Refinancing Frictions, Mortgage Pricing and Redistribution

David Berger† Konstantin Milbradt‡ Fabrice Tourre§
Joseph Vavra¶

January 2023

Abstract

There are large cross-sectional differences in how often US borrowers refinance mortgages. In this paper, we develop an equilibrium mortgage pricing model that allows us to explore the consequences of this heterogeneity. We show that equilibrium forces imply important cross-subsidies from borrowers who rarely refinance to those who refinance often. Mortgage reforms can potentially reduce these regressive cross-subsidies, but the equilibrium effects of these reforms can also have important distributional consequences. For example, many policies that lead to more frequent refinancing also increase equilibrium mortgage rates and thus reduce residential mortgage credit access for a large number of borrowers.

*We would like to thank Morris Davis and David Zhang (both discussants) for helpful discussions and the seminar participants at USC, UIUC, Northwestern University, UCLA, Copenhagen Business School, UIC, Cleveland Fed, Society of Economic Dynamics (2022), Chicago Fed Housing and Racial Bias Workshop (2022), Rio FGV (2022), TAU (2022), NYU Stern, and BU Finance for helpful discussions and feedback. Fabrice Tourre gratefully acknowledges financial support from the Danish Finance Institute and the Center for Financial Frictions (FRIC) (grant no. DNRF-102).

†Duke University and NBER; david.berger@duke.edu
‡Northwestern University and NBER; milbradt@northwestern.edu
§Copenhagen Business School; ft.fi@cbs.dk
¶University of Chicago and NBER; joseph.vavra@chicagobooth.edu
1 Introduction

There are large cross-sectional differences in how often US borrowers refinance their fixed-rate mortgages. Some “fast” borrowers refinance frequently. Other “slow” borrowers do not refinance despite substantial financial incentives.\footnote{See Keys, Pope, and Pope (2016) and Andersen et al. (2020) for evidence of low refinancing propensities on average and Gerardi, Willen, and Zhang (2020) and Zhang (2022) for evidence of cross-borrower heterogeneity.} In this paper, we argue that this heterogeneity has important equilibrium implications. We develop and characterize an equilibrium model of the mortgage market with persistent borrower heterogeneity, estimate it using US mortgage micro data, and show that heterogeneous refinancing leads to equilibrium forces that amplify inequality.

Institutional features of US mortgage markets limit lenders’ ability to price-discriminate, so borrowers with very different refinancing propensities face the same mortgage rates at origination. We show that this pooling equilibrium leads to substantial cross-subsidies from slow to fast borrowers: slow borrowers pay higher rates and fast borrowers lower rates at origination than if lenders could price-discriminate.

These equilibrium forces are also important for evaluating alternative mortgage market designs and policies. Since heterogeneous refinancing leads to substantial inequality, it is natural to think that policies leading to more frequent refinancing would improve borrower welfare and reduce inequality. However, we show that the same equilibrium forces that play an important role in the current market also matter for evaluating the distributional consequences of various policy counterfactuals. For example, “automatically refinancing” mortgages eliminate refinancing disparities across borrowers but would also lead lenders to charge higher rates on newly originated mortgages and thus reduce mortgage credit access for a large number of borrowers. It is important to account for these equilibrium effects in addition to the more commonly studied direct effects of policy reforms.

While our insight—the fact that the consequences of housing finance policies depend on equilibrium effects—is not new,\footnote{See, e.g., Campbell (2006).} systematic analysis of effects on inequality in equilibrium has been limited by the complexity of equilibrium environments with heterogeneity. Beyond our specific mortgage application, an important contribution of our paper is thus the development of a tractable framework that can be used to study equilibrium environments featuring ex ante household heterogeneity.
We develop this framework in three key steps. In the first step, we characterize in partial equilibrium the optimal refinancing decisions of borrowers facing the two main frictions identified by the past literature (e.g., Andersen et al. (2020)). Specifically, we allow both for “inattention” or other nonmonetary frictions (which generate time-dependent inaction) and for fixed monetary costs of refinancing (which generate state-dependent inaction) and solve for the optimal behavior of borrowers.

In the second step, we embed this household refinancing problem into an equilibrium model of the mortgage market under the assumption that borrowers are ex-ante identical. We assume that risk-neutral competitive investors purchase mortgage-backed securities (MBS), which pool together monthly payments made by borrowers (net of any intermediation fees), and we characterize the relative role of different frictions for new-issue MBS prices and resulting mortgage rates in equilibrium. We show that monetary fixed costs have small effects on equilibrium mortgage pricing since these costs primarily reduce refinancing for borrowers with small “rate gaps” (the difference between the coupon and current mortgage market rate), and closing small gaps via refinancing barely affects lender profits. In contrast, inattention has large effects on mortgage rates since it reduces refinancing even for borrowers with large gaps and this substantially changes lender profits and pricing.

The third step introduces heterogeneous refinancing frictions across borrowers into this equilibrium environment, which allows us to explore the redistributive effects of various mortgage market interventions. Consistent with institutional features of the US agency MBS market, we focus primarily on a “pooling” equilibrium in which lenders do not price-discriminate based on borrowers’ refinancing speed. We make two simplifying assumptions to increase tractability and approximate the cross-sectional distribution over households’ coupons and attention rates, which would otherwise arise as an additional state variable in the investors’ pricing problem, by a low-dimensional object. First, we assume that fixed costs are not paid up front but are instead capitalized into a higher interest rate for the new loan. This is a strong assumption, but it broadly aligns with actual US mortgage markets and greatly simplifies borrower decisions. Second, we assume that investors exhibit a simple form

\footnote{More than 80% of mortgage origination costs in the US are rolled into higher rates rather than paid up front (Zhang, 2022). In addition, our finding that up-front fixed costs have modest effects on equilibrium pricing suggests that modeling the remaining 20% of origination costs financed by borrowers via upfront closing costs would substantially complicate the analysis but have little quantitative impact.}
of bounded rationality: they value mortgages based on the average distribution of attention for those refinancing in the current rate environment but do not account for the entire history of rates when inferring this distribution. These two assumptions simplify the pooling equilibria substantially, allowing us to then characterize sufficient conditions for existence and many other important properties. For example, in equilibrium, borrower heterogeneity affects mortgage pricing through a simple covariance adjustment term.

We next turn to the model’s quantitative results. We estimate the cross-sectional distribution of borrower attention using a monthly borrower-level panel of mortgages from 2005 to 2017 and explore the implications of this heterogeneity in our equilibrium model. We start by testing whether the equilibrium mortgage rates implied by our model match the time series of actual mortgage rates in the data. We take US treasury yields from 2005 to 2017 together with estimates of intermediation costs from the literature and calculate the equilibrium mortgage rates and refinancing patterns implied by the model. These align well with the data, giving us confidence in the model’s implications.

We then explore the implications for inequality. We find that the current pooling equilibrium in the US results in substantial cross-subsidies from slow to fast borrowers. The fastest (slowest) borrowers incur mortgage interest expenses on average 80 bps lower (10 bps higher) than the average household. This inequality stems purely from ex post differences in refinancing propensities and is consistent with that documented in Gerardi, Willen, and Zhang (2020). This analysis, however, misses an important source of redistribution that arises from equilibrium effects: in a counterfactual “separating” equilibrium in which lenders could discriminate based on type, the fastest borrowers would face mortgage rates at origination 440 bps higher (on average) than the slowest borrowers and would incur an average mortgage interest expense 180 bps higher than in the pooling equilibrium. By the same token, the slowest households’ average mortgage interest spending would decrease by 25 bps from the corresponding average in the pooling environment. Thus, moving from a pooling to a separating equilibrium could substantially reduce inequality arising from the liability side of households’ balance sheets.

As usual, the quantitative importance of this assumption cannot be fully evaluated without solving the true (intractable) pricing problem. However, we provide some some evidence that this simplifying assumption likely has little quantitative effect on our conclusions.

The relevance of this counterfactual mortgage market equilibrium depends on the extent to
We then explore the equilibrium effects of a frequently discussed alternative contract design aimed at reducing mortgage inequality: the “automatically refinancing” mortgage. This contract refines automatically with no active borrower intervention when rates decline. Although automatically refinancing mortgages lead to much more refinancing for inattentive borrowers, they also lead to an increase in mortgage rates at origination of about 130 bps on average, offsetting some of these gains. Automatically refinancing mortgages yield individual time-paths of mortgage coupons that decline more rapidly but that start from a higher initial value. Over the loan life, these mortgages reduce average coupons by 30 bps for the slowest borrowers, but this decrease is substantially smaller than the 90 bps reduction that would arise without the equilibrium rate increase at origination. This rate increase is also likely to have important implications for access to housing markets: higher interest rates may force households that are at debt-to-income (DTI) limits to downsize their purchases or may exclude these households from the housing market entirely.\footnote{Although called DTI, these constraints cap the ratio of monthly payments to monthly income.} Borrowers benefit from the more frequent refinancing induced by automatic refinancing only if they are able to afford a mortgage at the initial higher rate in the first place.\footnote{Our model does not analyze initial home purchases and instead focuses on cross-subsidies across borrowers from refinancing, but a simple back-of-the-envelope calculation suggests that the increase in interest rates arising from a move to automatically refinancing mortgages might force approximately 20% of borrowers to select smaller homes requiring a smaller initial mortgage balance.}

In addition to evaluating alternative mortgage contracts, we show that other changes in the mortgage market that affect refinancing can have important equilibrium effects. For example, our data suggest that the rise of fintech and other nonbank lenders has spurred more frequent refinancing. Loans originated by nonbank lenders have 100 bps greater effective refinancing attention than loans originated by banks.\footnote{The patterns that we identify are cross-sectional correlations and do not necessarily isolate causal effects. However, some of this effect is likely causal since these lenders profit primarily from mortgage origination and so actively encourage refinancing.} If these same cross-sectional patterns were to hold with a move from a mortgage market dominated by banks to one dominated by nonbanks, it would have important equilibrium implications: a 100 bps increase in attention would lead equilibrium interest rates to rise by 50 bps on average.

