’s advising business. I argue that, when considering an informed trader, the bundling of these two activities in a single firm needs to be taken into account.

In future work, it would be useful to address these issues in an explicit dynamic moral hazard problem in which the informed trader’s capital is funded by a third party and the informed trader’s capital level and position limit vary endogenously over time based on profits and payments to and from the third party. An example of such an approach in a corporate finance context is given in Biais et al (2007), in which the agency problem is simpler than the one considered here.

Finally, this model offers a rationale for why some informed traders trade for their own book and others profit by selling information to clients, while still others mix the
two activities. It would be interesting to explore related explanations for the business plans chosen by informed traders. In particular, investment banks frequently combine multiple activities in which there is moral hazard and in which clients pay rents to the bank to achieve efficiency-wage-type enforcement. If shirking cannot be coordinated across activities, perhaps because of temporal asynchronicities, a given level of rents paid to the bank can provide incentives across a range of activities, economizing on the rents paid.\footnote{See Charles Ellis, “The Partnership: The Making of Goldman Sachs,” 2008, p. 421.}
Appendix: Proofs omitted from the text

**Proof of Proposition 25.** This proof proceeds in three steps. Step (i) shows that the proposed solution in Proposition (25) coincides with the solution given in Proposition (24) for \( x \leq x^L \) and \( x \geq \delta \frac{E[\theta|s=\beta]}{2\sigma} \). Step (ii) describes an alternate problem that provides an upper-bound for the payoff to the informed trader in the problem with ex-post uncertainty. I solve this alternate problem. Step (iii) shows that, for \( x \in (x^L, \delta \frac{E[\theta|s=\beta]}{2\sigma}) \), the solution to the alternate problem is the solution to the original problem.

*Step (i)*

I will show that the proposed solution in Proposition (25) coincides with the solution given in Proposition (24) for \( x \leq x^L \) and \( x \geq \delta \frac{E[\theta|s=\beta]}{2\sigma} \). Observe that \((2\alpha - 1) \leq q^{lie} - q^{true} \leq 1\). To see that the solutions coincide over the range \( x \leq x^L \), note that (2.16) and (2.17) imply that the monopoly profit is achieved for

\[
x \leq \frac{E[\theta|s=\beta]}{2\sigma} \frac{\delta}{q^{true} - q^{true}} (1 - \delta) + \delta
\]

The right-hand-side is increasing in \((q^{lie} - q^{true})\), so \((q^{lie} - q^{true}) \geq (2\alpha - 1)\) implies that the minimum value of the right-hand-side is \(x_L\). To see that they coincide over the range \( x \geq \delta \frac{E[\theta|s=\beta]}{2\sigma} \), note that (2.16) and (2.17) imply that only autarky is possible when

\[
x \geq \frac{E[\theta|s=\beta]}{2\sigma} \frac{\delta}{\delta + \frac{1-\delta}{q^{true} - q^{true}}}
\]

The right-hand-side is increasing in \((q^{lie} - q^{true})\), so \(q^{lie} - q^{true} \leq 1\) implies that the maximum values of the right-hand-side is \(\delta \frac{E[\theta|s=\beta]}{2\sigma}\). Thus, the remainder of the proof will focus on \( x \in (x^L, \delta \frac{E[\theta|s=\beta]}{2\sigma}) \).

*Step (ii)*

Consider an alternative problem, identical to the informed trader’s problem with ex-post uncertainty except that after lying, the informed trader can only trade \( d^{tie} = x \times sign\{s\} \).

By tying the liar’s hands, this assumption (weakly) reduces the gains from lying, making it easier to sustain cooperation. Thus, the payoff to the informed trader in the alternative problem is an upper-bound for the payoff to the informed trader in the original problem.
The alternative problem can be written as:

$$\max_{d^b,d^c,w,w^-} E[\Pi(d^b + d^c, 0) | s = \beta]$$

subject to

$$E[\Pi(x, -d^c) - w^- (\theta, -\beta, p(x, -d^c)) | s = \beta] \leq E[\Pi(d^b, d^c) - w^- (\theta, \beta, p(d^b, d^c)) | s = \beta]$$

$$E[\Pi(-x, d^c) - w^- (\theta, \beta, p(-x, d^c)) | s = -\beta] \leq E[\Pi(-d^b, -d^c) - w^- (\theta, -\beta, p(-d^b, -d^c)) | s = -\beta]$$

$$0 \leq w^- (\theta, s, p) \leq \frac{\delta}{1 - \delta}(V - \Pi_{aul}). \quad (2.26)$$

$$|d^b| \leq x$$

$$d^b \in \arg \max_{|d| \leq x} E[\Pi^b_{share}(d, d^c) | s = \beta]$$

Assign the Lagrange multipliers $\lambda_1$ and $\lambda_2$ to the first two constraints. Assign $\lambda_5$ to the fifth constraint. Ignore the sixth constraint momentarily. To the remaining constraints, assign to constraints the Lagrange multipliers $\lambda_3(\theta, \hat{\theta}, p)$ and $\lambda_4(\theta, \hat{\theta}, p)$. The Lagrange multiplier $\lambda_3(\theta, \hat{\theta}, p)$ applies to the first inequality in (2.26) and $\lambda_4(\theta, \hat{\theta}, p)$ applies to the second inequality in (2.26).

The first-order conditions with respect to $w^- (.)$ are:

$$\lambda_1 \phi \left( \frac{p - \sigma(x - d^c)}{\sigma_\epsilon} \right) - \lambda_2 (1 - \alpha) \phi \left( \frac{p - \sigma(-d^b - d^c)}{\sigma_\epsilon} \right) - \lambda_3 (\beta, -\beta, p) + \lambda_4 (\beta, -\beta, p) = 0$$

$$\lambda_1 (1 - \alpha) \phi \left( \frac{p - \sigma(x - d^c)}{\sigma_\epsilon} \right) - \lambda_2 \alpha \phi \left( \frac{p - \sigma(-d^b - d^c)}{\sigma_\epsilon} \right) - \lambda_3 (-\beta, -\beta, p) f(-\beta, -\beta, p) + \lambda_4 (-\beta, -\beta, p) = 0$$

$$-\lambda_1 \alpha \phi \left( \frac{p - \sigma(d^b + d^c)}{\sigma_\epsilon} \right) + \lambda_2 (1 - \alpha) \phi \left( \frac{p - \sigma(x + d^c)}{\sigma_\epsilon} \right) - \lambda_3 (\beta, \beta, p) + \lambda_4 (\beta, \beta, p) = 0$$

$$-\lambda_1 (1 - \alpha) \phi \left( \frac{p - \sigma(d^b + d^c)}{\sigma_\epsilon} \right) + \lambda_2 \alpha \phi \left( \frac{p - \sigma(x + d^c)}{\sigma_\epsilon} \right) - \lambda_3 (-\beta, \beta, p) + \lambda_4 (-\beta, \beta, p) = 0$$

The first FOC is for $w^- (\theta = \beta, \hat{\theta} = -\beta, p)$. The second, third and fourth FOC are for $w^- (\theta = -\beta, \hat{\theta} = -\beta, p)$, $w^- (\theta = \beta, \hat{\theta} = \beta, p)$ and $w^- (\theta = -\beta, \hat{\theta} = \beta, p)$, respectively.

I will conjecture, based on the symmetry of the $s = \beta$ and $s = -\beta$ cases, that

$$\lambda_1 = \lambda_2$$
and I will verify this below.

If
\[ d^b + d^c < d^b_m \]
then
\[ \lambda_1 = \lambda_2 > 0 \]

because otherwise, \( d^c \) could be increased by some small amount \( \Delta \) and a new penalty function \( \tilde{w}^-(\cdot) \) used such that all of the constraints remain satisfied and, indeed, the fourth constraint would slacken if it had been binding. This would increase the payoff. (The new penalty function would be \( \tilde{w}^-(\theta, \hat{\theta}, p) = w^-(\theta, \beta, p - \sigma \Delta \cdot \text{sign}(\hat{\theta})) \).)

Suppose
\[ d^b + d^c < d^b_m \]

Define \( p_N \) and \( p_Y \) as follows:

\[
\begin{align*}
\alpha \phi\left( \frac{p_N - \sigma(x - d^c)}{\sigma \varepsilon} \right) &= (1 - \alpha) \phi\left( \frac{p_N - \sigma(-d^b - d^c)}{\sigma \varepsilon} \right) \\
(1 - \alpha) \phi\left( \frac{p_N - \sigma(x - d^c)}{\sigma \varepsilon} \right) &= \alpha \phi\left( \frac{p_Y - \sigma(-d^b - d^c)}{\sigma \varepsilon} \right)
\end{align*}
\]

where \( p_N \) and \( p_Y \) are given by:

\[
\begin{align*}
p_N &= 2 \sigma^2 \ln\left( \frac{1 - \alpha}{\alpha} \right) - 2 \sigma^2 d^c(x + d^b) \\
p_Y &= 2 \sigma^2 \ln\left( \frac{\alpha}{1 - \alpha} \right) - 2 \sigma^2 d^c(x + d^b)
\end{align*}
\]

(2.27) (2.28)

Define
\[ w_{\text{max}} = \frac{\delta}{1 - \delta} (V - \Pi^b_{\text{aut}}) \]

Then cutoff function \( w^-() \) given in (2.19) is the unique function consistent with the first-order conditions with respect to \( w^-() \) and the third and fourth constraints of the problem. Given these \( w^-() \) functions, the first two constraints are the same, confirming that \( \lambda_1 = \lambda_2 \).
Now consider the FOC with respect to \( d^b \):

\[
d^b : \frac{1}{2}(E[\theta|s = \beta] - 2\sigma(d^b + d^c)) + \frac{1}{2}(-(E[\theta|s = -\beta] - 2\sigma(-d^b - d^c))
\]

\[
+ \lambda_1(E[\theta - \sigma(d^b + d^c) - \sigma d^b|s = \beta] - \frac{d}{d(d^b)}E[w^-(\theta, \beta, \sigma(d^b + d^c))|s = \beta])
\]

\[
+ \lambda_2(E[-(\theta - \sigma(-d^b - d^c)) - \sigma d^b|s = -\beta] - \frac{d}{d(d^b)}E[w^-(\theta, -\beta, \sigma(-d^b - d^c))|s = -\beta])
\]

\[
+ \int \lambda^3(\theta, \hat{\theta}, p)dF(\theta, \hat{\theta}, p)\frac{\delta}{1-\delta}(\frac{1}{2}(E[\theta|s = \beta] - 2\sigma(d^b + d^c)) + \frac{1}{2}(-(E[\theta|s = -\beta] - 2\sigma(-d^b - d^c))))
\]

\[
- \lambda^5 = 0
\]

Using the \( w^-(\cdot) \) functions derived above, we have

\[
E[w^-(\theta, \beta, \sigma(d^b + d^c))|s = \beta] = \alpha \Phi(\frac{-\rho_Y - \sigma(d^b + d^c)}{\sigma_\varepsilon}) + (1 - \alpha) \Phi(\frac{-p_N - \sigma(d^b + d^c)}{\sigma_\varepsilon}) \tag{2.29}
\]

\[
E[w^-(\theta, \beta, \sigma(d^b + d^c))|s = -\beta] = \alpha(1 - \Phi(\frac{-\rho_Y - \sigma(-d^b - d^c)}{\sigma_\varepsilon})) + (1 - \alpha)(1 - \Phi(\frac{p_N - \sigma(-d^b - d^c)}{\sigma_\varepsilon})) \tag{2.30}
\]

allowing us to calculate

\[
- \frac{d}{d(d^b)}E[w^-(\theta, \beta, \sigma(d^b + d^c))|s = \beta] = (\alpha \phi(\frac{-\rho_Y - \sigma(d^b + d^c)}{\sigma_\varepsilon}) + (1 - \alpha) \phi(\frac{-p_N - \sigma(d^b + d^c)}{\sigma_\varepsilon})) \frac{\sigma}{\sigma_\varepsilon} w_{\text{max}} > 0
\]

and

\[
- \frac{d}{d(d^b)}E[w^-(\theta, -\beta, \sigma(-d^b - d^c))|s = -\beta] =
\]

\[
(\alpha \phi(\frac{-\rho_Y - \sigma(-d^b - d^c)}{\sigma_\varepsilon}) + (1 - \alpha) \phi(\frac{p_N - \sigma(-d^b - d^c)}{\sigma_\varepsilon})) \frac{\sigma}{\sigma_\varepsilon} w_{\text{max}} > 0
\]

Thus if \((d^b + d^c) < d^b_m\), then \( \lambda_5 > 0 \) and \( d^b = x \). This is consistent with the sixth constraint, which we have thus far ignored.