While our paper focuses on mortgage markets, our modeling framework lends itself which borrower attention is observable and thus potentially priced ex ante and not just ex post. Evidence in Gerardi, Willen, and Zhang (2020) and our own analysis in Section XXX suggest that ex ante observable heterogeneity is indeed substantial.
to the analysis of many other settings and to future quantitative research. These environments all share the following features: on one side of the market, ex ante heterogeneous economic agents make dynamic discrete choices about entering into or renewing a long-term, non–state-contingent contract subject to some frictions, and the other side of the market is competitive but cannot, for informational or legal reasons, price-discriminate. For example, consider the classic labor market environment of Harris and Holmstrom (1982), in which risk-neutral firms set wages to insure risk-averse workers who cannot commit to turning down outside offers. We can use our framework to analyze the wage implications of heterogeneity in outside offer arrival rates. In a pooling wage equilibrium, workers with infrequent outside offers receive lower wages than they would in a separating equilibrium and thus effectively subsidize the wages of less loyal workers who receive frequent outside offers.

The remainder of the paper is structured as follows: Section 2 discusses the related literature. Section 3 characterizes households’ refinancing behavior given exogenous mortgage rates. Section 4 introduces the equilibrium, first with homogeneous and then with heterogeneous households. Section 5 lays out our key policy counterfactuals, Section 6 describes the data and estimation of household heterogeneity necessary to discipline these counterfactuals, and Section 7 quantifies the pricing and distributional consequences of this heterogeneity. Finally, Section 8 discusses other applications of our framework. The online supplementary materials contain all proofs.

## 2 Related literature

A growing literature provides evidence that households fail to refinance their mortgages optimally. Keys, Pope, and Pope (2016) argue that approximately 20% of US households fail to refinance when optimal even when they are eligible, and they provide some survey evidence supporting inattention and behavioral explanations. Agarwal, Rosen, and Yao (2016) provide empirical evidence that US households fail to refinance their mortgages optimally and correlate these patterns with various observable proxies for financial sophistication (see also Amromin et al. (2018)).

---

9There are many reasons the arrival rate of outside offers differs even for workers with identical productivity. Some workers are less “loyal” and solicit outside offers more aggressively, while others have constraints related to children’s schooling or spousal employment that make potential outside employers view them as less “movable.” Pooling is likely to arise both because “loyalty” cannot be directly observed and because it is illegal to set wages based on many of these characteristics.
sen et al. (2020) use even more detailed micro data from Denmark to show that both fixed costs and inattention are important for understanding individual refinancing patterns.

In works complementary to our paper, Fisher et al. (2021) and Zhang (2022) analyze the distributional impacts of heterogeneous refinancing rates. Fisher et al. (2021) analyze the UK mortgage market setting in which mortgages come with a short teaser rate that later resets to the market rate. Using a partial equilibrium consumption model, they estimate the distributional consequences of moving from this teaser system to a fixed-rate product that would generate the same revenue for lenders. Zhang (2022) uses US data to study cross-subsides arising from interactions between heterogeneous refinancing propensities and purchase points. He analyzes how closing fees change the equilibrium between mortgage originators and heterogeneous borrowers but takes MBS prices as fixed at their empirical values for the pooling equilibrium and computes the equilibrium only for the counterfactual separating environment. Our analysis is similarly motivated by household heterogeneity; however, we develop an equilibrium mortgage pricing framework that endogenizes MBS prices and mortgage rates and show that these equilibrium forces have important redistributive consequences, irrespective of the presence of up-front closing costs.

Two related papers study models with equilibrium mortgage pricing but without permanent borrower heterogeneity. Guren, Krishnamurthy, and McQuade (2021) study mortgage market reforms in an equilibrium model with borrower refinancing and risk-neutral competitive mortgage investors. Ex post heterogeneity arises in their model from income and moving shocks, but households are ex ante identical. This means their model cannot speak to the distributional issues that are the focus of our paper. The model in Berger et al. (2021) is most similar to ours, but they focus on entirely different questions. Our relative contribution is twofold: first, we more fully analyze the equilibrium and the importance of various frictions for pricing. Second, and more importantly, we study environments with permanent borrower heterogeneity and show that this heterogeneity generates important effects on inequality in equilibrium.

A large literature studies the impact of heterogeneity in capital returns on the asset side of households’ balance sheets (see, e.g., Benhabib, Bisin, and Zhu (2011), Bach, Calvet, and Sodini (2020) and Fagereng et al. (2016)). We complement this stream by showing that refinancing frictions contribute to wealth inequality via realized return
heterogeneity on the liability side of households’ balance sheets. This heterogeneity is more modest than return heterogeneity on the asset side but is very persistent and so can have a nonnegligible effect on wealth inequality.

3 Households’ refinancing behavior

In this section, we present a continuous time model of mortgage refinancing behavior with risk-neutral households. Given our focus on the US mortgage market, we study fixed-rate mortgage contracts that can be refinanced at any time. We consider households that face two types of potential refinancing frictions, which lead to state-dependent and time-dependent inaction. We initially take mortgage rates as given before endogenizing them in Section 4.

3.1 Setup

Time t is continuous. We consider a continuum of risk-neutral, long-lived households of measure 1, discounting utility flows at rate ρ. Each household carries a long-term fixed-rate prepayable mortgage with coupon rate c_t and constant unit notional balance. We denote as m_t the prevailing mortgage market interest rate, i.e., the rate that a household refinancing at time t can lock in. Households’ ability to refinance is hindered by two different frictions. First, they are inattentive and make decisions only at discrete points in time, modeled as i.i.d. Poisson events occurring with intensity χ. Second, they bear upfront closing costs ψ when refinancing. Last, households move from one house to another at intensity ν; when doing so, they must reset their mortgage coupon to the prevailing mortgage rate.10

Given our focus on a Markovian environment, we assume that the aggregate uncertainty is summarized by a latent state vector x_t, a possibly multidimensional, time-homogeneous Itô process with drift μ(x), diffusion σ(x) and infinitesimal generator L.11 The mortgage market interest rate is then a function m_t = m(x_t) of this latent state vector. For now, we assume that m(·) is continuous in x, and in Section 4,

10 ν can be viewed as the sum of a moving intensity and an amortization intensity—under the assumption that contractual mortgage balances amortize exponentially (an approximation of the actual amortization profile of a standard 30-year mortgage contract). Moving-related fixed costs could be added to the model without changing any of our conclusions.

11 L is defined over functions f of class C^2 via Lf(x) = μ(x) ⋅ ∂_x f(x) + 1/2 trace (σ'(x)∂_xx' f(x)σ(x)).
we prove that the equilibrium of our economy must satisfy this property.

Later, we consider households that might differ in their attention rate $\chi$ and the consequences of this heterogeneity for mortgage rates. For now, given our partial equilibrium focus, this heterogeneity is irrelevant, and our notation thus abstracts from the (potentially) household-specific nature of parameter $\chi$.

### 3.2 Interpreting the refinancing frictions

Households’ inability to make decisions continuously is sometimes referred to as *time-dependent* inaction, and it has featured in a large literature on rational inattention.\textsuperscript{12} The attention parameter $\chi$ should be viewed as a stand-in for various nonmonetary frictions. Some households, for example, cannot refinance even if it is beneficial for them to do so due to insufficient home equity or nonverifiable income (see Beraja et al. (2019)). Other households have low financial literacy and might only partially understand the mechanics of refinancing a mortgage. Thus, while we refer to households’ *inattention*, this friction should be understood as encompassing a wide set of environmental and behavioral factors.

The up-front closing cost borne by households when refinancing leads to *state-dependent* inaction, as it arises when a household’s incentives cause it to *decide* to stay put, and such incentives vary as the economic environment evolves. These up-front closing costs include application fees and the “points” payable out of pocket by borrowers on the transaction closing date; they also represent a component of the revenues collected by lenders upon mortgage origination.

Having discussed the refinancing frictions faced by households and their interpretation, we now solve the household decision problem.

\textsuperscript{12}For economic applications of the rational inattention modeling framework, see Reis (2006) in the context of inattentive consumers making consumption–savings decisions or Abel, Eberly, and Panageas (2007) in the context of inattentive investors in the stock market. See also Calvo (1983) in the context of sticky price models where firms make pricing decisions at discrete points in time.
3.3 Household optimal behavior

Let $V(x, c)$ be the valuation of all future mortgage liabilities for a household paying a coupon $c$ on its mortgage, when the latent state is $x$. Such a household solves

$$V(x, c) := \inf_{a \in \mathcal{A}} \mathbb{E}_{x,c} \left[ \int_{0}^{+\infty} e^{-\rho t} \left( c_t^{(a)} dt + a_t \psi dN_t^{(\chi)} \right) \right],$$

s.t. $dc_t^{(a)} = (m(x_t) - c_t^{(a)}) \left( a_t dN_t^{(\chi)} + dN_t^{(\nu)} \right)$,

where $\mathcal{A}$ is a set of progressively measurable binary actions $a = \{a_t\}_{t \geq 0}$ such that $a_t \in \{0, 1\}$ at all times, $N_t^{(\chi)}$ (resp., $N_t^{(\nu)}$) is a counting process with jump intensity $\chi$ (resp., $\nu$), $c_t^{(a)}$ is the coupon rate on the mortgage for a household following strategy $a$, and the subscript on the expectation indicates that it is conditional on the information available at time $t$. At the random points in time when the household pays attention, the household choice $a_t = 1$ represents a decision to refinance, while $a_t = 0$ means that the household chooses to keep its existing mortgage. $V$ captures all mortgage liabilities—related to household’s current mortgage (at rate $c$) and all future mortgages arising from future refinancing decisions. Going forward, let $z_t := c_t - m_t$ be the refinancing incentive, or rate gap, of a given household at time $t$. In the online appendix, we establish the following result:

**Proposition 1.** $V$ is twice continuously differentiable in $x$ and continuous and strictly increasing in $c$. It satisfies the Hamilton–Jacobi–Bellman (HJB) equation

$$(\rho + \nu + \chi) V(x, c) = c + \mathcal{L}V(x, c) + \nu V(x, m(x)) + \chi \min \left[ V(x, c), V(x, m(x)) + \psi \right].$$

The optimal refinancing choice satisfies

$$a^*(x, c) = 1_{\{c - m(x) \geq \theta(x)\}},$$

where the rate gap threshold $\theta(x)$ satisfies the value matching condition

$$V(x, m(x)) + \psi = V(x, m(x) + \theta(x)).$$

Our proof relies on standard results for stochastic control problems in continuous time. Proposition 1 holds for any arbitrary (continuous) mortgage function $m(\cdot)$,
i.e., not only the equilibrium one. It states that a household refinances optimally whenever its rate gap is above a state-dependent cutoff $\theta(x)$ and whenever it pays attention to mortgage rates. HJB (2) usually does not admit an analytical solution, except in special cases that we discuss now.