Substituting from (2.29-2.30) and (2.27-2.28) into the first constraint, which holds with equality, and using the conclusion that \( d^b = x \), we obtain a quadratic equation in \( d^c \), the solution to which is given in Proposition 25.
Step (iii)

I will show that, for \( x \in (x^L, \delta \frac{E[\theta | s = \beta]}{2\sigma}) \), the solution to the alternate problem is the solution to the original problem.

Given the penalty function (2.19), the solution to the problem in the IC constraints (2.12) and (2.13) is \( d_{\text{tie}}^b = x \cdot \text{sign}\{s\} \). To see this, note that

\[
\frac{d(E[\Pi^b(d_{\text{tie}}^b, -d^c) - w - (\theta, -\beta, p(d_{\text{tie}}^b, -d^c)) | s = \beta])}{d(d_{\text{tie}}^b)} = E[\theta | s = \beta] - 2\sigma d_{\text{tie}}^b + \sigma d^c - \frac{d(q_{\text{tie}}^b)}{d(d_{\text{tie}}^b)} w^{\text{max}}. 
\]

We know that, for \( x \in (x^L, \delta \frac{E[\theta | s = \beta]}{2\sigma}) \)

\[
\frac{dq_{\text{tie}}^b}{d(d_{\text{tie}}^b)} = \frac{\sigma}{\sigma_\epsilon} (\alpha \phi(-p_N - \sigma(d^c - x)) + (1 - \alpha) \phi(-p_Y - \sigma(d^c - x))) \leq \frac{1}{\sqrt{2\pi \sigma_\epsilon}} \frac{\sigma}{\sigma_\epsilon}
\]

\[
d_{\text{tie}}^b \leq \delta \frac{E[\theta | s = \beta]}{2\sigma}
\]

\[
w^{\text{max}} \leq \frac{\delta}{1 - \delta} \left[ E[\theta]^2 - x^L (E[\theta | s = \beta] - \sigma x^L) \right]
\]

The first statement follows from the properties of the normal pdf. The second statement follows the constraint \(|d_{\text{tie}}^b| \leq x\) and the restriction \( x \leq \delta \frac{E[\theta | s = \beta]}{2\sigma}\). The third statement follows from the definition of \( w^{\text{max}} \), given in (2.18), and the restriction \( x \geq x^L \). □
Bibliography

[1] Admati, Anat R. and Paul Pfleiderer, 1986. “A monopolistic market for information,” Journal of Economic Theory, Vol. 39, No. 2, pp. 400-438.

[2] Baker, George, Robert Gibbons, Kevin J. Murphy, 2002. “Relational Contracts and the Theory of the Firm,” The Quarterly Journal of Economics, Vol. 117, No. 1, pp. 39-84.

[3] Benabou, Roland and Guy Laroque, 1992, “Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility,” The Quarterly Journal of Economics, Vol. 107, No. 3, pp. 921-958.

[4] Brunnermeier, Markus K. and Lasse Heje Pedersen, 2005, “Predatory Trading,” The Journal of Finance, Vol. 60, No. 4, pp. 1825-1863.

[5] Brunnermeier, Markus K. and Lasse Heje Pedersen, 2009, “Market Liquidity and Funding Liquidity,” Review of Financial Studies, Vol. 22, No. 6, pp. 2201-2238.

[6] Crawford, Vincent P. and Joel Sobel, 1982, “Strategic Information Transmission,” Econometrica, Vol. 50, No. 6, pp. 1431-1451.

[7] Gromb, Denis and Dimitri Vayanos, 2002, “Equilibrium and welfare in markets with financially constrained arbitrageurs”, Journal of Financial Economics, Volume 66, Issues 2-3, pp. 361-407.

[8] He, Zhiguo and Arvind Krishnamurthy, 2010, “Intermediary Asset Pricing,” mimeo.

[9] Kondo, Jiro E. and Dimitris Papanikolaou, 2005, “Financial Relationships and the Limits to Arbitrage,” mimeo.
[10] Kyle, Albert S., 1985. “Continuous Auctions and Insider Trading,” Econometrica, Vol. 53, No. 6, pp. 1315-1335.

[11] Kyle, Albert S., 1989. “Informed Speculation with Imperfect Competition,” The Review of Economic Studies, Vol. 56, No. 3, pp. 317-355.

[12] Kyle, Albert S. and Wei Xiong, “Contagion as a Wealth Effect,” The Journal of Finance, Vol. 56, No. 4, pp. 1401-1440.

[13] Levin, Jonathan, 2003, “Relational Incentive Contracts,” The American Economic Review, Vol. 93, No. 3, pp. 835-857.

[14] Mitchell, Mark, Lasse Heje Pedersen and Todd Pulvino, 2007, “Slow Moving Capital,” American Economic Review, Vol. 97, No. 2, pp. 215-220.

[15] Rotemberg, Julio J. and Garth Saloner, 1986, “A Supergame-Theoretic Model of Price Wars during Booms,” The American Economic Review, Vol. 76, No. 3, pp. 390-407.

[16] Shleifer, Andrei and Robert W. Vishny, 1997, “The Limits of Arbitrage,” The Journal of Finance, Vol. 52, No. 1, pp. 35-55.

[17] Van Bommel, Jos, 2003, “Rumors,” The Journal of Finance, Vol. 58, No. 4, pp. 1499-1519.
Chapter 3

Preferred Habitat versus Limited Participation:
Irrelevance and Relevance Results for Open-Market Operations

3.1 Introduction

During and after the recent financial crisis, the Federal Reserve purchased long-term government bonds – financed by short-term borrowing – in an attempt to reduce long-term interest rates. Understanding how such maturity-lengthening open-market operations affect interest rates is an important challenge for macroeconomics.

According to one interpretation of the preferred-habitat model of Vayanos and Vila (2009), the existence of investors with preferred-habitat preferences implies that maturity-lengthening open-market operations (Federal Reserve purchases of long-term bonds financed by short-term borrowing) generate a reduction in long-term interest rates. In this paper, I clarify this interpretation by demonstrating that additional assumptions are needed for this interpretation.

In particular, in a Vayanos and Vila (2009) preferred-habitat model, I show that maturity-lengthening open-market operations have no effect on long-term interest rates if the agents in the economy ultimately receive the profits from the government’s portfolio via lump-sum...
taxes or transfers. This irrelevance result is similar to Ricardian equivalence and to the irrelevance results of Wallace (1981), Chamley and Polemarchakis (1984), and Eggertsson and Woodford (2003). The contribution of this paper with respect to these earlier papers is to show that maturity-lengthening open-market operations are irrelevant in the setting of Vayanos and Vila (2009) even though it has been asserted that the preferred-habitat preferences in Vayanos and Vila (2009) imply that maturity-lengthening open-market operations reduce long-term interest rates.

Next, I add an assumption to the Vayanos and Vila (2009) model such that maturity-lengthening OMO do result in a decrease in long-term rates. This assumption is a limited-participation assumption: namely, that the preferred-habitat investors cannot invest in short-term bonds. This is different than the Vayanos and Vila (2009) assumption that the preferred-habitat investors have preferences such that, without any maturity-lengthening OMO, these investors do not want to invest in short-term bonds. The limited-participation assumption is needed to break the irrelevance result because, although the preferred-habitat investors do not want to invest in short-term bonds if there are no maturity-lengthening OMO, they do want to do so if the government does engage in maturity-lengthening OMO. In particular, the preferred-habitat investors will invest in short-term bonds precisely to the extent necessary to undo the effects of the OMO on their consumption path; this is what generates the Ricardian equivalence result. The limited-participation assumption then breaks the irrelevance result by preventing – by assumption – the preferred-habitat investors from undoing the effects of the OMO by purchasing short-term bonds.

Indeed, it is the limited-participation assumption – and not the preferred-habitat preferences – that allow maturity-lengthening OMO to generate a decline in long-term interest rates in the resulting model. Assuming this type of limited participation may be less plausible than assuming preferred-habitat preferences; while it seems reasonable to posit the existence of some investors (like insurance or pension companies) who have a known long-term liability that they want to match with a safe long-term asset, it may seem less plausible that some investors cannot invest in short-term bonds or cannot invest in long-term bonds, even if they want to. Thus, perhaps the best justification for limited participation is based on transactions costs or a behavioral motivation; for example, there do appear to be many households that invest predominantly in short-term assets (like bank accounts or money-market funds) without considering – or paying the fixed cost of considering and acting upon
whether they would be better off investing in longer-duration assets.

Thus, I develop a model with limited participation (and without preferred-habitat preferences) in which maturity-lengthening OMO lead to a decrease in the long-term interest rate. The limited-participation assumption is that some households only invest in short-term assets and do not trade every period.

I then use the two limited-participation models to investigate whether maturity-lengthening OMO increase the vulnerability of the arbitrageurs to adverse shocks by shifting long-term safe assets – a hedge to many adverse shocks – from the balance sheet of arbitrageurs to the balance sheet of the government and the general public.

The way in which maturity-lengthening OMO affect the vulnerability of the arbitrageurs differs across the two models. In the model with a Vayanos and Vila (2009) setup with limited participation, the decrease in the net supply of long-term bonds generated by government OMO is met by arbitrageurs who short long-term bonds (relative to their position under no government intervention), in return for a risk premium.¹ In the event of a negative technology shock, interest rates fall and the arbitraguer experiences losses on his short position. I characterize the joint distribution of technology shocks and arbitrageurs’ terminal wealth. I show that the covariance of productivity and arbitrageurs’ terminal wealth is increasing (linearly) in the dollar amount of OMO purchases.

The model with limited participation without preferred-habitat preferences emphasizes wealth accumulation of the arbitrageurs. Wealth accumulation plays no formal role in the Vayanos and Vila (2009) literature, even in versions of the model with more than three periods, because arbitrageurs are assumed to have mean-variance preferences over instantaneous profits. In the model with limited participation but without preferred-habitat preferences, OMO serve both to tilt downward the consumption profile of the arbitrageurs and to raise the overall level of the consumption profile (because the arbitrageurs sell long-term bonds to the government at a high price). The former effect increases the vulnerability of the arbitrageurs, while the later effect decreases it. I characterize the conditions under which the later effect dominates the former and hence OMO decrease the vulnerability of the arbitrageurs.

¹I write “the decrease in the net supply of long-term bonds that is generated by government OMO,” rather than simply “government purchases of long-term assets” because, due to the general-equilibrium, Ricardian-equivalence logic developed in this paper and in the Wallace (1981) literature, these two quantities in quotation marks may not be the same. Indeed, a main point of the paper is that without limited participation, the first quantity is zero even when the second quantity is positive.
Related literature and modelling approach

Preferred-habitat literature. This paper builds on – and critically examines – the preferred-habitat literature of Vayanos and Vila (2009), Greenwood and Vayanos (2010) and Greenwood, Hanson and Stein (2010).

This literature features short-term and long-term bonds, with an exogenous and stochastic short-term interest rate. The key element of the models are preferred-habitat investors who consume only in the long-term and who are infinitely risk averse (i.e., they have max-min preferences and “nature” plays against them). Hence, they invest their entire endowment in the long-term bond corresponding to their intended consumption date, regardless of the cost of this bond. These preferences are meant to capture the investment needs of pension funds and insurance companies that invest inelastically in long-term assets.

The resulting demand for long-term assets are met by arbitrageurs – or, in the case of Greenwood, Hanson and Stein (2010), by corporations acting like arbitrageurs. One virtue of this literature is that arbitrageurs look exactly like proprietary trading desks and hedge funds: when long-term rates are high relative to short-term rates, arbitrageurs buy long-term bonds financed by selling and rolling over short-term bonds. This trading entails a risk, because the short-term rate varies exogenously. Since arbitrageurs are risk averse, they charge a risk premium in order to bear this risk. Naturally, this risk premium is endogenously increasing in the absolute value of the net inelastic demand of the preferred-habitat investors. Thus, an increase in the demand from preferred-habitat investors leads to a decrease in the interest rate.\(^2\)

Vayanos and Vila (2009) and Greenwood and Vayanos (2010) use this model to explain the effect of government bond-market trades that lengthen the maturity of the government debt. In particular, Greenwood and Vayanos (2010) study the effect of government purchases of long-term debt financed by purchases of short-term debt. This analysis is motivated by historical episodes such as Operation Twist, a 1962-64 program to shorten the maturity of government debt, and the 2000-2002 buyback of long-term US government debt, both of which were associated with decreases in the long-term interest rate. This analysis is also motivated by recent Fed policy of buying long-term government debt.

The analysis of Vayanos and Vila (2009) and Greenwood and Vayanos (2010) has some-
times been taken to imply that the recent Fed policy of buying long-term government debt should have a negative effect on long-term interest rates; see, for example, Swanson (2011) and Doh (2010).