First, consider the environment where households do not bear any up-front closing costs. In this case, households optimally refinance as soon as they pay attention and their contractual coupon is above the mortgage market rate. This environment will soon become the main focus of our paper.

**Corollary 1.** Absent upfront closing costs ($\psi = 0$), the rate gap threshold is $\theta(x) = 0$, and the optimal refinancing choice is $a^*(x, c) = 1_{\{c \geq m(x)\}}$.

Next, consider the case where the mortgage rate is a Brownian motion, i.e., $m_t = \sigma B_t + m_0$. This simplified environment allows us to derive analytic expressions for the value function and rate gap threshold and leads to several important insights.

**Proposition 2.** Assume that $m_t$ is a Brownian motion with volatility $\sigma$. Introduce the constants

$$
\eta_x := \sqrt{2(\rho + \nu + \chi)} / \sigma, \quad \epsilon_x := (\rho + \nu) (\eta_0 + \eta_x) / \chi.
$$

The household’s value function satisfies $V(m, c) = \xi + v(z)$, with $z = c - m$ and $v$ admitting the functional form showed in the online appendix. The (state-independent) rate gap threshold $\theta > 0$ satisfies the implicit equation

$$
e^{-\eta_0 \theta} + (\eta_0 + \epsilon_x) \theta = 1 + (\eta_0 + \epsilon_x) (\rho + \nu) \psi. \quad (5)
$$

A second-order Taylor expansion around $\theta = 0$ yields the following approximation $\hat{\theta}$:

$$
\hat{\theta} = \sqrt{\frac{2}{\eta_0} \left(1 + \frac{\epsilon_x}{\eta_0}\right) (\rho + \nu) \psi + \left(\frac{\epsilon_x}{\eta_0}\right)^2 - \frac{\epsilon_x}{\eta_0}}. \quad (6)
$$

Moreover, $\theta$ is an increasing function of the attention rate $\chi$, and asymptotically:

$$
\lim_{\chi \to 0} \theta = (\rho + \nu) \psi \quad (7)
$$

$$
\lim_{\chi \to +\infty} \theta = \frac{1}{\eta_0} \left[1 + \eta_0 (\rho + \nu) + W(-\exp(-1 - \eta_0 (\rho + \nu)))\right], \quad (8)
$$
where $W$ is the Lambert $W$ function.

Our proof relies on the observation that the value function can be decomposed into the sum of (a) the present value of all future interest payments $c/\rho$ (based on the current mortgage coupon) plus (b) a prepayment option $v(z)$ that, given the unit root behavior of mortgage rates, depends only on the rate gap $z$. Proposition 2 generalizes the results of Agarwal, Driscoll, and Laibson (2013) (hereafter ADL) to the case where households are inattentive. Formula (8) shows that in the limit where households are infinitely attentive, the rate gap threshold at which they refinance converges to the ADL threshold. Instead, in the limit where households do not pay attention, formula (7) implies that the rate gap threshold converges to the annuity value of the upfront closing cost $\psi$, computed at the effective discount rate $\rho + \nu$. Most importantly, a decrease in $\chi$ reduces the rate gap threshold: when the attention rate decreases, the household exercises its refinancing option “sooner” when the opportunity arises. Figure 1 illustrates the sensitivity of the cutoff $\theta$ to the household’s attention rate.

In our quantitative applications, we will show that the parameter $\chi$ in the data is approximately 19% per year, i.e., $\log(\chi) \approx -1.66$. At this level of attention, the “effective” refinancing threshold implied by our model is only 30% of that computed using the ADL formula.

This analysis sheds new light on empirical studies focusing on mistakes made by households in connection with their refinancing decisions. Agarwal, Rosen, and Yao (2016) and Fuster et al. (2019), for instance, conclude that borrowers refinance at rate gaps that are on average too small relative to the ADL threshold. Once we take into account the fact that households exhibit inattention, households’ optimal threshold is reduced significantly; rather than making refinancing mistakes (by refinancing at excessively low rate gaps, as these empirical studies suggest), households may act rationally—refinancing aggressively when they have the chance—subject to their attention friction.

Proposition 2 relies on the assumption that the mortgage rate $m_t$ is a random walk. Instead, we want to focus on the equilibrium of our economy, where the driving process $x_t$ affects the term structure of interest rates and $m_t$ is endogenously determined, with dynamic properties that differ from those of the Brownian motion—in particular.

---

13 Agarwal, Rosen, and Yao (2016) find that households refinance at rate gaps with an average of 121 bps, relative to the average ADL threshold of 158 bps. Similarly, Fuster et al. (2019) find that among refinancing households, more than half of the refinancings are executed at rates that are too small when assessed against the ADL threshold.
particular, equilibrium mortgage rates in our environment are mean-reverting. In this case, refinancing decisions are necessarily state dependent, i.e., \( \theta = \theta(x) \). To what extent does mortgage rate mean reversion alter a household’s rate gap threshold? The online appendix addresses this question by focusing on mortgage rates that follow an Ornstein–Uhlenbeck process, varying the degree of persistence. Mean reversion introduces a nonnegligible amount of state dependence in \( \theta(x) \), which then has a positive slope: the rate gap threshold is lowest at low mortgage rates and highest at high mortgage rates. Intuitively, when mortgage rates are low, a household that manages to refinance can lock in a low coupon for a long time, without the need to refinance (and thus to incur up-front closing costs) in the near future. Instead, at high mortgage interest rates, rates are drifting downwards, and households might be reluctant to incur the up-front costs today, given the projected future path of rates. Given the degree of persistence of mortgage rates observed in the data, each percentage-point increase in the current mortgage rate increases the rate gap threshold \( \theta(x) \) by 4 bps, so while this effect is nonnegligible, it is quantitatively modest.
4 Mortgage market equilibrium

We now introduce mortgage investors and discuss the resulting equilibrium environment. While households pay coupon $c_t$ on their mortgage, investors receive only $c_t - f$, with a wedge $f$ capturing the fees charged by intermediaries for providing various services.\footnote{The fee rate $f$ captures (a) the ongoing portion of G-fees paid to the GSEs and (b) the servicing fee paid to mortgage servicers. However, the servicing fee is usually sold off separately by the originator and is thus captured by the gain on sale $\pi$. To avoid double-counting, we therefore drop the part of $f$ that arises from servicing fees. For simplicity, we assume that these fees are uniform across households. See also Footnote 15.} At the time of origination, mortgage pools are sold by the initial lender to (secondary market) investors at a price of $1 + \pi$, where the “gain on sale” $\pi$ represents revenues generated by the original lender (in addition to those arising from up-front closing costs $\psi$ paid by households).\footnote{Total revenues—the up-front closing cost $\psi$ and the gain on sale $\pi$—compensate the lender for all costs incurred in connection with mortgage origination. These origination costs include (a) legal and underwriting, (b) broker commissions, (c) hedges of mortgage locks, (d) future servicing, and (e) the portion of guarantee fees related to “loan-level price adjustment” (for Fannie Mae) or “credit fees for mortgages with special attributes” (for Freddie Mac) and payable up front by the original lender. See Fuster et al. (2013) for a detailed description of mortgage lenders’ costs of origination.} The sum $\pi + \psi$ is the marginal origination cost, i.e., the price of intermediation, as defined in Fuster, Lo, and Willen (2017).

We start with a general environment that includes both state- and time-dependent inaction. However, we numerically show that, for empirically relevant values, up-front closing costs $\psi$ have only small effects on equilibrium mortgage rates. Thus, the choice to include up-front closing costs (or not) has only small effects on our conclusions. In practice, most households do not pay closing costs up front and instead roll them into higher rates. Together, these observations motivate us to abstract from state-dependent frictions, which dramatically simplifies our subsequent analysis.

We initially focus on households that are ex ante homogeneous in their attention rate $\chi$. As we discuss in more detail in Section 6, both US and Danish data reject the hypothesis of homogeneous attention. Nevertheless, this homogeneous environment serves as an important building block for the empirically relevant case in which households exhibit ex ante attention heterogeneity.
4.1 Homogeneous households

In this section, all households share the same attention parameter $\chi$. When pricing mortgage debt, investors take households’ refinancing decisions as given. Let $P(x, c; \chi)$ denote the market price of a unit mortgage with coupon $c$ whose borrower has attention intensity $\chi$, when the latent state is $x$:

$$P(x, c; \chi) := \mathbb{E}_x \left[ \int_0^\tau e^{-\int_0^t r(x_s)ds} (c - f) dt + e^{-\int_0^\tau r(x_s)ds} \right],$$

(9)

where $\tau$ is the (random) prepayment time. Competitive mortgage lenders must break even when extending a new loan and immediately selling it to secondary market investors and consequently need to generate a gain on sale $\pi$ at the time of loan origination to recoup their costs. This yields the equilibrium condition

$$P(x, m(x); \chi) = 1 + \pi.$$  

(10)

We are now equipped to define an equilibrium in this environment.

**Definition 1.** A Markov perfect equilibrium (MPE) is defined as (i) a household value function $V$ that satisfies (2), (ii) the associated optimal refinancing policy satisfying (3), (iii) a pricing function $P$ defined via (9) and (iv) a mortgage rate function $m(x)$ that satisfies (10).