An example of the use of Vayanos and Vila (2009) to think about the effects of the recent Fed policy of buying long-term government debt is Krishnamurthy and Vissing-Jorgensen (2011), “The Effects of Quantitative Easing on Interest Rates.” In this paper, they write:

Vayanos and Vila (2009) offer a theoretical model for the duration risk channel. The model produces a risk premium on a bond of maturity $t$ that is the product of the duration of a maturity $t$ bond and the aggregate amount of duration risk borne by the marginal bond market investor. By purchasing long-term Treasuries, Agency debt, or Agency MBS, policy can reduce the duration risk in the hands of investors and thereby alter the yield curve, particularly reducing long maturity bond yields relative to short-maturity yields. To deliver these results the model departs from a frictionless asset pricing model. The principal departure that generates the duration risk premium result is the assumption that the bond market is segmented and that there is a subset of risk-averse investors who bear (all or a lot of) the interest rate risk in owning bonds.

My paper studies a Vayanos and Vila (2009) model with preferred-habitat preferences. I show that if profits or losses from a policy of purchasing long-term government debt are passed to taxpayers via lump-sum taxes or transfers, then this policy will not alter the yield curve. This is a version of the standard Wallace irrelevance result, which I show applies here even with the inclusion of Vayanos and Vila (2009) preferred-habitat investors.

The key to this result is that the Fed eventually transfers the profits or losses associated with its balance sheet to the US Treasury. These are in turn transferred (in the model, via lump-sum taxes or transfers) to the general public. Thus, ultimately, the preferred-habitat investors or the arbitrageurs, or some combination thereof, are claimants on the Fed’s profits or losses. Therefore, the preferred-habitat investors or arbitrageurs will choose to exactly undo the effect of the Fed’s open-market operations, for the same reason that underpins Ricardian equivalence or Miller-Modigliani irrelevance. For example, if the Fed transfers its profits or losses from maturity-lengthening open-market operations to the preferred-habitat investors, these preferred-habitat investors will go long short-term debt and short long-term debt, relative to their previous portfolio allocation, so as to exactly undo the effect of the Fed’s trades on their eventual consumption.
This intuition is familiar from the irrelevance papers of Wallace (1981), Chamley and Polemarchakis (1984), and Eggertsson and Woodford (2003). It has recently been emphasized in Curdia and Woodford (2011). What is new here is to show the preferred-habitat preferences in Vayanos and Vila (2009)'s model do not overturn this irrelevance intuition.

This result clarifies the argument in Greenwood and Vayanos (2010) that the preferred-habitat view and Ricardian-equivalence view are “at odds.” Preferred-habitat preferences alone do not break Ricardian equivalence; limited participation or other additional assumptions are required to break Ricardian equivalence in the Vila and Vayanos (2009) model.

Besides limited participation, one way to break Ricardian equivalence is to add government consumption to the model, and to suppose that government consumption is pure waste. Suppose further that the profits or losses from the Fed’s OMO were offset by changing government consumption, rather than through lump-sum taxes or transfers. Then, maturity-lengthening OMO would indeed lead to a decrease in the long-term interest rate. Thus, one interpretation of Greenwood and Vayanos (2010)'s model is that there is an implicit assumption that government spending is wasteful and that government spending changes one-for-one with windfall profits or losses on the Fed’s portfolio.

Another way that Fed purchases of long-term debt might alter the yield curve is if taxation is not lump-sum. However, if distortionary taxation is the mechanism by which long-term debt purchases affect the yield curve, this should be explicitly acknowledged and modeled. Moreover, a mechanism which depends on the distortionary effects of taxation does not seem to capture the “portfolio-balance” channel (see Bernanke, 2010) which policymakers invoke and which Vayanos and Vila (2009) aim to capture.

For these reasons, this paper then turns to limited-participation – prohibitions (due to transactions costs or behavioral motivations) on certain agents trading certain parts of the yield curve. Limited participation can break the irrelevance result by preventing agents from undoing the effects of the OMO. And although it may be hard to justify limited participation in certain parts of the yield curve for insurance companies or pension funds, it is reasonable to assume limited-participation in the long part of the yield curve for some households. Ultimately, what is needed to break the irrelevance result is that some portion

---

3 Greenwood and Vayano (2010), p. 1.
4 Another interpretation is that utility from government spending enters additively into agents’ utility functions and that government spending changes one-for-one with windfall profits or losses on the Fed’s portfolio.
of the profits or losses from the OMO must be borne by agents who cannot perfectly undo these profits or losses through trading on their own.

Including a limited-participation assumption for the preferred-habitat investors in the Vayanos and Vila (2009) model, I show that maturity-lengthening OMO reduce long-term interest rates. It turns out, however, that investors with preferred-habitat preferences are neither necessary nor sufficient for policy to have this effect. I show in a simple model how limited participation alone – in the model, there are households who either do not trade long-term debt or do not participate in financial markets at all, and there are no preferred-habitat preferences – is sufficient for maturity-lengthening OMO to reduce interest rates.

One interpretation of this paper is that this paper spells out a set of conditions under which open-market operations are relevant, or irrelevant, in the Vayanos and Vila (2009) model.

*The Fed’s policy response to the recent crisis.* This paper is also related to the growing literature on the Fed’s policy response to the recent crisis.

While overviews of the Fed’s response can be found in Reis (2009) and elsewhere, I discuss here how the Fed changed the maturity structure of its balance sheet and how these details impact the modelling approach taken in this paper and by others.

The purchase of long-term Treasury and mortgage-backed securities was an important part of the Federal Reserve’s response to the 2007-2009 financial crisis. Between July 2007 and March 2011, the Fed increased its holdings of Treasury securities from $790 billion to $1.3 trillion. Moreover, there was a shift in the composition of its holdings toward longer maturities, as shown in Table 1.

| Table 1. Federal Reserve holdings of Treasury securities by maturity |
|-----------------|----------------|-------------|-------------|-------------|----------------|
|                  | <1             | 1-5 years   | 5-10 years  | >10 years   | Total         |
| Holdings, July 2007 | $399           | $235        | $75         | $82         | $791         |
| Holdings, March 2011 | 100            | 538         | 452         | 176         | 1,266        |
| Increase         | -299           | 304         | 378         | 93          | 476          |
| Percent increase | -75%           | 129%        | 506%        | 113%        | 60%          |

Notes: Amounts in billions of dollars. Data from Federal Reserve.

Over this same period, the Fed increased its holdings of mortgage-backed securities from zero to $950 billion and of agency debt from zero to $143 billion. Essentially all of the
mortgage-backed securities had maturities greater than 10 years.

Fed Chairman Bernanke has argued that these open-market operations should affect long-term interest rates because of a “portfolio balance channel.” In particular, he told August 2010 Jackson Hole conference:

I see the evidence as most favorable to the view that such purchases work primarily through the so-called portfolio balance channel, which holds that once short-term interest rates have reached zero, the Federal Reserve’s purchases of longer-term securities affect financial conditions by changing the quantity and mix of financial assets held by the public. Specifically, the Fed’s strategy relies on the presumption that different financial assets are not perfect substitutes in investors’ portfolios, so that changes in the net supply of an asset available to investors affect its yield and those of broadly similar assets. Thus, our purchases of Treasury, agency debt, and agency MBS likely both reduced the yields on those securities and also pushed investors into holding other assets with similar characteristics, such as credit risk and duration. For example, some investors who sold MBS to the Fed may have replaced them in their portfolios with longer-term, high-quality corporate bonds, depressing the yields on those assets as well.

The logic of the portfolio balance channel implies that the degree of accommodation delivered by the Federal Reserve’s securities purchase program is determined primarily by the quantity and mix of securities the central bank holds or is anticipated to hold at a point in time (the ‘stock view’), rather than by the current pace of new purchases (the ‘flow view’).

It is important to model this view explicitly and understand exactly what elements are needed for this approach to work.

A key institutional feature in this paper’s analysis is that the Fed rolls over its short-term borrowing (daily in the case of interest-bearing reserves) or unwinds part of its long-maturity position and transfers to the U.S. Treasury any net interest income plus realized profits or losses on its portfolio. In the five years prior to the crisis, the Fed transferred between $18.1 and $29.1 billion to the Treasury each year. As the Fed balance sheet expanded, so did its profits and transfers to the Treasury: these totalled $34.6, $31.7, $47.4 and $78.4 billion dollars in the years from 2007 through 2010. Given the Fed’s long-maturity position, it is quite possible that the Fed – and thus the Treasury – will realize losses if the Fed unwinds its position during a period of rising long-term rates. It is the transfer of these profits or losses to the preferred-habitat investors and the arbitrageurs that will lead to the
Ricardian-equivalence result in the model without limited participation.

3.3 Model with preferred-habitat investors

Consider a three-period version of Vayanos and Vila (2009), similar to Greenwood and Vayanos (2010) and Greenwood, Hanson and Stein (2010). Each period is labeled \( t \in \{1, 2, 3\} \). There is a long-term bond that pays out one unit in period 3; the price of the long-term bond in period 1 is \( P \). Short-term bonds are also traded. The interest rate on these short-term bonds is stochastic. The interest rate between periods 1 and 2 is \( R_1 \). This is known in period 1. The interest rate between periods 2 and 3 is \( R_2 \). This is unknown in period 1. In Vayanos and Vila (2009), the short rate “could be determined by the Central Bank and the macro-economic environment,” but they do not model this determination. Later, I will assume that agents have access to a linear technology, such that if they invest \( k_t \) in period \( t \), they will receive \( A_t k_t \) in period \( t + 1 \). Hence \( R_t = A_t \).

The period-1 good is the numeraire. I sometimes refer to the units of period-1 good as dollars, but there is no liquidity or transactions value associated with this term.

There are three actors: arbitrageurs, preferred-habitat investors, and the government. The arbitrageurs consume only in period 3. They have mean-variance preferences.\(^5\)

Preferred-habitat investors have \( L \) units of wealth in period 1. As in Vayanos and Vila (2009), they can trade short-term and long-term bonds; denote their holdings of long-term bonds by \( k \). The preferred-habitat investors have min-max preferences (i.e., infinite risk aversion). In particular, preferred-habitat investors’ demand function for long-term bonds is given by:

\[
k(P) \in \max_k \min_{R_2} u(k + (L - Pk)R_1R_2)
\]

Hence, preferred-habitat investors demand \( k(P) = \frac{L}{P} \).\(^6\)

\(^5\)This could be micro-founded by assuming that arbitrageurs have CARA preferences and that \( A_2 \) is normally distributed. This of course comes with the possibility of negative consumption and productivity.

\(^6\)In the Vayanos and Vila (2009) literature, the preferred-habitat investors demand \( L \) units of the long-term bond, where \( L \) does not depend on the price. This is motivated by assuming that preferred-habitat investors are infinitely risk averse consume only at period 3. Of course, as Vayanos and Vila (2009) note, if preferred-habitat investors have endowment \( L \), their preferences and budget constraint would imply that preferred-habitat investors demand \( \frac{L}{P} \) units of the long-term bond, but Vayanos and Vila (2009) ignore this to simplify their already-complex continuous-time analysis. Here, since I am interested in the optimizing behavior of the preferred-habitat investors in response to government policy, I will hold true to Vayanos and Vila (2009)’s preferences and budget constraint for preferred-habitat investors and hence the preferred-
The government supplies $G$ units of the long-term bond. For simplicity, assume that these bonds are backed by some asset that government owns, with a period-3 pay out of $G$.

The arbitrageur buys $h \in \mathbb{R}$ long-term bonds, financing this purchase with short-term debt, which is rolled over at the stochastic interest rate $R_2$. The arbitrageur has zero initial wealth. Hence, the arbitrageurs’ terminal wealth is

$$w = h[1 - PR_1R_2]$$

with mean $h[1 - PR_1E[R_2]]$ and variance $h^2P^2R_1^2Var[R_2]$. Optimization implies

$$h(P) \in \max_h h[1 - PR_1E[R_2]] - \frac{\gamma}{2}h^2P^2R_1^2Var[R_2]$$

The first-order condition is:

$$\frac{1}{P} - R_1E[R_2] = \gamma h(P)PR_1^2Var[R_2]$$

or

$$h(P) = \frac{\frac{1}{P} - R_1E[R_2]}{\gamma R_1^2Var[R_2]}$$

Market-clearing for the long-term bond in period-1 requires

$$h(P^*) = G - \frac{L}{P^*}$$

This implies

$$\frac{1}{P^*} - R_1E[R_2] = \gamma (GP^* - L)R_1^2Var[R_2] \quad (3.1)$$

The left-hand-side is the definition of the risk-premium associated with buying one dollar of long-term debt and financing this with short-term debt. The right-hand-side is the equilibrium value of this risk premium necessary for the arbitrageurs to bear the duration risk that results from the equilibrium net supply of long-term bonds, $G - \frac{L}{P^*}$, from the government and preferred-habitat investors.

Habitat investors will demand $\frac{L}{P^*}$ long-term bonds. I will also consider alternative preferences giving rise to preferred-habitat behavior, in which preferred-habitat investors demand $L$ units of the long-term bond; I will show that my analysis applies in this case as well.
Note that this is a quadratic in $P^*$. Hence, $P^*$ may not be uniquely defined. In a moment, this will be addressed by assuming that $G = 0$. However, this assumption is not needed for the following result:

**Lemma 31 (Vayanos and Vila; Greenwood, Hanson and Stein)** The expectations hypothesis holds (that is, $\frac{1}{P^*} = R_1E[R_2]$) if and only if one or more of the following conditions holds: (i) $\gamma = 0$; (ii) $\text{Var}[R_2] = 0$; (iii) $G = LR_1E[R_2]$.

Lemma 1 follows directly from (3.1). Lemma 1 says that there is no risk premium for holding duration risk if and only if one of the following conditions holds: (i) arbitrageurs are risk-neutral; (ii) there is no duration risk associated with a portfolio that is long long-term bonds and short short-term bonds; or (iii) there is no net supply of duration risk that needs to be held for markets to clear.

For the remainder of the paper, I assume that $G = 0$. The next result is the key preferred-habitat result: when the preferred-habitat investors’ demand for the long-term bond increases, its price increases.

**Lemma 32** $P^*$ is increasing in $L$, the endowment of the preferred-habitat investors.

Lemma 2 also follows directly from (3.1). When the endowment $L$ of the preferred-habitat investors increases, their demand curve for the long-term bond, $k(P) = \frac{L}{P}$, shifts to the right. In order for the arbitrageurs to accommodate this increased demand, the risk premium charged to hold short-term debt and finance this by selling long-term debt increases. Hence the risk premium on long-term debt decreases, and its price increases.

Now, consider the policy experiment featured in Greenwood and Vayanos: the government buys some long-term debt, and finances this purchase by selling short-term debt. The crucial question is: who bears the resulting profits or losses on the government’s position? Greenwood and Vayanos do not address this question: it is as if they assume that the profits or losses disappear into thin air. While this assumption may be appropriate for some

---

7For tractability, I could have alternatively assumed that the government supplies long-term bonds in an amount equal to $g$ dollars, but this assumption moves away from an economy specified in terms of endowments and preferences of the consumers. The important question to ask when we think about an increase in government debt is: what backs up the debt? If it is an endowment of the government – an asset that pays out in a given period, as assumed here – than an increase in government debt is tantamount to a change in endowments. If it is backed up by lump-sum taxes, than a Ricardian-equivalence irrelevance result will apply, as the remaining analysis shows.
settings, it does not seem to reflect what happens to net interest income and capital gains or losses from the Fed’s portfolio, which are transferred to the Treasury.

**Definition 33** Define a maturity-lengthening open-market operation as follows. The government purchases $z$ dollars of long-term debt and finances it by selling short-term debt. The government redeems share $\phi$ of the short-term debt in period 2, and rolls over share $(1 - \phi)$. To redeem the short-term debt in period 2, the government raises $\phi R_1 z$ from lump-sum taxes. In particular, the arbs bear $\alpha_2 \in [0, 1]$ share of the lump-sum taxes, and the preferred-habitat investors bear share $(1 - \alpha_2)$. In period 3, the government receives a payout $\frac{z P}{R}$ from its long-term holdings and requires $(1 - \phi) R_1 R_2 z$ to redeem the rolled-over short-term debt. The government transfers $\alpha_3 \in [0, 1]$ share of the net payout $\frac{z P}{R} - (1 - \phi) R_1 R_2 z$ to the arbs and $(1 - \alpha_3)$ share to the preferred-habitat investors; if the net payout is negative, this transfer takes the form of lump-sum taxes.

A maturity-lengthening open-market operation reflects a policy (or a model) without government consumption and with proportional division of government taxes and transfers, in which this proportion can change over time. In the case of the Fed’s recent policy actions, it seems reasonable to model this proportion as unchanging ($\alpha_2 = \alpha_3$), but this is not necessary for the next result. This policy is referred to as a maturity-lengthening operation because it lengthens the maturity of the Fed’s assets; it shortens the maturity of debt directly in the hands of the public. The aim of maturity-lengthening open-market operations is to increase the price of long-term debt.

Of course, this definition of the Fed’s open-market operations also includes an assumption that the government can lump-sum tax and that the net interest income and realized capital gains or losses on the Fed’s balance sheet are divided according to some (potentially time-varying) proportions across the arbitrageurs and the preferred-habitat investors. These assumptions seem to be a good starting point because the mechanism by which OMO are supposed to work in a preferred-habitat or “portfolio balance” setting have not been articulated to rely on the distortionary effects of taxation; indeed, the preferred-habitat literature has ignored the role of taxation completely.

The next result is the key irrelevance result of this part of the paper:

**Proposition 34** In the Vayanos and Vila preferred-habitat model, maturity-shortening open-market operations have no effect on equilibrium prices.
The key intuition is this: The government takes on some duration risk through OMO, and passes some proportion of the risk onto arbitrageurs and some proportion onto preferred-habitat investors. If arbitrageurs and preferred-habitat investors did not alter their demand curves in response to an increase in OMO, an increase in OMO would decrease the risk premium for long-term bonds and raise the price. However, arbitrageurs and preferred-habitat investors will change their demand curves, to undo the effect of the OMO on their consumption path. This is simplest to explain for the preferred-habitat investors: they are infinitely risk averse and do want to bear any risk. So if they bear risk on the government’s portfolio through the tax system, and if they have instruments to undo this risk, they will undo this risk, no matter the price. Similarly, if at price P, arbitrageurs are willing to demand \( h(P) \) units of long-term debt funded by short-term debt, and if the government is implicitly making them claimants on the profits from an additional unit of long-term debt funded by short-term debt, the arbitrageurs will reduce their demand for long-term debt funded by short-term debt by one unit. Hence the OMO has no effect.

One way to break this irrelevance result is to assume that the preferred-habitat investors cannot trade short-term bonds; with this assumption, the preferred-habitat investors cannot undo the supply and demand effects of the Fed’s purchases. Although this may seem a stark assumption, the purpose here is illustrate the type of assumption needed for OMO to affect long-term interest rates; as the previous proposition showed, preferred-habitat preferences are not enough to do so.

**Proposition 35** If preferred-habitat investors cannot trade the short-term bond, maturity-shortening OMO has an effect on the price of the long-term bond if and only if
\[
\alpha_2 \phi + \alpha_3 (1 - \phi) \neq 1.
\]
That is, if preferred-habitat investors are exposed to part of the risk on the government’s balance sheet, then and only then, \( P \) is increasing in \( z \).

We can use this model, with limited participation, to study how the distribution of the arbitrageurs’ terminal wealth is related to \( z \), the dollar amount of OMO. In particular, it is interesting to study the covariance of arbitrageurs’ terminal wealth with the technology shock, \( A_2 \). I will show that the arbitrageur does poorly precisely when the technology shock is positive.

With limited-participation by the preferred habitat investors, we can write the terminal wealth of the arbitrageurs as
\[ w = d^i[R_1(R_2 - E[R_2]) + \gamma R_1^2 Var[R_2]d^i] \]
\[ + \frac{z}{T} [\alpha_3 - \alpha_2] \phi \]

where
\[ d^i = L + z[1 - (\alpha_2 \phi + \alpha_3(1 - \phi))] \]

is the inelastic demand for long-term bonds in dollars. In equilibrium, the arbitrageurs must absorb the inelastic demand \( d^i \). For each unit of long-term bond absorbed by the arbitrageurs, they are exposed to the following zero-expected-value risk \( R_1(R_2 - E[R_2]) \). To bear this risk, they demand a per-unit risk premium equal to \( \gamma R_1^2 Var[R_2]d^i \). This per-unit risk premium is linearly increasing in the amount of inelastic demand they must meet through short selling; the total risk premium earned is hence quadratic in the total inelastic demand they must meet. In addition, if the proportion of per-period government taxes borne by the arbitrageurs is changing over time, they earn an implicit transfer equal to \( \frac{z}{T} [\alpha_3 - \alpha_2] \phi \). The next Lemma follows directly from the previous analysis.

**Lemma 36** Suppose preferred-habitat investors cannot trade the short-term bond. Then the covariance of arbitrageurs’ terminal wealth \( w \) and productivity \( A_2 \) is given by \( d^i R_1 Var[R_2] \) and is linearly increasing in value \( z \) of long-term bonds purchased by the Fed via OMO.

It is also the case that the variance of arbitrageurs’ terminal wealth is increasing in \( z \) and, further supposing that \( \alpha_2 = \alpha_3 \), the expectation of arbitrageurs’ terminal wealth is increasing in \( z \). This is because an increase in \( z \) requires arbitrageurs to have a larger short position in long-term bonds. This implies more risk (hence the increase in variance of terminal wealth) and requires a larger risk premium (hence the increase in expected terminal wealth). However, these results about the expectation and variance of terminal wealth depend on the assumption that there is no supply of long-term bonds from outside the model. If, for example, some new agent inelastically supplied \( y \) dollars of the long-term bond, with \( y > L + z[1 - (\alpha_2 \phi + \alpha_3(1 - \phi))] \), then an increase in \( z \) would lead to a decrease in the expectation and variance of terminal wealth. In contrast, the result that covariance of \( w \) and \( A_2 \) is increasing in \( z \) is robust to the introduction of a new agent inelastically supplying \( y \) dollars of the short-term bond.
3.4 Limited participation without preferred-habitat preferences

In the previous discussion, it was shown that adding limited-participation (a restriction on the ability of some agents to trade certain parts of the yield curve) to a model with Vayanos and Vila (2009) preferred-habitat investors leads to maturity-lengthening OMO having a negative effect on the long-term interest rate. In this model, what was crucial for the relevance of the OMO was not the agents with preferred-habitat preferences, but the assumption of limited participation.

With this in mind, I formulate here a new model of OMO with limited participation, and without preferred-habitat preferences. This brings limited participation front and center, showing that preferred-habitat preferences are neither necessary nor sufficient for the relevance of OMO.

This model also allows for a more realistic form of limited participation. In the previous model, I assumed that preferred-habitat investors (insurance companies and pension funds) could not invest in short-term assets. Here, instead, I will suppose that there are some households who invest only in short-term assets and do not trade every period. At least anecdotally, there do appear to be many households who invest inelastically in short-term instruments, like savings accounts and money-market funds. There also appear to be households who do not save at all or who participate in financial markets only sporadically. These phenomena may be due to technological or cognitive transactions costs. Regardless, assuming some households cannot invest in long-term bonds seems more appealing than assuming that insurance companies cannot invest in short-term bonds.

One might question whether there is a meaningful difference between my limited-participation investors and the Vayanos and Vila (2009) preferred-habitat investors: after all, both investors, in the absence of open-market operations, appear to be investing in a “preferred” part of the yield curve. The key difference is that limited-participation cannot invest in certain or all assets, by assumption, whereas the Vayanos and Vila (2009) preferred-habitat investors choose not to invest in short-term assets. However, the Vayanos and Vila (2009) preferred-habitat investors will want to – and will – invest in short-term assets in the presence of open-market operations; limited-participation investors will want to – but, by assumption, cannot – invest in certain instruments in the presence of open-market...
3.4.1 Two-period version

Consider the following very simple model. There are two periods, \( t \in \{1, 2\} \), and two types of households. The first type, whom I will call arbitrageurs, can trade a one-period bond; the name arbitrageurs is fairly arbitrary, but it is meant to reflect that these households are present in the market and trade with the government. The second type cannot trade the bond; this second type of household simply does not participate in financial markets. Each household receives an endowment of \( e_t \) in period \( t \).

The arbitrageurs have access to a private-savings technology \( f(k) \); I will assume that \( f(k) \) satisfies the standard neoclassical properties. If instead \( f(k) = 0 \), we have an endowment economy. There is no uncertainty in this economy.

Suppose in this economy, the government performs the following OMO: it buys one-period bonds worth \( x \) period-1 goods and finances this by lump-sum taxes. I assume that the period-1 lump-sum taxes are borne exclusively by the non-trading households, and that the non-trading households are the exclusive recipients of the period-2 flows from government security holdings. I could also have assumed that the lump-sum taxes and transfers are divided across the two types, but this would not change the qualitative results. The only important feature is that some of the lump-sum tax be borne by the non-trading household.