Since the definition of $P$ in (9) implicitly depends on a mortgage rate function $m(x)$ (via the prepayment time $\tau$), and since the equilibrium condition (10) defines $m(x)$ implicitly via the function $P$, the MPE is a fixed-point problem. Our equilibrium concept is then analogous to the Markov perfect equilibria studied in the sovereign or dynamic corporate debt literature.\(^{16}\) In these environments, the existence and uniqueness of the equilibrium frequently depend on various assumptions. In the context of mortgage prepayments, the special case without up-front closing costs allows us to derive several sharp results. We will shortly argue that in equilibrium, the general setup with up-front closing costs does not depart materially from this special case. Consequently, we will rely on this simplified environment when studying the impact of household heterogeneity on equilibrium outcomes.

\(^{16}\)See Chatterjee and Eyigungor (2012) for an example of MPE in the context of a sovereign default model or DeMarzo and He (2021) in the context of a corporate dynamic capital structure model.
Proposition 3. Assume finite attention rate (i.e., $\chi < \infty$) and assume that short-term rates $r_t$ are positive and bounded. Absent up-front closing costs (i.e., $\psi = 0$),

i. if the gain on sale $\pi = 0$, there exists a unique MPE;

ii. if the gain on sale $\pi > 0$ and if $x$ is unidimensional with $r(\cdot)$ being monotone increasing, there exists a unique MPE in which the mortgage rate function $m(x)$ is increasing in $x$.

In our model, households optimize over their refinancing decisions, subject to various frictions, taking as given the behavior of mortgage market rates. Investors price mortgages competitively, taking as given households’ refinancing behavior. Proposition 3 tells us that this fixed-point problem, absent closing costs, always admits a unique solution. In this special setup, we can disentangle the households’ decision problem from the investors’ pricing problem: irrespective of how rates evolve, households want to refinance whenever their coupon is above the current mortgage rate.\footnote{This observation holds irrespective of the gain on sale $\pi \geq 0$. When households must pay up-front closing costs ($\psi > 0$), they solve a difficult option pricing problem, as in ADL; instead, when they roll such costs into a higher rate, the problem simplifies to comparing their current mortgage coupon to the market rate. Mortgage market rates must then adjust so as to generate a sufficient gain on sale $\pi$ to recover origination costs.}

This simplifying assumption also allows us to deliver the following comparative static result.

Proposition 4. Under the assumptions of Proposition 3 for which a unique MPE exists, the mortgage market interest rate $m(\cdot)$ is increasing in the attention rate $\chi$.

Proposition 4 implies that higher household attention is worse from the point of view of mortgage market investors. With higher attention, households tend to exercise their prepayment option more optimally, and since mortgage investors are short this option, these investors react by raising mortgage market interest rates. The left-hand side of Figure 2 illustrates the sensitivity of the equilibrium mortgage rate function to a range of attention parameters $\chi$ that are below and above our estimated average value (see Section 6.3); when $\chi$ increases from 50% to 150% of this value, the ergodic average of equilibrium mortgage rates increases by 34 bps.

When inattentive households bear up-front closing costs, we can study numerically the extent to which these costs influence equilibrium mortgage rates. The right-hand side of Figure 2 illustrates this sensitivity for a range of realistic fixed costs; going
Figure 2: **Equilibrium mortgage rates vs. $\chi$ and $\psi$.** The left figure shows the sensitivity of the mortgage rate to the attention rate $\chi$ when $\psi = 0$. The right figure shows the sensitivity of the mortgage rate to the fixed-cost parameter $\psi$ when the attention rate $\chi$ equals its estimated average value 19% (see Section 6.3). All other parameters are given in Table 1. The ergodic distribution of $r$ is shown in the shaded pink area.

Figure 2 suggests that equilibrium mortgage rates have low sensitivity to up-front closing costs. Intuitively, while both fixed costs and inattention play a role, omitting fixed costs overestimates the occurrence of small-rate-gap refinancings, while omitting inattention underestimates the occurrence of large-rate-gap refinancings. In equilibrium, large-rate-gap refinancings have a greater impact on investors’ profits—and thus on mortgage rates—than refinancings with small rate gaps. We have verified the robustness of this observation for various short rate processes and attention rates.\(^1\)

Thus, up-front closing costs have little quantitative effect on equilibrium mortgage rates from $\psi = 0$ to $\psi = 2\%$ (of the mortgage balance) causes equilibrium mortgage rates to decrease on average by 13 bps.

\(^1\)In more detail, with mortgage rates held fixed, an increase in the up-front closing costs $\psi$ raises the rate gap threshold $\theta(x)$, dampening refinancing activity, particularly when interest rates are high, given the state-dependent nature of $\theta(x)$ documented in the online appendix. In equilibrium, this reduced refinancing activity lowers and flattens the mortgage rate function $m(x)$. A mortgage rate function with a smaller slope leads to lower mortgage rate volatility, particularly at high interest rate states; this, in turn, leads households to exercise their refinancing option “earlier,” i.e., at a lower rate gap threshold, thus dampening the initial upward shift.
rates in the environment with homogeneous attention. Moreover, up-front closing costs represent only a small fraction of total origination costs in practice: as documented in Zhang (2022), 80.5% of these costs are rolled into a higher mortgage coupon. Lenders’ origination costs are then recouped via the gain on sale $\pi$. Beyond simple motivations related to liquidity, this choice to roll fixed costs into the rate can also be cast as an optimal decision in an environment with bounded rationality: with sufficiently limited information processing capacity, this choice leads to a simple refinancing problem, which dominates the alternative—paying up-front closing costs and then solving a more complex option exercise problem.

These two observations prompt us to make the following simplifying assumption:

**Assumption 1.** *Households do not face any up-front closing costs (i.e., $\psi = 0$).*

Assumption 1 will simplify our numerical computations in the case of ex ante heterogeneous households and will apply for the remainder of the paper.

We end this section with a discussion of the interpretation of the MPE introduced in Definition 1, connecting the homogeneous environment that we have studied until now to the heterogeneous environment that we will deal with next. When households are heterogeneous in their attention rate but investors can screen on $\chi$, mortgage prices and mortgage market interest rates are type specific, i.e., $m(x, \chi)$, with each type’s mortgage price determined by equation (9), and mortgage market interest rates are determined by the break-even condition (10). Thus, we will sometimes refer to the MPE in the homogeneous case as the *separating MPE*. In the alternative case, when investors do not observe $\chi$ (i.e., in a “pooling” environment), significant complexities emerge.

### 4.2 Heterogeneous households

Suppose now that there is a cross-section of attention types in the population, with cumulative distribution $H(\chi)$ and associated density $h$. Crucially, we assume that investors—for whatever reason—do not or cannot screen on $\chi$. We discuss this assumption and relate it to institutional features of the US agency MBS market in Section 4.3.2. We will interpret this environment as if mortgages are traded in pools, formed at the time when mortgages are originated.
4.2.1 Infinite-dimensional problem

Let $F_t(c, \chi)$ be the joint cumulative distribution over outstanding coupon rates $c$ and types $\chi$ in the population at time $t$, with associated joint density $f_t(c, \chi)$. The relevant state of the economy, from the point of view of mortgage investors who cannot observe the type of each given household, is $S_t := (x_t, F_t)$, consisting of the exogenous latent state $x$ describing the current level of short rates and the infinite-dimensional endogenous cross-sectional distribution $F$ over current coupons and types. The mortgage market interest rate is then $m_t = m(S_t)$.

Let $V(S, c; \chi)$, as defined in (1), be the valuation of all future mortgage liabilities for a type-$\chi$ household with current mortgage coupon $c$ when the state of the economy is $S$. Under Assumption 1, the optimization problem solved by households yields refinancing decisions identical to those in the homogeneous case: households refinance whenever they pay attention and $m_t \leq c_t$.

In a Markov perfect equilibrium, we need to specify the dynamics of the state vector $S_t$. $x_t$ is exogenous and follows a time-homogeneous Markov process. The density $f_t$, instead, evolves endogenously over time with households’ refinancing decisions, according to equations described in our online appendix.

We denote as $P(S, c; \chi)$ the shadow price of a mortgage with coupon $c$, conditional on the knowledge that the related household has attention rate $\chi$, as defined in (9). We will refer to $P(S, c; \chi)$ as a shadow price since investors do not observe $\chi$ and thus cannot trade conditional on $\chi$.

The rate for newly originated mortgages depends on the characteristics of households refinancing at time $t$. These households have a type distribution with density

$$g_t(\chi) = \frac{\int_c (\nu + \chi 1_{c > m_t}) f_t(c, \chi) dc}{\int_{\chi} \int_c (\nu + \chi 1_{c > m_t}) f_t(c, \chi) dcd\chi}$$

with corresponding cumulative distribution function $G_t(\chi)$. To build intuition for how the attention distribution of refinancers $G_t$ differs from the distribution of permanent types $H$, consider the case where everyone’s refinancing option is in the money at time $t$. The origination distribution $g_t$ is then given by

$$g_t(\chi) = \frac{(\nu + \chi) h(\chi)}{\int_y (\nu + y) h(y) dy} = \left( \frac{\nu + \chi}{\nu + \bar{\chi}H} \right) h(\chi),$$

$$18$$
with $\tilde{\chi}_H := \mathbb{E}^H[\chi]$ being the average degree of attention in the population. Conversely, when no one’s refinancing option is in the money at time $t$, the origination distribution $g_t$ then coincides with the population distribution,

$$g_t(\chi) = h(\chi).$$

The attention distribution $G_t$ of refinancers is thus tilted towards higher-attention types relative to the distribution of permanent types $H$ in the population, particularly in low-rate states.

Our perfect competition assumption imposes the following restriction on the mortgage rate function $m(S)$:

$$\mathbb{E}^{G_t} [P(S_t, m(S_t); \chi)] := \int \chi P(S_t, m(S_t); \chi) dG_t(\chi) = 1 + \pi,$$

subject to $g_t$ given by (11) and where the superscript on the expectation indicates the distribution of household types $\chi$ over which the cross-sectional average is computed. We can then define a pooling Markov perfect equilibrium of this economy as follows.

**Definition 2.** A pooling Markov perfect equilibrium (pooling MPE) is defined as (i) a household refinancing policy satisfying (3), (ii) a shadow pricing function $P$ defined via (9), (iii) a joint density $f_t$ with evolution consistent with households’ refinancing decisions, (iv) a mortgage rate function $m(S)$ that satisfies (14), with (v) an origination distribution $G$ that satisfies (11).

Such a pooling MPE, which features an infinite-dimensional state space, is reminiscent of problems in heterogeneous agent consumption–savings models in macroeconomics (see Krusell and Smith (1998) or more recently Ahn et al. (2018)), or sticky-price models in the monetary policy literature (see, for instance, Golosov and Lucas Jr (2007)), with the additional complexity of a zero-profit pricing condition. Rather than addressing the computation of the pooling MPE in general, we will instead make simplifying assumptions that yield tractability while still capturing the main economic forces underlying the mortgage market equilibrium.

### 4.2.2 Simplifying assumption

The equilibrium computation in the pooling environment is significantly more complex than in the separating MPE, as it involves the determination of a fixed point in
the space of functions of infinite-dimensional objects. To make progress, and for the remainder of the paper, rather than attempting to find such a fixed point, we make the following simplifying assumption:

**Assumption 2.** Regardless of the path of \( r_t \), investors price mortgages assuming a cross-sectional origination distribution that is either (i) a constant \( G(\chi) \) or (ii) a state-dependent function \( G(\chi|x) \).

Assumption 2 restricts the cross-sectional origination distribution used for pricing purposes to be dependent at most on the latent state \( x \) rather than on the full time-varying density \( f_t \), and it thus reduces substantially the dimensionality of the relevant state space. While we make this assumption largely for computational tractability, it can also be justified when investors are engaged in \( k \)-level thinking, so that they understand the impact of refinancing incentives on prepayments but do not fully consider how this prepayment behavior then affects the attention distribution of refinancers over time. The strength of Assumption 2 depends on how much the actual origination distribution \( G_t \) dynamically changes and differs from the distribution \( G \) assumed for pricing purposes. When we turn to the equilibrium definition, we will impose a consistency condition, in that \( G \) is either the (i) unconditional or (ii) conditional ergodic average origination distribution \( G_t \); this ensures that investors, while potentially making gains or losses upon their mortgage purchases at each point in time, break even on average.\(^{19}\)

### 4.2.3 Mortgage pricing in the simplified environment

Under Assumption 2, the only relevant aggregate state variable for the investors’ pricing problem is the latent state \( x_t \), and we thus write the mortgage market interest rate \( m_t = m(x_t) \). We continue to use \( P(x, c; \chi) \) for the shadow price of a type-\( \chi \) mortgage. Let \( \bar{P}_t(x, c) \) be the expectation of \( P(x, c; \chi) \) under the origination distribution \( G \), and

\(^{19}\)Our approach resembles the approximation method developed in Krusell and Smith (1998); when \( G \) is the unconditional (conditional) origination distribution over attention, households’ decisions feed back into an aggregate behavior that leads to a constant (state-dependent) distribution \( f \) over coupons and types \((c, \chi)\) and thus a constant (state-dependent) attention distribution for newly originated mortgages \( G \). Our algorithm thus leads us to solve a fixed-point problem, as in Krusell and Smith (1998); in the version of our approximation where the origination distribution is \( G(\chi|x) \), the dynamics of this distribution are nonlinear, whereas the dynamics of the first moments of the relevant cross-sectional distribution in Krusell and Smith (1998) are assumed to be log-linear.
let the market price of a newly issued mortgage pool be

$$\bar{P}_G(x, c) := \mathbb{E}^G[P(x, c; \chi)].$$  \hfill (15)\]

Under Assumption 2, the market equilibrium condition is given by

$$\bar{P}_G(x, m(x)) = 1 + \pi.$$ \hfill (16)\]

Finally, households’ optimal refinancing behavior combined with the mortgage rate function \(m(\cdot)\) implies an ergodic cross-sectional distribution \(f_{\infty}(x, c, \chi)\) and thus an ergodic marginal type distribution for refinancers. The unconditional origination distribution is given by

$$g(\chi) = \frac{h(\chi) \int_x \left( \nu + \chi \int_{c \geq m(x)} f_{\infty}(c | x, \chi) dc \right) f_{\infty}(x) dx}{\int_x h(\chi) \int_x \left( \nu + \chi \int_{c \geq m(x)} f_{\infty}(c | x, \chi) dc \right) f_{\infty}(x) d\chi dx},$$ \hfill (17)\]

while the conditional origination distribution is given by

$$g(\chi | x) = \frac{h(\chi) \left( \nu + \chi \int_{c \geq m(x)} f_{\infty}(c | x, \chi) dc \right)}{\int_x h(\chi) \left( \nu + \chi \int_{c \geq m(x)} f_{\infty}(c | x, \chi) dc \right) d\chi}.$$ \hfill (18)\]

These distributions, as well as Assumption 2, help us build our equilibrium definition:

**Definition 3.** An approximate pooling Markov perfect equilibrium (approximate pooling MPE) is defined as (i) a household refinancing policy satisfying (3), (ii) a shadow pricing function \(P\) defined via (9), (iii) an ergodic joint density \(f_{\infty}(x, c, \chi)\) and its corresponding ergodic marginal density over refinancers \(g\) satisfying either consistency condition (17) (in the unconditional case) or (18) (in the conditional case), (iv) a newly originated pool pricing function \(\bar{P}_G\) defined via (15), and (v) the break-even condition (16).

The separating MPE and the approximate pooling MPE are similar in that they both have a single aggregate state variable, \(x_t\). However, they differ in two aspects. First, the break-even condition of originators in the heterogeneous case is a cross-sectional expectation version of that in the homogeneous case. Second, and most importantly, our approximate pooling MPE requires a consistency condition: the
cross-sectional origination distribution $G$ used by investors when pricing new issue mortgages needs to be consistent with the marginal density of refinancers, as implied by households’ behavior and the corresponding joint density $f_\infty$ over the latent state $x$, coupon $c$ and inattention $\chi$. Last, we introduce the concept of “monotonicity” of an equilibrium as follows.

**Definition 4.** When $x$ is unidimensional and $r(\cdot)$ is increasing, an approximate pooling MPE is “monotone” if the mortgage function $m(x, G)$ is increasing in $x$.

The approximation imposed by Assumption 2 allows us to establish some useful theoretical results and simplifies our numerical calculations.

**Proposition 5.** Let $x$ be unidimensional and $r(\cdot)$ be monotone increasing. Define the candidate mortgage rate

$$m(x; G) := f + \frac{1 + \pi - \mathbb{E}^G [PO(x; \chi)]}{\mathbb{E}^G [IO(x; \chi)]},$$

where $G$ is a distribution defined in (17) (unconditional) or (18) (conditional), where

$$IO(x; \chi) := \mathbb{E}_x \left[ \int_0^{\tau_{x,\chi}} e^{-\int_0^s r_x ds} ds \right], \quad PO(x; \chi) := \mathbb{E}_x \left[ e^{-\int_0^{\tau_{x,\chi}} r_x ds} \right],$$

and where, for any arbitrary $x$, $\tau_{x,\chi}$ is a stopping time with arrival intensity $\nu + \chi \mathbb{1}_{\{x_t \leq x\}}$. If $m(x; G)$ is monotone in $x$, there exists a unique monotone approximate pooling MPE of this economy, with $m(x; G)$ being the equilibrium mortgage rate.

Our proof is constructive; if a monotone equilibrium exists, we compute its related unconditional and conditional origination distributions $G(\chi)$ and $G(\chi|x)$, and we show that the related equilibrium mortgage rate must satisfy (19). Conversely, if the object $m$, defined in (19), is monotone increasing in $x$, then an approximate pooling MPE exists and is unique. What are the properties of the equilibrium mortgage rates in this environment with permanent attention heterogeneity? Our next proposition allows us to be more specific about the impact of this cross-sectional heterogeneity on mortgage rates in the case of the unconditional approximate pooling MPE.

**Proposition 6.** In a monotone unconditional approximate pooling MPE, the pool price $\bar{P}_G$ satisfies

$$\bar{P}_G(x, c) = P(x, c; \bar{X}_G) - \mathbb{E}_x \left[ \int_0^T e^{-\int_0^t r(x_s) ds} \mathbb{1}_{\{m(x_t) \leq c\}} \text{Cov}^G(\chi, P(x_t, c; \chi)) dt \right].$$

22
where $\tau$ is the prepayment time for a household with attention rate $\bar{\chi}_G := \mathbb{E}^G[\chi]$.

Thus, the pool price $\bar{P}_G$ in the market behaves as if it were made up of homogeneous households with attention $\bar{\chi}_G$, with an adjustment equal to the average (conditional on the rate gap being positive) discounted cross-sectional covariance $\text{Cov}^G(\chi, P(x_t, c; \chi))$ between (a) shadow mortgage prices and (b) attention rates. If the shadow price $P$ is “on average” decreasing in $\chi$ whenever the prepayment option is in the money, this correction term is positive. This yields the following corollary:

**Corollary 2.** In a monotone unconditional approximate pooling MPE, if the average (conditional on the rate gap being positive) discounted cross-sectional covariance between (a) shadow mortgage prices and (b) attention rates is negative, then the equilibrium mortgage rate $m(\cdot)$ when households have a nondegenerate origination distribution $G$ is lower than the corresponding equilibrium mortgage rate when households are homogeneous with attention $\bar{\chi}_G$.

In all our numerical computations of the approximate pooling MPE, we find that the correction term in equation (20) is indeed positive. Intuitively, with the average attention rate $\bar{\chi}_G$ held constant, faster households have a shorter effective maturity than slower households. Investors make money off slower households while making losses off faster households, but since the average maturity of slower households is higher than that of faster households, a mean-preserving spread benefits investors by increasing $\int_x P(x, c; \chi) dG(\chi)$. The zero-profit condition then forces investors to pass this benefit on to households in the form of lower mortgage rates.

We end this section by discussing how the interaction between the current interest rate and heterogeneity affects mortgage pricing and the state dependence of mortgage interest rates.