Market-clearing implies that consumption for the arbitrageurs in period 1 and 2 satisfy:

\[
    c_{1a} = e_1 + x - k \\
    c_{2a} = e_2 - \frac{1}{p_2} x + f(k)
\]

Optimization for the arbitrageur further implies:

\[
    k(p_2) = f'^{-1}(\frac{1}{p_2}) \\
    p_2 = \beta \frac{u'(c_{2a})}{u'(c_{1a})} = \beta \frac{u'(e_2 - \frac{1}{p_2} x + f(f'^{-1}(\frac{1}{p_2})))}{u'(e_1 + x - f'^{-1}(\frac{1}{p_2}))}
\]

Denote the equilibrium bond price as a function of \( x \) by \( p_2(x) \).
Lemma 37 \( p_2(x) > p_2(0) \) and \( k_2(x) > k_2(0) \) if and only if \( x > 0 \).

The intuition is that, relative to an absence of government intervention, maturity-lengthening open-market operations \( x \) result in the arbitrageurs selling long-term debt to the government and buying short-term debt from the government; this shifts period 2 consumption down and period 1 consumption up. Of course, the open-market operations also have an effect on how much the arbitrageurs invest in the private-savings technology. However, if the price of the long-term bond were to weakly decrease, optimal investment (3.2) would decrease, and consumption would be even further tilted toward period 1 and away from period 2. This is not consistent with a weak decrease in the price.

Is \( c_2(x) \) greater than \( c_2(0) \)? This depends on a combination of income and substitution effects. On one hand, OMO \( x \) represents a transfer to the arbitrageurs, because they sell to the government at a high price. One the other hand, OMO \( x \) generates an increase in the price of period-2 consumption. Whether the income and substitution effects dominate depends on the particular assumptions about preferences, endowments and technologies.

3.4.2 A model with wealth accumulation

This section develops a three-period version of the previous model, in order to study the effect of maturity-lengthening OMO on wealth accumulation in a limited-participation model.

One advantage of the three-period version is that the limited participation assumption can take a less restrictive form: rather than assuming that some households do not participate in financial markets at all, I will assume only that some households do not trade long-term debt and do not trade in every period. Like the above two-period limited-participation model without preferred-habitat preferences, this three-period model is motivated by the observation that limited-participation, rather than preferred-habitat preferences, are crucial for the relevance of OMO, as shown in Section 3.

In the model, there are households and arbitrageurs. The arbitrageurs can trade short-term and long-term bonds. So can some of the households. But there are some households that trade only in period 1, and only trade the short-term bond.\(^8\) I label the

---

\(^8\)An alternative setup might be developed along the following lines. There are three types of households. The arbitrageurs can trade in all three periods. A second and third type of households can only trade one-period debt. The second type trades only during the first period; in the second period, he does not participate in financial markets. The third type trades only during the second period; in the first period,
limited-participation households Type L (for limited), the arbitrageurs Type A, and the full-participation households Type F (for full).

For simplicity, I assume that the profits or losses from the government portfolio flow through lump-sum taxes or transfers only to the limited-participation households, but what is important is only that some of profits or losses flow to the limited-participation households.

This model allows us to study wealth accumulation by the arbitrageurs, how this wealth accumulation is affected by government bond purchases, and the resulting implications for the vulnerability of the arbitrageurs to an endowment shock.

3.4.3 Benchmark: Ricardian equivalence

Consider an environment in which there is only one type of agent, households. There are three periods, \( t = \{1, 2, 3\} \). In each period, the endowment of households is \( e_t \). The households have total mass equal to one, and each household is infinitesimal. The households trade securities that pay off in a given date. Security \( b_t \) is a claim to one good in state \( t \), and period-1 goods are the numeraire. There is no uncertainty.

In this environment, households choose a portfolio \( \{b_t\} \) to maximize their utility

\[
u(b_1) + \beta u(b_2) + \beta^2 u(b_3)
\]

subject to the budget constraint

\[
b_1 + p_2 b_2 + p_3 b_3 \leq e_1 + p_2 e_2 + p_3 e_3
\]

Market clearing implies that there is no trade: \( b_t = e_t \), for all \( t \). Combining this with the first-order conditions of the household’s problem, the following prices are obtained

\[
p_t = \beta^{t} \frac{u'(e_t)}{u'(e_1)}
\]

for \( t \in \{2, 3\} \).

In this environment, I now introduce a government that buys long-term (i.e., period 3) securities and finances these purchases by selling short-term (i.e., period 2) securities.

he does not participate in financial markets.
Specifically, the government buys long-term debt worth $x$ and sells short-term debt worth $x$. In period 3, the government is repaid $x \frac{1}{p_3}$ and rebates this to households. In period 2, the government pays $x \frac{1}{p_2}$ to bond holders and raises this amount in lump-sum taxes.

**Lemma 38 (Ricardian Equivalence)** With complete markets, market prices and allocations do not depend on $x$, the amount of government long-term asset purchases.

Thus, in this complete-markets setting, the government cannot affect the term structure of interest rates through purchases of long-term assets financed by selling short-term assets. This is of course a case of Ricardian equivalence, similar to Proposition 4, which addressed the preferred-habitats case; in the monetary literature, these Ricardian equivalence results are similar to Wallace (1981), Chamley and Polemarchakis (1984), and Eggertsson and Woodford (2003).

### 3.4.4 Adding limited participation

Essentially, to break the Ricardian equivalence of the earlier setup, some type of market incompleteness has to be introduced that prevents the agents from undoing the effects of government policy.

Here, I introduce limited participation. The arbitrageurs and the non-limited households trade in the securities for all three dates. The limited-participation households, in contrast, trade in the securities only for dates 1 and 2; moreover, it is the limited-participation households who are taxed to redeem the government’s obligations and these same households who receive the proceeds of the government’s investments. I assume that each agent has endowment $e_t$ and that the measure of arbitrageurs, non-limited households and limited households is $(\frac{1}{7}, \frac{1}{7}, 1)$.

In this section (4.4), the arbitrageur and the non-limited household are identical except in name, so I will subsume the non-limited households into the arbitrageurs.

How is Ricardian equivalence broken in this model? When the government purchases long-term securities, the only investors who can be on the other side of the trade are the arbitrageurs and the non-limited households (Types A and F). As a result, mechanically, purchases of long-term securities by the government reduce the period-3 consumption of the arbitrageurs and non-limited households (who hold fewer claims to period-3 consumption)
and increase the period-3 consumption of the limited-participation households (who are rebated the proceeds of the period-3 claims purchased by the government). The arbitrageurs and households have endowments such that, were there complete markets and no government asset purchases, there would be no trade. Hence, with government asset purchases, the arbitrageurs and households would like to trade to undo the effect of the asset purchases on their consumption allocations. However, the limited-participation households cannot trade period-3 assets, by assumption. Hence, to induce the arbitrageurs and non-limited households to sell period-3 assets to the government, the price for period-3 assets rises (i.e., the long-term interest rate falls).

The problem of the arbitrageur is:

$$\max u(b_{1a}) + \beta u(b_{2a}) + \beta^2 u(b_{3a})$$

subject to

$$b_{1a} + p_2b_{2a} + p_3b_{3a} \leq e_1 + p_2e_2 + p_3e_3$$

The limited-participation household’s problem can be written as

$$\max u(b_{1l}) + \beta u(c_{2l}) \quad (3.3)$$

subject to

$$b_{1l} + p_2c_{2l} \leq e_1 + p_2e_2 - x \quad (3.4)$$

where $c_{2l} = b_{2l} - \frac{1}{p_2}x$ is the household’s period-2 consumption.

The market-clearing conditions are

$$b_{3a} = e_3 - \frac{1}{p_3}x$$
$$b_{2a} + b_{2l} = 2e_2 + \frac{1}{p_2}x$$
$$b_{1a} + b_{1l} = 2e_1$$

Using the period-3 market-clearing condition $b_{3a} = e_3 - \frac{1}{p_3}x$, it can be seen that the arbitrageur’s optimal choice for period 1 and 2 security purchases is the solution to

$$\max u(b_{1a}) + \beta u(b_{2a}) \quad (3.5)$$
subject to
\[ b_{1a} + p_2 b_{2a} \leq e_1 + p_2 e_2 + x \quad (3.6) \]

Thus, conditional on the price of period 2 goods (which may vary with government policy \( x \)), the period-1 and period-2 allocations for the limited-participation household and the arbitrageur are the same as if the agents lived only for 2 periods and the limited-participation household received a lump-sum tax of \( x \) and the arbitrageur received a lump-sum subsidy of \( x \).\footnote{Note that the market-clearing condition for period 2 can be re-written \( b_{2b} + c_{2a} = 2e_2 \).} This subsidy and tax reflect that the government policy essentially forces the limited-participation households to buy period-3 consumption from the arbitrageurs. The arbitrageurs sell period-3 assets to the government, and use the proceeds to buy consumption in periods 1 and 2 from the limited-participation households; the limited-participation households sell period 1 and 2 goods to the arbitrageur to finance the limited-participation households’ period 2 taxes, which the government uses to buy period-3 consumption for these households from the arbitrageur.

Thus, the period-1 and period-2 equilibrium prices and allocations do not depend on \( e_3 \) and depend on \( x \) only through the implicit subsidy as reflected in budget constraints (3.4) and (3.6). This leads to the following result, which is similar to Lemma 7. It shows that in this limited-participation setting, maturity-lengthening open-market operations decrease the long-term interest rate. This is because maturity-lengthening OMO in the presence of limited-participation households force the arbitrageurs to consume less in period 3 and more in periods 1 and 2; the limited-participation households would want to sell period-3 securities and buy period-2 securities, offsetting the open-market operations, but by assumption they cannot.

**Lemma 39** With government intervention \((x > 0)\), the equilibrium price of period-3 securities is higher than the equilibrium price without government intervention \((x = 0)\).

In much of the remainder of this paper, I will focus on the case of agents with constant elasticity of intertemporal substitution \( \theta \), for tractability. However, for the next lemma, constant elasticity of intertemporal substitution is only one of the sufficient conditions for the results. This lemma says that, under certain conditions, long-term interest rates are strictly decreasing in open-market operations \( x \) and that open-market operations do not affect short-term rates.
Lemma 40 Suppose $e_1 = e_2$ or that agents have constant elasticity of intertemporal substitution. Then:

(i) $p_3$ is strictly increasing in $x$ and $p_2 = \beta \mu'(e_2) / u'(e_1)$ does not depend on $x$; and

(ii) $b_{1a}$ and $b_{2a}$ are strictly increasing in $x$, and $b_{1l}$ and $c_{2l}$ are strictly decreasing in $x$.

This lemma also shows that, under constant elasticity of intertemporal substitution or constant period-1 and period-2 endowments, the period-1 and period-2 allocations are exactly the same as if the government were to provide a lump-sum subsidy of $x$ to the arbitrageurs and a lump-sum tax of $x$ for households and there were no period 3. Before, without the assumption about constant elasticity of intertemporal substitution or equality of endowments over period 1 and 2, the later statement could be made if the statement “conditional on the price of period-2 assets” were added. With CRRA preferences or constant period 1 and 2 endowments, this conditioning statement is no longer needed, because $p_2$ does not depend on government intervention $x$.

Thus, we see that with this limited-participation setup, government maturity transformation can depress the long-term interest rate. Under constant endowments or constant elasticity of intertemporal substitution, “Operation Twist” twists the yield curve not by raising short-term rates and lowering long-term rates, but lowering long-term rates and leaving short-term rates unaffected. This example shows how government maturity transformation can achieve its goal of lowering the long-term interest rate in a limited-participation setup, in contrast to the irrelevance of government maturity transformation that obtained under complete markets.

With an additional assumption about preferences – a restriction to constant elasticity of intertemporal substitution – equilibrium allocations and prices can be further characterized. This characterization will be helpful for studying how open-market operations change the consumption plan and wealth accumulation of the agents.

Lemma 41 Suppose that agents have constant elasticity of intertemporal substitution $\theta$. Then period 3 equilibrium prices and allocations can be characterized as follows:

\[
p_3 = \beta^2 (e_3 - \frac{1}{p_3} x)^{-\frac{1}{\theta}} (\frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\theta})^{-\theta}})^{\frac{1}{\theta}}
\]

\[
b_{3a} = e_3 - \frac{1}{p_3} x
\]

\[(3.7)\]

\[(3.8)\]
In accordance with Lemma 10, the equilibrium prices for constant elasticity of intertemporal substitution preferences satisfy \( \frac{dp_3}{dx} = 0 \) and \( \frac{dp_3}{dx} > 0 \), and the equilibrium allocations satisfy \( \frac{db_{3a}}{dx} > 0 \) and \( \frac{db_{3a}}{dx} > 0 \).\(^{10}\)

In particular, how planned period-3 consumption for the arbitrageur changes with open-market operations \( x \) can be studied by examining (3.7) and (3.8). The next lemma shows that if \( \theta \geq 1 \), an increase in open-market operations decreases period-3 consumption for the arbitrageur: the substitution effect dominates the income effect. However, it is possible that period-3 consumption increases for the arbitrageur in response to an increase in \( x \). Intuitively, this occurs if and only if the price elasticity of the period-3 bond with respect to open-market operations \( x \) is greater than unit. And, as the next lemma states, this occurs if \( \theta < 1 \) and open-market operations are sufficiently large relative to endowments.

The empirical evidence on the elasticity of intertemporal substitution suggests a value for \( \theta \) around or above one (Angeletos 2007).

**Lemma 42** The arbitrageurs’ equilibrium holding of long-term bonds, \( b_{3a} \), is increasing in \( OMO \) \( x \) if the elasticity of intertemporal substitution \( \theta \) is greater than or equal to one. If \( \theta < 1 \), then

\[
\frac{\partial b_{3a}}{\partial x} =
\begin{cases}
< 0 & \text{if } x < \frac{e_1 + p_2 e_2}{\theta - 1} \\
= 0 & \text{if } x = \frac{e_1 + p_2 e_2}{\theta - 1} \\
> 0 & \text{if } x > \frac{e_1 + p_2 e_2}{\theta - 1}
\end{cases}
\]

Again, this is useful not only for understanding how open-market operations \( x \) affect long-term interest rates, but also for understanding how open-market operations affect planned consumption paths, which will be important for the following discussion of vulnerability to endowment shocks.

### 3.4.5 Government asset purchases, wealth accumulation, and shocks

In this section, I theoretically evaluate whether an increase in long-term asset purchases by the government, in the context of this model, can increase the vulnerability of the arbitrageurs to adverse shocks.

---

\(^{10}\)Prices and allocations for periods 1 and 2 and, for the case of log utility, closed-form solutions for \( p_3 \) and \( b_{3a} \), are given in Appendix B.
In particular, I study how the economy responds to a low endowment realization for the arbitrageurs in period 2.

I analyze two types of shocks: shocks that could not have been anticipated by the agents (shocks that, from the agents’ ex-ante point of view, have zero probability) and shocks that agents know might occur. The former type of shock is more interesting to study in this model, because it can capture the notion that the arbitrageurs tilt their consumption profile downward in response to OMO, and then, in the event of an unanticipated shock, they have fewer assets with which to buffer this shock. The zero-probability shock corresponds to the notion that there are some shocks on which market-participants put zero weight, even though an outsider (such as the government) or the true probability measure might put positive weight on them. Of course, a welfare analysis under such an assumption is tricky, but here the focus is a positive characterization.

I show that under certain assumptions, the income effect from OMO dominates the substitution effect, and OMO make arbitrageurs less vulnerable to a zero-probability shock. With a shock that can be anticipated, the income and substitution effects push in the same direction, toward higher period-2 arbitrageur consumption in the event of a shock.

A loose interpretation is that the Federal Reserve may worry that, by reducing long-term interest-rates and making wealth-accumulation more expensive, the arbitrageurs may become more vulnerable to a shock in the medium-run and that political or other constraints may prevent further intervention at this date.

In the case of a shock that cannot be anticipated, the pre-shock equilibrium has the following features. Government long-term asset purchases force the arbitrageurs to have low period-3 consumption\textsuperscript{11}. The arbitrageurs then use the wealth they gain from selling period-3 assets to boost consumption in periods 1 and 2; and the arbitrageurs spread the wealth across the two periods, by trading with the households.

Now suppose a zero-probability shock hits at period 2, in the form of an unexpectedly low period-2 endowment for the arbitrageurs. The arbitrageurs have already used part of their wealth from selling period-3 consumption to purchase period-1 consumption, and this wealth is now gone. They will still sell some of their period-3 assets to the non-limited

\textsuperscript{11}This assumes the amount of asset purchases is less than a threshold defined below. If the asset purchases are greater than this threshold, it is possible that period-3 consumption for the specialists is increasing in the amount of government long-term asset purchases, because the elasticity of the price of period-3 assets with respect to the amount of government purchases is greater than one. This is discussed more below.
households, in return for period-2 consumption. Nonetheless, they have less period-3 assets to sell than if there were no government intervention.

Does this imply that period-2 consumption in the event of a zero-probability shock is decreasing in the size of government purchases of long-term assets during period 1? I show that under certain conditions – such as intertemporal elasticity of substitution equal to one, or intertemporal elasticity of substitution less than one and sufficiently small government purchases – the answer is no, because the income effect dominates the substitution effect.

The intuition, for intertemporal elasticity of substitution equal to one, is that the arbitrageurs, in the event of a zero-probability shock, sell some period-3 assets for period-2 assets to offset the shock – but how much the arbitrageurs offset the shock is independent of their holdings of period-3 assets. Hence, because period-1 government purchases of long-term assets increase the arbitrageurs’ initial period-2 holdings, these government purchases increase period-2 consumption in the event of a zero-probability shock, and hence do not add to the volatility of the arbitrageurs’ consumption. The intuition for intertemporal elasticity of substitution less than one is similar.

What about intertemporal elasticity of substitution greater than one? Here, an increase in government purchases unambiguously reduces the arbitrageurs’ initial period-3 asset holdings and, in the event of a shock, the amount that the arbitrageurs offset the shock (by selling period-3 assets) is increasing in the amount of initial period-3 assets. This mechanism suggests that government purchases increase the volatility of period-2 arbitrageur consumption. However, it is still possible that an increase in government purchases increases period-2 consumption in the event of a shock. The reason is that, with government purchases of long-term assets during period 1, the financial arbitrageurs sell their period-3 consumption for a high price. The arbitrageurs are selling, and the households are (implicitly) buying, at a high price. The resulting portfolio allocations can be such that period-2 consumption in the event of a shock is increasing in period-1 government purchases, even though the arbitrageurs already allocated some of their period-3 wealth to period-1 consumption as a result of the government purchases.

I continue with the assumption that each agent puts a probability of one on an endowment \((e_1, e_2, e_3)\) for herself and for all other agents. Under this assumption, the equilibrium prices \((p_2, p_3)\) and allocations \((b_{1l}, b_{1a}, c_{2l}, b_{2a}, b_{3a})\) will be same as given in Lemma 11. In particular, the non-limited households will choose the same allocation as the arbitrageurs.
I then study what happens if, during period 2, everyone finds out that the arbitrageurs have a lower endowment than anticipated. In particular, the new period-2 endowment of the arbitrageurs is $e_2 - \delta \geq 0$. Since, in equilibrium $b_{2a} \geq e_2$, there is no risk (on the equilibrium path) that it will be infeasible for the arbitrageurs to honor their obligations.

At this point, the arbitrageurs and the third type will trade, as the arbitrageur would like to offset the shock by exchanging period-3 consumption for period-2 consumption. The households will not trade, as they cannot trade period-3 securities.

In period 2, with this shock, the setup is equivalent to a two-period economy in which the arbitrageurs have endowment $(b_{2a} - \delta, b_{3a})$ and the third type has endowment $(b_{2a}, b_{3a})$. Denote the new asset positions and consumption allocations of the arbitrageur and the third type, respectively, as $(s_{2a}, s_{3a})$ and $(s_{2f}, s_{3f})$. Denote the new period-3 price as $p_{3s}$.

This problem is identical to the period 1 and 2 problem solved in Lemma 11, except that the endowments and budget adjustments have changed. Using those results, we can write

$$ s_{2a} = \frac{b_{2a} + p_{3s}b_{3a} - \delta}{1 + p_{3s}(-\frac{b_{3a}}{b_{3s}})^{\theta}} \quad (3.9) $$

and

$$ p_{3s} = \beta \left( \frac{b_{2a} - \delta}{b_{3a}} \right)^{\frac{1}{\theta}} \quad (3.10) $$

The question of interest is how period-2 arbitrageur consumption $s_{2a}$ varies with period-1 government long-term asset purchases $x$. To understand this as intuitively as possible, I first study how $s_{2a}$ varies with $b_{2a}$ and $b_{3a}$, because we know that government long-term asset purchases increase the former and how they affect the latter from Lemma 10 and 12, respectively.

**Lemma 43** $s_{2a}$ is strictly increasing in $b_{3a}$ if and only if $\theta > 1$:

$$ \frac{\partial s_{2a}}{\partial b_{3a}} > 0 \quad \text{if } \theta > 1 $$

$$ \frac{\partial s_{2a}}{\partial b_{3a}} = 0 \quad \text{if } \theta = 1 $$

$$ \frac{\partial s_{2a}}{\partial b_{3a}} < 0 \quad \text{if } \theta < 1 $$

$s_{2a}$ is strictly increasing in $b_{2a}$.

Combining Lemmas 11 through 13 immediately yields the following result.

---

\[12\] In Lemma 11, the budget adjustments were a lump-sum subsidy of $x$ for the specialists and a lump-sum tax of $x$ for the households, and here the budget adjustment is a lump-sum loss of $x$ for the specialists.
Proposition 44 The arbitrageur’s period-2 consumption in the event of a low endowment realization, $s_{2a}$, is increasing in period-1 government long-term asset purchases, $x$, if: (i) $\theta = 1$; or (ii) if $\theta < 1$ and $x < \frac{e_1 + p e_2 x}{1 - \beta}$. 

The intuition is that, under conditions (i) or (ii), an increase in government intervention leads to an increase in planned period-2 consumption, $b_{2a}$, and a decrease in planned period-3 consumption, $b_{3a}$, according to Lemmas 11 and 12, respectively. Moreover, the decrease in $b_{3a}$ does not effect $s_{2a}$ under condition (i) and serves to increase $s_{2a}$ under condition (ii).

For the special case of $\theta = 1$, the equilibrium allocations can be solved explicitly, with some algebra. The equilibrium allocations are given in the following lemma.

Lemma 45 The arbitrageur’s period-2 consumption in the event of a low endowment realization ($\delta > 0$) is increasing in period-1 government long-term asset purchases: $\frac{ds_{2a}}{dx} > 0$. The equilibrium allocations are as follows:

\[
s_{2a} = b_{2a} - \frac{2 + \beta}{2 + 2\beta} \delta < b_{2a}
\]

\[
s_{3a} = b_{3a} - \frac{\beta^2 \delta e_1 e_3 (e_1 (1 + \beta) + x)}{((\beta^2 (1 + \beta) e_1 + x(1 + \beta + \beta^2)) (2 e_2 (e_1 (1 + \beta) + x) - (1 + \beta) \delta e_1))} < b_{3a}
\]

Thus, with log utility, an increase in OMO $x$ serves to increase the arbitrageurs’ planned period 2 consumption, and decrease the arbitrageurs’ planned period 3 consumption. In the event of a negative endowment shock $\delta$, the arbitrageur will shift some consumption from period 3 to period 2. However, as shown in (3.11), the amount of consumption shifted from period 3 to period 2 equals $\frac{2 + \beta}{2 + 2\beta} \delta$, which does not depend on OMO $x$. Thus, knowing that OMO $x$ serves to increase the arbitrageurs’ planned period 2 consumption is sufficient to conclude that $x$ results in a higher period-2 consumption for the arbitrageur in the event of a shock.

What if agents could have anticipated the shock, as opposed to being unable to anticipate the shock as previously assumed? That is, agents know that with some probability $p$, the arbitrageurs’ period 2 endowment will be $e_2 - \delta$ and with probability $1 - p$, it will be $e_2$. In this case, an increase in OMO $x$ will unambiguously lead to an increase in arbitrageur consumption in the state in which the consumption realization is low. The intuition is the same as the intuition for Lemma 10. Under constant elasticity of intertemporal substitution, maturity-lengthening open-market operations $x$ have an effect on period-1 and 2 arbitrageur...
allocations that is equivalent to a lump-sum transfer $x$ without changing period-2 prices. Thus, period-2 arbitrageur consumption, in the shock state and the no-shock state, are increasing in $x$.

### 3.5 Conclusion

Understanding the effects of maturity-lengthening open-market operations, such as those undertaken by the Federal Reserve in response to the 2007-2009 financial crisis, is an important challenge for macroeconomics. It has been traditionally held that there are two views about a government policy of purchasing long-term debt financed by short-term borrowing. One view, based on the intuition of Ricardian equivalence and developed in the context of monetary policy by Wallace (1982), holds that open-market operations are irrelevant for the yield curve because agents can undo them through trading. Another view, popular with policymakers and modeled explicitly in Vayanos and Vila (2009) and Greenwood and Vayanos (2010), emphasizes preferred-habitat investors as a reason why maturity-lengthening open-market operations can reduce long-term interest rates. According to Greenwood and Vayanos (2010), these two views are “at odds.”