**Proposition 7.** When a monotone approximate pooling MPE exists, denote as $\underline{x}$ the lowest attainable latent state. The lowest mortgage rate $m(\underline{x})$ is invariant to the distribution over permanent heterogeneity $H$.

The proof for Proposition 7 relies on the observation that if a household has managed to lock in the lowest attainable mortgage coupon $m(\underline{x})$, it will never refinance for strategic reasons, only due to an exogenous move. This necessarily means that such a lowest-coupon mortgage has a shadow price that is independent of the attention rate $\chi$. The break-even condition at $x = \underline{x}$ allows us to conclude that $m(\underline{x})$ is invariant to
*H. Proposition 7* gives us some intuition of how the mortgage rate function \( m(\cdot) \) might change as the variance of the distribution \( H \) increases: we expect \( m \) to be relatively insensitive to such variance when rates are low but substantially more sensitive in high-interest-rate states.

## 4.3 Discussion

In this section, we discuss several of our modeling assumptions and highlight how our framework endogenously generates the phenomenon of burnout.

### 4.3.1 Default risk

Since our model is primarily geared toward studying prepayment risk and mortgage rates, we have intentionally abstracted from default risk. Thus, our framework should be seen as an approximation of the US conforming mortgage market, in which the GSEs’ credit guarantee isolates lenders and mortgage investors from default risk.

### 4.3.2 Pooling equilibrium in the US conforming mortgage market

Our focus on a pooling (rather than separating) equilibrium stems from empirical evidence and certain institutional features of the US agency MBS market.

First, using our micro data (see Section 6), conditioning on the time a mortgage is originated, on the FICO score, and on the LTV ratio soaks up a large fraction of the cross-sectional variation in mortgage coupons, suggesting that mortgage originators do not price-discriminate on any dimension other than these two key variables (see also Hurst et al. (2016) for additional evidence on the lack of spatial heterogeneity of mortgage interest rates).

Second, the majority of mortgage lending in the US is funded through the agency MBS market. Fuster, Lo, and Willen (2017) document that between 2009 and 2014, only 20% of loans originated were kept on banks’ balance sheets. As of 2020, 70% of conforming mortgages were originated by specialty mortgage lenders rather than deposit-taking institutions; these finance companies’ sole objective is to originate conforming mortgages and immediately distribute them to investors via the agency MBS market (see Jiang (2019) or Buchak et al. (2018)). To hedge their pipeline, these finance companies sell their origination book forward via the to-be-announced (TBA) market. TBA buyers do not know the exact mortgage pool that they will
receive at settlement. Rather, they know only 5 characteristics of the pool: the agency (Fannie Mae or Freddie Mac), the average coupon rate, the maturity, the face value, and the settlement month. Thus, interest rates on mortgages originated by those finance companies are indirectly linked to the TBA market, in which prices take into account the fact that investors do not know the specific pool characteristics beyond those described above. These considerations provide the rationale for our argument that a pooling equilibrium is relevant in the context of the US conforming mortgage market.

4.3.3 Burnout

Prepayment models used by the financial industry typically include, above and beyond the rate incentive, various factors that capture seasonality, seasoning (the fact that newly originated mortgages tend to have slower prepayment speeds than older ones), and burnout (the fact that pools with higher past cumulative prepayments tend to have slower prepayment speeds than otherwise-comparable pools with lower past cumulative prepayments). Our model with heterogeneous attention rates endogenously produces the burnout effect: in a given mortgage pool, faster households tend to refinance earlier and thus leave the pool faster than slower households, causing a decrease in prepayment speeds as the pool factor declines.

4.4 Redistribution via the mortgage market

We then consider the distributional effects of the pooling equilibrium by type of household. Let \( m(x, G) \) (\( m(x, \chi) \)) be the equilibrium mortgage rate in the approximate pooling MPE (separating MPE for type-\( \chi \) households).

In a pooling environment, fast (slow) households face lower (higher) mortgage market interest rates than in the presence of price discrimination based on type \( \chi \); given that investors break even on average, the pooling environment necessarily leads to cross-subsidies among households. To quantify the extent of this redistribution, one can, for example, focus on \( m(x, \chi) - m(x, G) \), i.e., the difference in mortgage rates.

---

20See Fuster, Lucca, and Vickery (2022) for a detailed discussion on the institutional features of the US MBS market and, in particular, the role of the TBA market.
21See, for instance, Spahr and Sunderman (1992).
22This effect has also been noted by Stanton (1995) in a model where households have heterogeneous fixed costs of prepaying.
that a type-χ household faces in the separating vs. approximate pooling MPEs. When this difference is positive, type-χ households benefit from a subsidy, and it coincides with investors losing money on the specific mortgage \( P(x, m(x, G); \chi) < 1 + \pi \).

However, investors’ actual profit or loss on a particular mortgage provides only a partial picture of the cross-subsidies in the model since attention is a permanent household attribute. Indeed, households with different attention rates have different effective maturities of their current mortgages, and they benefit from (if they are of the fast type) or are hurt by (if they are of the slow type) the difference between \( P(x, m(x); \chi) \) and the initial mortgage price \( 1 + \pi \) not just in their current mortgage but also every time they refinance in the future. As an alternative measure of redistribution, we thus consider \( \mathbb{E}[c_t|\chi, \text{pooling}] \), the ergodic average coupon paid by type-χ households in the approximate pooling MPE relative to its cross-sectional average \( \mathbb{E}[c_t|\text{pooling}] \). This calculation takes into account not only the subsidies/taxes obtained by a household for a given mortgage but also those obtained for all future mortgages on average. The difference in these ergodic averages across household types stems from ex post differences in refinancing rates rather than from equilibrium forces. If one wants to factor equilibrium effects into our measure of redistribution, one needs to instead consider \( \mathbb{E}[c_t|\chi, \text{separating}] \)—i.e., the ergodic average coupon rates paid by household types in the separating MPE.

These various measures of redistribution will be considered when we study policy proposals and perform counterfactual calculations, as described next.

## 5 Policy evaluations and counterfactuals

We now turn to various policy proposals put forth in the academic literature and in policy circles, primarily aimed at improving mortgage borrowers’ welfare by reducing costs induced by widespread financial mistakes and the lack of financial literacy. Our structural model allows us to evaluate these proposals via counterfactual analysis. We lay out the various counterfactuals of interest and discuss the theoretical mechanisms

---

23 Alternatively, assume that household type \( \chi \) is not directly observable in the data, but suppose instead that one could separate households along observables into \( i \in I \) distinct groups with group distribution \( G_i \), so that \( G(\chi) = \sum_{i \in I} w_i G_i(\chi) \), with \( w_i \) being the weight of each respective group. Then, \( \Delta_i m(x) := m(x, G_i) - m(x, G) \) captures the cross-subsidy to group \( i \).

24 The expectation operator \( \mathbb{E} \) integrates over the ergodic density of the dynamic system, conditional on the equilibrium being the approximate pooling MPE.
at play, and in Section 6 and Section 7, we take this analysis to the data.

5.1 Automatically refinancing mortgages

Consider the introduction of automatically refinancing mortgages (auto-RMs), as suggested for instance by Keys, Pope, and Pope (2016) or Campbell et al. (2011). With an auto-RM, a household pays the minimum realized mortgage rate since the mortgage’s inception at time $\tau$:

$$m_t \equiv \min_{\tau \leq s \leq t} \{m_s\}. \quad (21)$$

The contractual rate of the product is thus tied to the minimum process of the mortgage market interest rate. On its face, the auto-RM seems like a great idea for inattentive or financially unsophisticated households since it reduces the impact of financial mistakes on the cost of these households’ lifetime mortgage liabilities. However, discussions around this proposal are usually cast in terms of the partial equilibrium and thus fail to take into account the equilibrium response of mortgage rates. Our model can speak to this response.

To ensure the existence of an MPE in this environment, we make the following smart-contract assumption:

**Assumption 3.** No origination costs are incurred at the time of automatic rate resets. The equilibrium rate $m_t$ is such that the price of a newly issued auto-RM is equal to $1 + \pi$.

Under Assumption 3, refinancing events occur automatically; a change in rates can then be viewed as a rate reset, just like under adjustable rate mortgages. However, unlike under an adjustable rate mortgage, this adjustment process is asymmetric: rates adjust down when the market rate declines but do not adjust up when the market rate rises. Origination costs are incurred only at the time households move, prepay the principal balance of their existing mortgage, and take on a new mortgage at the current auto-RM market rate, denoted (with an abuse of notation) $m(x, \infty)$.

---

25 See Agarwal, Rosen, and Yao (2016) for a quantification of the cost of these mistakes.

26 Under Assumption 3, $m(x, \infty)$ is the limit of the separating MPE’s mortgage market interest rate $m(x, \chi)$ as $\chi \rightarrow +\infty$ when we set the gain on sale $\pi = 0$, and thus, we also use this notation for $\pi > 0$, cognizant of the fact that we assume no origination costs are incurred upon rate reset under Assumption 3. Without this assumption, we would not have an equilibrium in the limit, as discussed in the online appendix.
We relegate all technical details to the online appendix.

We make three observations about this environment. First, even though households may still have heterogeneous attention rates, this heterogeneity is irrelevant for pricing purposes due to the mortgage contract design. This case is thus equivalent to an economic environment in which households no longer face any refinancing frictions, i.e., an environment without cross-subsidies.

Second, traditional fixed-rate prepayable mortgages trigger origination costs upon refinancing that are recovered by lenders via a combination of (i) up-front closing costs $\psi$ paid by borrowers and (ii) the gain on sale $\pi$ extracted from secondary market mortgage investors. If these various costs are dead-weight losses, then, under Assumption 3, the auto-RM must be a more efficient contract since it removes the incidence of these costs at rate resets.

Third, since mortgage market investors continually receive coupons according to the minimum realized rate, borrowers almost always “underpay” for mortgage financing relative to their cost in the case where they obtain funds via a more traditional adjustable rate mortgage. Lenders optimally charge a rate $m(x, \infty) \geq r(x) + f$ to break even against the backdrop of future refinancings at lower rates.\(^{27}\) This discussion is formalized in our next proposition, which we prove in the online appendix:

**Proposition 8.** The auto-RM rate satisfies $m(x, \infty) \geq r(x) + f$ for all $x$.

Next, consider what the introduction of the auto-RM does to an environment with cross-sectional heterogeneity in $\chi$. Initially, all households have a traditional fixed-rate prepayable mortgage. In the resulting pooling MPE, at the time of a refinancing, the slowest households overpay for their mortgage, whereas the fastest households underpay. Thus, the former will find it beneficial to migrate to the auto-RM when the opportunity arises since they can obtain an actuarial “fair” rate with no inherent cross-subsidies. As the slowest households migrate towards the auto-RM, the effective attention rate of households left on the traditional mortgage contract increases, pushing those mortgage rates higher in equilibrium. The slowest households remaining in the traditional mortgage contract now subsidize the fastest ones and will thus find it beneficial to migrate to the auto-RM, further pushing up traditional

\(^{27}\)Our auto-RM framework shares many similarities with models of wage determination with stochastic productivity, a risk-averse worker and one-sided commitment by the firm—see, for instance, Harris and Holmstrom (1982) and Section 8 for a greater discussion on the mapping between the two models.
mortgage rates and skewing the attention distribution of households remaining on the traditional contract even more towards higher and higher attention rates. This unraveling continues until only the highest type is left in the traditional mortgage market. This discussion leads to the following proposition:

**Proposition 9.** With heterogeneous attention rates and the ability of households to use either (i) traditional fixed-rate prepayable mortgages or (ii) auto-RMs, absent any other financial constraints, all households migrate to the auto-RM.

What could undo this unraveling, even in the presence of the natural advantage that the auto-RM enjoys over a traditional mortgage due to Assumption 3? First, some households might not fully understand or be able to value the refinancing option embedded in the auto-RM. When faced with rates $m(r, G) < m(r, \infty)$, they simply gravitate towards the cheaper rate, even though the expected net present value of future interest costs is lower under the auto-RM than under the traditional mortgage.\(^{28}\)

Second, some households might be financially constrained, so that picking the cheaper rate associated with a traditional mortgage may just allow them to purchase their target home, while the auto-RM rate might lead to initial mortgage payments on their target home that are too high. In other words, the disutility of a suboptimal home allocation might outweigh the cross-subsidies and deadweight costs associated with refinancing inherent in the traditional mortgage.

Third, the auto-RM might not be the most desirable option if a household is risk averse, given the high associated cash-flow volatility in comparison to that under a more traditional mortgage. The welfare impact of this type of contract, in the presence of risk-averse households, depends on the comovement of income with rates.

### 5.2 Improving financial decisions

The attention distribution $H$ is a stand-in for a range of frictions, one of which is financial literacy. The cross-subsidies embedded in the pooling MPE—from inattentive to attentive households—suggests that policies raising financial literacy for the least sophisticated households may appear attractive to policymakers interested in reducing mortgage inequality. Moreover, if financial literacy positively correlates with income

\(^{28}\)Absent any required gain on sale ($\pi = 0$), $m(r, G) < m(r, \infty)$ is a natural outcome. In contrast, when gains on sale are required for originators to recoup their costs (i.e., $\pi > 0$) and under assumption Assumption 3, this might not hold when $\bar{\chi}_G$ is large enough.
(as we will argue later), such policies will also have efficiency consequences—as they will tend to reduce the average mortgage interest expense (and thus increase disposable income) for households with a relatively high marginal propensity to consume. Our framework, however, emphasizes that this type of intervention—even when properly targeted—has equilibrium effects that tend to hurt untreated households, via an increase in equilibrium mortgage rates.

A rightward shift in the attention distribution $H$ can also occur absent policy intervention, for instance with the evolution of the mortgage market. Buchak et al. (2018) document, over the last decade, a significant increase in the share of all mortgages originated by specialty finance companies, reaching 50% in 2015. Financial technology (fintech) lenders’ market share has also been rising during this time period, accounting for 8% of total US mortgage issuance as of 2016 (see Fuster et al. (2019)). In Section 7.4.2, we argue that these recent trends in mortgage origination and servicing have improved the effective attention rate of the borrower population, with an impact on equilibrium mortgage rates that can be quantified through the lens of our model.

6 Household attention in mortgage refinancing data

We now estimate the level and degree of cross-sectional heterogeneity of attention rates in the population of US conforming mortgage borrowers, which is necessary for quantifying the equilibrium consequences of the mortgage market interventions discussed in the previous section. We use two separate datasets for this purpose: (1) a month–household panel dataset extracted from Equifax Credit Risk Insight Servicing McDash (CRISM), which allows us to track various mortgage statistics for a sample of approximately 250,000 households, and (2) a month–mortgage panel obtained from Fannie Mae’s Single-Family Loan Performance (SFLP) datasets, which offers a longer sample period and more detailed covariates than the CRISM dataset but does not allow us to track individual households. Our online appendix provides greater details on the data.
6.1 Individual attention rates

Using CRISM, for each borrower $i$, we compute the number $t_i$ of “effective” periods—i.e., the number of months across a borrower’s loans during which the loan’s coupon rate is $\theta$ bps above the effective mortgage rate that month, i.e., $\text{gap} > \theta$.\(^{29}\) Next, we sum the number of monthly refinancing events for each borrower across the sample, $s_i$. Under our assumptions, each borrower $i$’s successes $s_i$ in $t_i$ trials follows a binomial distribution with a probability of attention and thus refinancing in a given effective month equal to $p_i = p(\chi_i) := 1 - \exp(-\chi_i/12)$. Given the “Calvo” assumption implicit in our modeling of household inattention, for a given $t_i$ and $\chi_i$, the realization of $s_i$ is i.i.d.\(^{30}\) The maximum likelihood estimate (MLE) for an individual’s $p(\chi_i)$ is then $\hat{p}_i = s_i/t_i$, the number of successes over the total number of trials.

The left panel of Figure 3 shows the empirical cumulative distribution function (CDF) of refinancing propensities $\hat{p}_i = s_i/t_i$ for different rate gap cutoffs, all weighted by households’ average outstanding loan balance. The right panel shows the corresponding CDF for the attention rate $\hat{\chi}_i = -12 \cdot \log(1 - \hat{p}_i)$. The first—striking—observation is the large number of households that are completely inattentive: at least 70% of them have $\hat{p}_i = \hat{\chi}_i = 0$. While the distribution over $\hat{\chi}_i$ appears to be spread out, this observation in and of itself does not allow us to draw conclusions about the presence of heterogeneity in attention rates in the data.

The observed heterogeneity in $\hat{\chi}_i$ could reflect either (i) an underlying type difference in $\chi_i$, i.e., true attention heterogeneity, or (ii) randomness in the binomial distribution, i.e., sampling noise. In other words, even if our population were homogeneous

\(^{29}\)We reintroduce the threshold $\theta$ here for empirical accuracy. As discussed previously, given a distribution of attention rates, fixed costs have only second-order effects on pricing, since omitting them leads our model to overestimate the number of “small” rate refinancings and these refinancings do not alter the equilibrium pricing function substantially. However, in measuring the distribution of attention rates, getting small refinancings wrong does impact the estimated attention distribution. In the absence of up-front closing costs, any threshold $\theta > 0$ results in a consistent estimation of attention rates. However, in the presence of up-front closing costs $\psi > 0$, a balance has to be struck to estimate attention rates—high $\theta$ yields more consistent attention rates but rapidly reduces the sample size, while low $\theta$ underestimates attention rates by attributing inaction in periods of small positive rate gaps to inattention instead of fixed costs.

\(^{30}\)In this preliminary approach, we treat the effective individual sample length $t_i$ as an exogenous input in our MLE procedures. We thus ignore the fact that for a given interest rate path $\{m_t\}_{t\geq0}$, households with higher attention rate $\chi$ will have on average a lower number of “effective periods” $t_i$ than households with lower $\chi$. In other words, even though we treat the number of “effective periods” $t_i$ as exogenous for our estimation of $\chi$, within the model, the two are linked once we move away from the homogeneity assumption. Later, our GMM estimation of the attention distribution will address this shortcoming rigorously.
in the attention rate, randomness in $s_i$ and $t_i$ would still lead to a cross-sectional distribution in observed $\hat{\chi}_i$. We thus need to disentangle the true heterogeneity from sampling noise.

6.2 Testing for a unique homogeneous group

We first check whether our refinancing data contain some heterogeneity in attention by testing whether the data could have been generated by a homogeneous group of households with common attention rate $\chi$. We perform a two-sided Kolmogorov–Smirnov test between the empirical distribution of $\hat{p}_i$ and a simulated distribution of $\hat{p}^{\text{sim}}_i$. This simulated distribution utilizes the maximum likelihood estimator for $\chi$ under the assumption that the refinancing data are generated by a single group, which is simply $\hat{p}_{N=1} = \sum_i^{N_{HH}} s_i / \sum_i^{N_{HH}} t_i$. We generate the distribution by repeatedly sampling the binomial distribution with probability $\hat{p}_{N=1}$ for a given vector of $t_i$s.\footnote{We numerically derive the theoretical distribution of $\hat{p} = s/t$ given $t$ by taking 10,000 samples from the binomial distribution with $p = 1 - e^{-\chi dt}$ over the vector of trials $t$ and then merging the samples to form the distribution of $\hat{p}^{\text{sim}}$.}

The Kolmogorov–Smirnov test rejects our assumption of household homogeneity in attention rates. In other words, not all of the observed cross-sectional heterogeneity in $\hat{p}_i$ can be attributed to sampling noise.
6.3 Measuring the attention distribution $H(\chi)$

Next, we use a clustering algorithm to quantify the degree of cross-sectional attention heterogeneity in the data. Our approach allows us to explicitly account for sampling noise by using all the information in an observation $(s_i, t_i)$ rather than relying only on the ratio $\hat{p}_i = s_i/t_i$.\footnote{In particular, our estimation treats an observation of $s_i = 1$ refinancing success out of $t_i = 2$ trials differently from an observation of $s_i = 10$ refinancing successes out of $t_i = 20$ trials, even though they both result in the same observed $\hat{p} = 1/2$.} For a given number $N$ of homogeneous groups, we use maximum likelihood (ML) to estimate the group-specific rates $\chi_k$ for $k \in \{1, ..., N\}$, and we allocate each individual $i$ into a group $k$ to derive the estimated weight of each group.\footnote{The optimal number $N$ of groups is an open question, and for the moment, we abstract from it.} We choose $N = 5$ and $\text{gap} > 0.5\%$, and in the online appendix, we verify the robustness of our conclusions to alternative choices. Figure 4 (left panel, blue bars) displays the results of our estimation. A total of 81.1\% of households in our sample are estimated to be almost completely inattentive, while about 1.4\% of households are estimated to be very attentive, paying attention to mortgage markets (and refinancing when their prepayment option is in the money) once every 2.3 months. The remainder of households—approximately 17.5\%—have attention rates between these two extremes and fall into the remaining three groups. The resulting average attention rate is $\bar{\chi}_H = 19\%$, yielding an average $\mathbb{E}^H[p(\chi)] = 1.42\%$ monthly attention probability.