In this paper, I clarify this statement. In particular, in a Vayanos and Vila (2009) preferred-habitat model, I show that maturity-lengthening open-market operations have no effect on long-term interest rates if the agents in the economy ultimately receive the profits from the government’s portfolio via lump-sum taxes or transfers. The reason is that both arbitrageurs investors with preferred-habitat preferences will take into account their exposure to duration risk as claimants on the government’s risky portfolio. Thus, even with preferred-habitat investors, who under no government intervention prefer to hold only long-term debt, Ricardian equivalence applies.

I show that one way to break this Ricardian equivalence is to assume limited participation. In particular, if preferred-habitat investors cannot invest in short-term debt, then maturity-lengthening open-market operations will decrease the long-term interest rate. I further show that it is limited participation, and not preferred-habitat preferences, that drive this result. I do this by developing a model with limited participation, but without preferred-habitat preferences, and in which open-market operations decrease long-term rates. This model is based on a more-plausible form of limited participation, in which some...
households only invest in short-term instruments and do not trade every period.

Finally, in the models in which open-market operations are relevant, I characterize how open-market operations affect the vulnerability of arbitrageurs to negative technology or endowment shocks.

In the future, it would useful to understand the role of distortive taxation in these models and to develop a dynamic model to better understand how maturity-lengthening open-market operations affects wealth accumulation by arbitrageurs and how wealth accumulation by arbitrageurs in turn determines the effects of open-market operations. For example, an interesting question is: If agents know that the Federal Reserve will engage in maturity-lengthening open-market operations in the event of crisis, how does this affect portfolios and asset prices ahead of the crisis, and what does this imply for the impact of maturity-lengthening open-market operations in the event of a crisis?

In addition, a quantitative version of Vayanos and Vila (2009) can be developed; Hamilton and Wu (2009) present estimates of a discrete-time version of a model in the style of Vayanos and Vila (2009). It would be useful to understand the quantitative impact of maturity-lengthening open-market operations in a model that reasonably takes into account the assumptions necessary to break Ricardian equivalence. For example, if Ricardian equivalence is broken by appealing to the limited-participation of certain households – how much effect will open-market operations have if the income and assets of limited-participation households are small?
Bibliography

[1] Angeletos, George-Marios, 2007. “Uninsured idiosyncratic investment risk and aggregate saving,” *Review of Economic Dynamics*, Volume 10, Issue 1, January, pp. 1-30.

[2] Bernanke, Ben S., 2010. “The Economic Outlook and Monetary Policy,” Federal Reserve Bank of Kansas City Economic Symposium, Jackson Hole, Wyoming, August 27.

[3] Chamley, Christophe and Herakles Polemarchakis, 1984. “Assets, General Equilibrium and the Neutrality of Money,” *The Review of Economic Studies*, Vol. 51, No. 1 (Jan., 1984), pp. 129-138.

[4] Curdia, Vasco and Michael Woodford, 2011. “The Central Bank Balance Sheet as an Instrument of Monetary Policy,” *Journal of Monetary Economics*, Vol. 58, No. 1, p. 47-74.

[5] D’Amico, Stefania and Thomas B. King, 2010. “Flow and Stock Effects of Large-Scale Treasury Purchases,” Federal Reserve Board, Finance and Economics Discussion Series, No. 52, September.

[6] Doh, Taeyoung, 2010. “The Efficacy of Large-Scale Asset Purchases at the Zero Lower Bound,” *Economic Review*, Federal Reserve Bank of Kansas City, pp. 5-33.

[7] Greenwood, Robin, Sam Hanson and Jeremy Stein, 2010. “A Gap-Filling Theory of Corporate Debt Maturity Choice,” *The Journal of Finance*, Vol. 65, pp 993–1028.

[8] Greenwood, Robin and Dimitri Vayanos, 2010. “Bond Supply of Excess Bond Returns,” mimeo.
[9] Hamilton, James D. and Jing Cynthia Wu, 2010. “The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment,” mimeo.

[10] Krishnamurthy, Arvind and Annette Vissing-Jorgensen, 2011. “The Effects of Quantitative Easing on Interest Rates,” mimeo.

[11] Swanson, Eric T, 2011. “Let’s Twist Again: A High-Frequency Event-Study Analysis of Operation Twist and Its Implications for QE2,” Spring 2011 conference of the Brookings Papers on Economic Activity.

[12] Vayanos, Dimitri and Jean-Luc Vila, 2009. “A Preferred-Habitat Model of the Term Structure of Interest Rates,” mimeo.

[13] Wallace, Neil, 1981. “A Modigliani-Miller Theorem for Open-Market Operations,” The American Economic Review, Vol. 71, No. 3, June 1981, pp. 267-274
Appendix A: Proofs

Proof of Proposition 4

In period 1, the arbitrageur purchases \( h \) units of long-term debt and finances this purchase by selling \( hP \) in short-term debt. In period 2, the arbitrageur’s needs to pay \( R_1hP \) to redeem his short-term debt. In addition, the arbitrageur pays a lump-sum tax equal to \( \alpha_2\phi R_1z \). The arbitrageur rolls over \( R_1[hP + \alpha_2\phi z] \) via short-term debt. In period 3, the arbitrageur’s terminal wealth is

\[ w = h - R_1 R_2 [hP + \alpha_2 \phi z] + \alpha_3 \left[ \frac{z}{P} - (1 - \phi) R_1 R_2 z \right]. \]

We can re-write terminal wealth as

\[ w = h + \alpha_3 \frac{z}{P} - R_1 R_2 [hP + \alpha_2 \phi z + \alpha_3 (1 - \phi) z] \]

The expectation of terminal wealth is

\[ E[w] = h + \alpha_3 \frac{z}{P} - R_1 E[R_2] [hP + \alpha_2 \phi z + \alpha_3 (1 - \phi) z] \]

and the variance of terminal wealth is

\[ Var[w] = R_1^2 Var[R_2] [hP + \alpha_2 \phi z + \alpha_3 (1 - \phi) z]^2 \]

The arbitrageur solves:

\[ \max_h E[w] - \frac{\gamma}{2} Var[w] \]

The first-order condition is:

\[ 1 - R_1 E[R_2] P - \gamma R_1^2 Var[R_2] [hP + \alpha_2 \phi z + \alpha_3 (1 - \phi) z] P = 0 \]

Hence

\[ h(P; z, \alpha_2, \alpha_3, \phi) = h^0(P) - \frac{z}{P} [\alpha_2 \phi + \alpha_3 (1 - \phi)] \]

Thus, at a given price \( P \), the arbitrageurs demand exactly \([\alpha_2 \phi + \alpha_3 (1 - \phi)] \frac{z}{P}\) less long-term debt than before.

The preferred-habitat investors purchase \( k \) units of the long-term debt and invest the remainder of their wealth \((L - kP)\) in the short-term bond. In period 2, their short-term
position returns $R_1(L - kP)$ and they owe $(1 - \alpha_2)\phi R_1 z$ in lump-sum taxes. Their net period-2 wealth is rolled over, so that their terminal wealth is $k + R_1 R_2[L - kP - (1 - \alpha_2)\phi z] + (1 - \alpha_3)[\frac{z}{P} - (1 - \phi)R_1 R_2 z]$. We can write terminal wealth as:

$$w = k + (1 - \alpha_3)\frac{z}{P} + R_1 R_2[L - P(k + \frac{z}{P}[(1 - \alpha_2)\phi + (1 - \alpha_3)(1 - \phi)])]$$

Given their max-min preferences, the preferred-habitat investors will completely hedge their risk, by setting

$$k(P; z, \alpha_2, \alpha_3, \phi) = \frac{L}{P} - \frac{z}{P}[(1 - \alpha_2)\phi + (1 - \alpha_3)(1 - \phi)]$$

Thus, at a given price, the preferred-habitat investors demand exactly $[(1 - \alpha_2)\phi + (1 - \alpha_3)(1 - \phi)]\frac{z}{P}$ less long-term debt than before.

Market-clearing for the long-term bond requires:

$$h(P^*; z, \alpha_2, \alpha_3, \phi) + k(P^*; z, \alpha_2, \alpha_3, \phi) = -\frac{z}{P}$$

or

$$h^0(P) + \frac{L}{P} = 0$$

which does not depend on $z$. The equilibrium price is thus independent of $z$.

**Proof of Proposition 5**

With limited participation, the preferred-habitat investors demand for the long-term bond is $k(P; z, \alpha_2, \alpha_3, \phi) = \frac{L}{P}$, regardless of $(z, \alpha_2, \alpha_3, \phi)$. Market-clearing for the long-term bond requires:

$$h(P^*; z, \alpha_2, \alpha_3, \phi) + k(P^*; z, \alpha_2, \alpha_3, \phi) = -\frac{z}{P}$$

or

$$\frac{1}{P} - \frac{R_1 E[R_2]}{\gamma R_1^2 Var[R_2]} = -L - z[1 - (\alpha_2 \phi + \alpha_3(1 - \phi))]$$

Thus, $P$ is strictly increasing in $z$ if and only if $\alpha_2 \phi + \alpha_3(1 - \phi) < 1$.

**Proof of Lemma 7**

154
Suppose that \( p_2(x) \leq p_2(0) \) for some \( x > 0 \). Then, when \( x > 0 \), capital is weakly lower and the arbitrageurs’ bond holdings are strictly lower, so period-2 consumption is strictly lower, relative to the levels when \( x = 0 \). Moreover, because arbitrageurs’ bond holdings are lower, the price of bonds is weakly lower, and capital is weakly lower, it must be that period-1 consumption is strictly higher. Hence \( p_2(x) > p_2(0) \), a contradiction.

**Proof of Lemma 8**

The household’s problem is now:

\[
\max_{\{b_i\}} u(b_1) + \beta u(b_2 - \frac{1}{p_2} x) + \beta^2 u(b_3 + \frac{1}{p_3} x)
\]

subject to

\[
b_1 + p_2b_2 + p_3b_3 \leq e_1 + p_2e_2 + p_3e_3
\]

The first-order conditions of the households’ problem are

\[
\begin{align*}
u'(b_1) &= \lambda \\
\beta u'(b_2 - \tau_2 x) &= \lambda p_2 \\
\beta^2 u'(b_3 + \tau_3 x) &= \lambda p_3
\end{align*}
\]

Market clearing requires:

\[
\begin{align*}
b_1 &= e_1 \\
b_2 &= e_2 + \frac{1}{p_2} x \\
b_3 + \frac{1}{p_3} x &= e_3
\end{align*}
\]

Combining the first-order and market-clearing conditions:

\[
\begin{align*}
u'(e_1) &= \lambda \\
\beta u'(e_2) &= \lambda p_2 \\
\beta^2 u'(e_3) &= \lambda p_3
\end{align*}
\]
Prices, allocations, and the Lagrange multiplier do not depend on x.

**Proof of Lemma 9**

A comparison of the re-written problems of the household ((3.3) and (3.4)) and the arbitrageur ((3.5) and (3.6)) shows that if \( x > 0 \), the following hold:

\[
b_{1a} > e_1 > b_{1l}\]

and

\[
b_{2l} > e_2 > c_{2l}\]

If \( x = 0 \), these inequalities hold as equalities. This is because, for any \( p_2 \), the problems of each Type L are identical except for a subsidy to the arbitrageur and a tax on the household, both equal to x.

Moreover, market-clearing requires

\[
b_{3a} = e_2 - \frac{1}{p_3}x.
\]

From the arbitrageur’s first-order condition

\[
p_3 = \beta^2 \frac{u'(b_{3a})}{u'(b_{1a})}.
\]

According to the previous analysis, if \( x > 0 \), then the arbitrageur’s period-3 [period-1] consumption is lower [higher] than it would be if there were no government intervention. The assumption of decreasing marginal utility then gives the lemma.

**Proof of Lemma 10**

First, I will show that \( e_1 = e_2 \) implies that \( p_2 \) does not depend on x. Since the agents can freely trade period-1 and period-2 securities, the ratio of the marginal utility of arbitrageurs to the marginal utility of households must be the same in periods 1 and 2. This can be observed from the first-order conditions. With \( e_1 = e_2 \), the only way for this to hold and for markets to clear is to have constant consumption: \( b_{1a} = b_{2a} \) and \( b_{1l} = c_{2l} \). This implies that \( p_2 = \beta \) and hence does not depend on x.
Next, I will show that $p_2$ does not depend on $x$ if agents have constant elasticity of intertemporal substitution. With constant elasticity of intertemporal substitution, the first-order conditions imply

$$\frac{b_{2a}}{b_{1a}} = \frac{c_{2I}}{b_{1I}}.$$ 

Thus, for any $x$, the consumption of a given Type L's a share of the total endowment is the same in periods 1 and 2. Hence the ratio of period-2 consumption to period-1 consumption for either type is given by the ratio of period-2 endowments to period-1 endowments. With constant elasticity of intertemporal substitution, $p_2$ depends only on the former ratio, which, by its equality to the later ratio, does not depend on $x$.

Therefore, if $e_1 = e_2$ or agents have constant elasticity of intertemporal substitution, then $p_2$ does not depend on $x$. Hence, based on the re-written form of the agents’ problems, $b_{1a}$ and $b_{2a}$ are increasing in $x$.

Note that market clearing requires $b_{3a} = e_3 - \frac{1}{p_3} x$. Moreover,

$$p_3 = \beta \frac{u'(b_{3a})}{u'(b_{2a})} = \beta \frac{u'(e_3 - \frac{1}{p_3} x)}{u'(b_{2a})}.$$ 

(3.12)

Suppose, by contradiction, that there exists $x'$ and $x'' > x'$ such that $p_3(x'') \leq p_3(x')$. Since $b_{2a}$ is strictly increasing in $x$, $u'(b_{2a}(x'')) < u'(b_{2a}(x'))$. Moreover, given the assumption that $p_3(x'') \leq p_3(x')$, it must be that

$$b_{3a}(x'') = e_3 - \frac{1}{p_3(x'')} x'' < b_{3a}(x') = e_3 - \frac{1}{p_3(x')} x',$$

and hence $u'(b_{3a}(x'')) > u'(b_{3a}(x'))$. Hence $p_3(x'') > p_3(x')$, a contradiction.

**Lemma 11: Additional results**
Then period 1 and 2 equilibrium prices and allocations are given by

\[ p_2 = \beta \left( \frac{e_1}{e_2} \right)^{\frac{1}{\theta}} \]

\[ b_{1a} = \frac{e_1 + p_2 e_2 + x}{1 + p_2 e_2} \]

\[ b_{1l} = \frac{e_1 + p_2 e_2 - x}{1 + p_2 e_2} \]

\[ b_{2a} = \frac{e_2 e_1 + p_2 e_2 + x}{e_1 (1 + p_2 e_2)} \]

\[ b_{2l} = \frac{e_2 e_1 - p_2 e_2}{e_1 (1 + p_2 e_2)} \]

In the special case of log utility, (3.7) and (3.8) can be solved to express period-3 prices and allocations in closed form. In particular, we have

\[ p_3 = \beta^2 \frac{e_1}{e_3} + \frac{1 + \beta + \beta^2}{1 + \beta} \frac{x}{e_3} \]

\[ b_{3a} = \frac{\beta^2 (e_1 (1 + \beta) + x)}{\beta^2 (1 + \beta) e_1 + x (1 + \beta + \beta^2)} \]

**Proof of Lemma 11**

From the agents’ first-order conditions, we have

\[ p_2 = \beta \frac{u'(b_{2a})}{u'(b_{1a})} = \beta \left( \frac{b_{1a}}{b_{2a}} \right)^{\frac{1}{\theta}} \]

\[ p_2 = \beta \left( \frac{b_{1l}}{b_{2l}} \right)^{\frac{1}{\theta}} \]

We can solve these equations for period-1 allocations as a function of period-2 prices and allocations.

\[ b_{1a} = \left( \frac{p_2}{\beta} \right)^{\theta} b_{2a} \]

\[ b_{1l} = \left( \frac{p_2}{\beta} \right)^{\theta} c_{2l} \]

Substituting into the market-clearing constraint for period 1, we have

\[ (b_{2a} + c_{2l}) \left( \frac{p_2}{\beta} \right)^{\theta} = 2e_1 \]
Substituting from the market-clearing constraint for period 2, we have

\[
\frac{p_2}{\beta}^\theta = \frac{2e_1}{2e_2} \quad \Rightarrow \quad p_2 = \beta \left( \frac{e_1}{e_2} \right)^{\frac{1}{\theta}}
\]

(3.13)

Now we can solve for \(b_{1a}\) and \(b_{2a}\) as a function of \(p_2\) using the arbitrageurs’ first-order conditions and budget constraint. The first-order conditions are:

\[
u'(b_{1a}) = \lambda_b = b_{1a}
\]

\[
\beta u'(b_{2a}) = \lambda_b p_2 = \beta b_{2a}
\]

Thus, we can write \(b_{1b}\) and \(b_{2b}\) as follows:

\[
\lambda_b^{-\theta} = b_{1a}
\]

\[
\lambda_b^{-\theta} (\frac{p_2}{\beta})^{-\theta} = b_{2a}
\]

Substituting into the arbitrageurs’ budget constraint, one obtains:

\[
\lambda_b = \left( \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}} \right)^{-\frac{1}{\theta}}
\]

and

\[
b_{1a} = \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}}
\]

\[
b_{2a} = (\frac{p_2}{\beta})^{-\theta} \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}}
\]

Substituting for \(p_2\) from (3.13), one obtains the expressions for \(b_{1a}\) and \(b_{2a}\) given in the Lemma.

The period-3 price is characterized by substituting the resulting expression for \(b_{2a}\) and the market-clearing condition \(b_{3a} = e_3 - \frac{1}{p_3} x\) into (3.12).

**Proof of Lemma 12**

To see that \(\frac{db_{3a}}{dx} < 0\) if \(\frac{1}{\theta} \leq 1\) or if \(x\) is sufficiently small, totally differentiate (3.8) with
respect to x. This yields:

\[ \frac{db_{3a}}{dx} = \frac{1}{p_3} \left[ \frac{x \, dp_3}{dx} - 1 \right] \]

which implies that \( b_{3a} \) is strictly increasing in \( x \) if and only if the elasticity of the third-period price with respect to \( x \) is greater than one.

To investigate this elasticity, totally differentiate (3.7) with respect to \( x \). This yields

\[ \frac{dp_3}{dx} = \beta^2 \left[ -\frac{1}{\theta} (e_3 - \frac{1}{p_3} x) - \frac{1}{\theta} \left( \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}} \right) \right] \frac{x}{p_3} \frac{dp_3}{dx} - \frac{1}{p_3} \]

or, with some algebra,

\[ \frac{dp_3}{dx} = \frac{1}{p_3^2} \left( e_3 - \frac{1}{p_3} x \right) - \frac{1}{\theta} \left( \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}} \right) \frac{x}{p_3} \frac{dp_3}{dx} - \frac{1}{p_3} \]

Using this last expression, note that, with some algebra,

\[ \text{sign} \left[ \frac{dp_3}{dx} \right] = \frac{1}{p_3^2} \left( e_3 - \frac{1}{p_3} x \right) - \frac{1}{\theta} \left( \frac{e_1 + p_2 e_2 + x}{1 + p_2 (\frac{p_2}{\beta})^{-\theta}} \right) \frac{x}{p_3} \frac{dp_3}{dx} - \frac{1}{p_3} \]

\[ = \text{sign} \left[ - \left( e_1 + p_2 e_2 \right) + \left( \frac{1}{\theta} - 1 \right) x \right] \]

**Proof of Lemma 13**

Substituting (3.9) into (3.10) and taking the derivative of \( s_{3a} \) with respect to \( b_{3a} \), one obtains:

\[ \frac{\partial s_{3a}}{\partial b_{3a}} = \frac{\beta (1 - \frac{1}{\theta}) \left( \frac{b_{2a} - \frac{\delta}{b_{3a}}} {b_{3a}} \right)^{\frac{1}{\theta} + 1}} {b_{3a} \left( \left( \frac{b_{2a} - \frac{\delta}{b_{3a}}} {b_{3a}} \right)^{\frac{1}{\theta} + 1} + \left( \frac{b_{2a} - \frac{\delta}{b_{3a}}} {b_{3a}} \right)^{2} \right)} \]

Since \( \delta < b_{2a} \), we have that \( \text{sign} \left[ \frac{\partial s_{3a}}{\partial b_{3a}} \right] = \text{sign} \left[ (1 - \frac{1}{\theta}) \right] \).
Taking the derivative of $s_{3a}$ with respect to $b_{2a}$, one obtains:

$$\frac{\partial s_{3a}}{\partial b_{2a}} = (1 - \frac{1}{\theta})\beta(\frac{b_{2a} + b_{3a}B(b_{2a} - \frac{\delta}{2})^{1-\theta}/\theta - \delta}{b_{2a} - \frac{\delta}{2}})\left(\frac{b_{2a} - \delta}{b_{3a}}\right)^{\frac{1}{\theta} - 1} + (1 + \beta(\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1})(1 + \beta(\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1})$$

The denominator is clearly positive; thus, with some algebra,

$$\text{sign}\left[\frac{\partial s_{3a}}{\partial b_{2a}}\right] = \text{sign}(1 - \frac{1}{\theta})\beta(\frac{b_{2a} + b_{3a}B(b_{2a} - \frac{\delta}{2})^{1-\theta}/\theta - \delta}{b_{2a} - \frac{\delta}{2}})\left(\frac{b_{2a} - \delta}{b_{3a}}\right)^{\frac{1}{\theta} - 1} +$$

$$\left(1 + \beta(\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1}(1 + \beta(\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1})\right)$$

$$= \text{sign}[\beta(1 + \frac{b_{2a} - \delta}{b_{2a} - \frac{\delta}{2}}) + \frac{1}{\theta}\beta(1 - \frac{b_{2a} - \delta}{b_{2a} - \frac{\delta}{2}}) + \beta^2(\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1} + (\frac{b_{2a} - \delta}{b_{3a}})^{\frac{1}{\theta} - 1}]$$

Because $b_{2a} \geq e_2 > \delta$, each term in parentheses in the last equation is positive. Hence

$$\frac{\partial s_{3a}}{\partial b_{2a}} > 0.$$
Appendix B

This Appendix describes how the maturity structure of the Federal Reserve’s balance sheet changed during the crisis. It completes the discussion of the Fed’s maturity-lengthening open-market operations in Section 2.

July 2007-June 2008: Sale of (short-term) Treasury securities

Prior to the financial crisis, the Fed’s Treasury securities were mostly of short duration. In July 2007, as markets showed the first signs of dislocation, the Fed held $399 billion in Treasuries due in less than one year; this constituted more than half of the Fed’s total holdings of Treasury securities, $791 billion (see Figure 1). In the ensuing months, the Fed’s holdings of Treasuries declined, at first slowly and then more quickly, until reaching $478 billion in June 2008. During this period, the Fed sold Treasury securities in order to attain the target Federal Funds rate, which remained at or above 2 percent (see Figure 2). During this period, the introduction of liquidity programs like the Term Auction Facility (TAF) resulted in an increase in the supply of reserves, and the sale of Treasuries aimed to offset this. Sales of Treasury securities with maturities less than one year accounted for most of the sales; the Fed’s holdings of Treasury securities with maturities greater than 5 years slightly increased during this time.

July 2008-March 2009: Beginning of Agency and (long-term) MBS purchases

Between July 2008 and March 2009, the Fed’s holdings of Treasury securities remained between $470 and $480 billion, with a slight shift toward longer maturities. During this time, the Fed began its purchases of mortgage-backed and agency securities. Prior to the crisis, the Federal Reserve did not hold any MBS. The Fed’s purchases of MBS began in January 2009 and increased (at a generally decreasing rate) until reaching $1.13 trillion in June 2010. Throughout this period, almost all (at least 99.9 percent) of the Fed’s MBS holdings had maturities of 10 years or more. The Fed’s purchases of agency securities began in September 2008 and reached $169 billion in March 2010.

April 2009-: Purchases of (long-term) Treasury securities

In April 2009, the first of two rounds of purchases of Treasury securities began. These rounds are popularly known as “QE1” and “QE2.” The first round began in April 2009 and ended in October 2009; during this time, Treasury holdings increased by almost $300
billion. This included a $42 billion increase in holdings with maturity greater than ten years; a $109 billion increase in holdings with maturity between five and ten years; and a $156 billion increase in holdings with maturity between one and five years. Holdings of Treasury securities with maturity less than one year slightly declined. The second round began in August 2010 and remains ongoing, as of April 2011. Between August 2010 and March 2011, Treasury holdings increased $489 billion. This included a $32 billion increase in holdings with maturity greater than ten years; a $236 billion increase in holdings with maturity between five and ten years; and a $207 billion increase in holdings with maturity between one and five years. The Federal Reserve has announced plans to purchase a total of $600 billion in Treasury securities in this ongoing second round.
Figure 1. Federal Reserve holdings of Treasury securities by maturity

Data: Federal Reserve. Millions of dollars.

Figure 2. Federal Funds target rate

| Date             | Rate  |
|------------------|-------|
| December 11, 2007| 4.25  |
| January 22, 2008 | 3.5   |
| January 30, 2008 | 3     |
| March 18, 2008   | 2.25  |
| April 30, 2008   | 2     |
| October 8, 2008  | 1.5   |
| October 29, 2008 | 1     |
| December 16, 2008| 0-0.25|

Figure 3. Federal Reserve holdings of MBS with maturity greater than 10 years

Data: Federal Reserve. Millions of dollars.