We then recover the ergodic unconditional (conditional) distribution of refinancers $G(\chi)$ ($G(\chi|x)$) in our approximate pooling MPE using the estimated distribution $H$, under the assumption that the interest rates and other model parameters are those used in Section 7 and described in Table 1, with the interest rate being unidimensional $r(x) = x$.\footnote{We verify that the resulting approximate pooling MPE is monotone for both (i) the unconditional and (ii) the conditional case and thus unique in both (i) and (ii). Our numerical derivation of $G$ closely follows our theoretical derivation from the proof of Proposition 5.}

The left panel of Figure 4 shows the ergodic unconditional origination distribution $G(\chi)$ that we will use for our quantitative analysis in Section 7, while the right panel shows the corresponding conditional distribution $G(\chi|r)$, which we will use to test the robustness of our conclusions to relaxation of Assumption 2. Both origination distributions overrepresent high-$\chi$ types and underrepresent low-$\chi$ types relative to the population distribution $H$, as discussed in Section 4.2.1. This observation holds
Figure 4: Ergodic origination distribution $G(\cdot)$ implied by $H(\cdot)$. The left panel shows the unconditional population (origination) distribution $H(G)$ in the left blue bars (right orange bars). The estimation focuses on households and months with gap > 0.5%, weighted by the average loan amount. The right panel shows the conditional origination distribution $G(\chi|r)$ (solid lines) and the unconditional population distribution $H(\chi)$ (thick dots).

for the conditional distribution $G(\chi|r)$ for all but the highest-interest-rate state, at which point no one voluntarily refinances. As the right panel of Figure 4 shows, the distortion in representation is especially strong in low-interest-rate environments, in which the most inattentive households drop from a population weight of 81.1% to 40.7% while the most attentive households increase from a population weight of 1.4% to 9.3%. The unconditional origination distribution $G$ is also skewed towards more attentive households, with $\bar{\chi}_G = 53\%$ that is greater than $\bar{\chi}_G = 19\%$.

Our clustering procedure thus illustrates the substantial degree of cross-sectional heterogeneity in attention rates in the population. In our online appendix, using a similar procedure, we investigate the observable household characteristics that are related to households’ attention. We show that high-FICO borrowers tend to exhibit greater attention than borrowers with a lower credit score, potentially reflecting a greater degree of financial frictions faced by these latter households. Similarly,

---

$^{35}$The state $r = \bar{r}$ corresponds to the situation described by (13); the ergodic average conditional distribution $G(\chi|\bar{r})$ is thus equal to the distribution $H(\chi)$ of permanent heterogeneity.
borrowers with larger loan balances tend to be more attentive than borrowers with smaller loan balances; for these latter households, the cost of making mistakes in refinancing is lower than that for households with larger mortgages.

7 Quantitative implications

In this section, we use our equilibrium model of mortgage rate determination to study quantitatively the various policies and counterfactuals discussed in Section 5.

7.1 Estimation/calibration of remaining model parameters

We use the cross-sectional attention distribution estimated in Section 6.3 via our clustering procedure. The short-term interest rate $r_t$ follows a one-factor, square-root diffusion process as in Cox, Ingersoll Jr, and Ross (1985). We take as the relevant short rate the 3-month treasury rate and estimate the parameters of our term structure model via MLE on a sample covering 1971 to 2021.\(^{36}\)

The parameter $\nu$ can be interpreted as the sum of a moving and maturity intensity. We set the moving intensity $\nu_{mov} = 4.1\%$, consistent with the estimate in Berger et al. (2021). Since our empirical work focuses on 30-year mortgages, we assume a maturity intensity $\nu_{mat} = 3.3\%$. Thus, the “forced” prepayment intensity is assumed to be $\nu = \nu_{mov} + \nu_{mat} = 7.4\%$.

We set the wedge between mortgage payments made by households and cash receipts by mortgage investors to $f = 0.45\%$, consistent with the estimated ongoing portion of G-fees paid to the GSEs as of 2019 (see the 2019 Federal Housing Finance Agency (FHFA) report on guarantee fees).\(^{37}\) Last, since we assume no closing costs borne by the household ($\psi = 0$), since 80% of origination costs are financed via higher rates, and since the average cost of mortgage intermediation is 4.6% (see Zhang 2021).

\(^{36}\)In other words, we set $r(x) = x$, $\mu(x) = \kappa(\mu - x)$ and $\sigma(x) = \sigma \sqrt{x}$. The parameters to estimate are thus the long-run mean $\mu$, the speed of mean reversion $\kappa$, and the volatility parameter $\sigma$.

\(^{37}\)In principle, the agency MBS coupon is lower than the average loan pool interest rate due not only to the ongoing portion of the G-fees but also to the 25 bps base servicing fee. However, this servicing fee belongs to the originator, who can monetize it by selling the “mortgage servicing rights” (MSRs). In our quantitative application, the market value of the MSR is included in the gain on sale $\pi$ realized by mortgage lenders when they sell a loan that they originate. A portion of the gains on sale stems from the premium to par in the TBA market, and the balance stems from the value of the MSR. See Fuster, Lo, and Willen (2017) for a similar discussion on the decomposition of the gain on sale into a TBA premium and the MSR value.
(2022)), we set the gain on sale to \( \pi = 80\% \times 4.6\% = 3.68\% \). Table 1 summarizes our parameter choice. We solve our model using a standard finite-difference method.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>0.035</td>
<td>Long-run short rate mean</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>0.13</td>
<td>Mean reversion coefficient</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.06</td>
<td>Volatility</td>
</tr>
<tr>
<td>( \nu )</td>
<td>0.074</td>
<td>Total unconditional prepay rate</td>
</tr>
<tr>
<td>( f )</td>
<td>0.0045</td>
<td>Ongoing portion of G-fees</td>
</tr>
<tr>
<td>( \pi )</td>
<td>0.0368</td>
<td>Gain on sale</td>
</tr>
</tbody>
</table>

Table 1: Baseline parameter values

7.2 Validation of our equilibrium mortgage pricing

We first compare our model-implied mortgage rate function \( m(x, G) \) with its data counterpart. Since we estimated the household attention distribution using a sample of households observed between 2005 and 2021, we use the same time period to make this comparison. Given our one-factor term structure model of interest rates, specifying the yield at a single maturity characterizes the entire term structure and reveals the latent state \( x \). We choose to focus on the 10-year constant maturity zero-coupon Libor swap rate and retrieve the implied short-term interest rate \( r(x_t) \), which we then use to compute the relevant model-implied mortgage rate \( m_t = m(x_t, G) \).\(^{38}\)

In Figure 5, we plot the time series of the model-implied mortgage rates and the 30-year mortgage rate from Freddie Mac’s primary mortgage market survey, as reported by the St. Louis Fed. Our model-implied mortgage rate not only is highly correlated with its data counterpart but also has a time-series average that is only slightly below what we measure in the data—with a difference of 17 bps pa, which can be attributed to our assumption of a zero option-adjusted spread.\(^{39}\) The only notable

\(^{38}\)We choose the Libor swap curve as our benchmark interest rate since agency MBSs trade at an option adjusted spread (OAS) relative to the Libor swap curve. In our model, we implicitly assume that the OAS is zero. To construct the 10-year constant maturity zero-coupon Libor swap rate, we add 35 bps to the 10-year constant maturity treasury rate reported by the St. Louis Fed, where the 35 bps corresponds to the average 10-year swap spread over the sample period 2000–2017.

\(^{39}\)See Boyarchenko, Fuster, and Lucca (2019) for an extensive discussion of the concept of the option-adjusted spread in the agency MBS market.
difference between our model-implied rates and their empirical counterpart appears for the financial crisis period of 2008 and early 2009, when stress in financial markets and questions surrounding the implicit government backing of the GSEs resulted in very wide mortgage spreads. We could better capture this episode by modeling a credit spread factor as in Chernov, Dunn, and Longstaff (2018), but modeling this agency risk is outside the focus of our analysis. Overall, the model shows a good fit to the actual mortgage data given the simple one-factor term structure interest rate process that we assume.

### 7.3 Mortgage rates and redistribution

We next study the quantitative impact of the cross-sectional attention heterogeneity on mortgage rates and its redistributive consequences.

The left panel of Figure 6 shows the mortgage rate \( m(r, G) \) in the approximate pooling MPE and the counterfactual rate \( m(r, \bar{\chi}_G) \) in the hypothetical environment where households are homogeneous with attention rate \( \bar{\chi}_G \). Both mortgage functions are increasing in the short rate, with identical values at \( r = 0 \) but with \( m(\cdot, \bar{\chi}_G) \)
Figure 6: **Equilibrium mortgage rates and ergodic coupons.** The left panel shows equilibrium mortgage rates for (i) the approximate pooling MPE $m(r, G)$ (solid blue), (ii) the MPE with homogeneous households with attention $\bar{\chi}_G = 53.14\%$, $m(r, \bar{\chi}_G)$ (dotted green), (iii) the separating MPE $m(r, \chi)$ for type $\chi = \chi_1$ (dot-dashed purple) and $\chi = \chi_5$ (dashed red), and (iv) the auto-RM equilibrium $m(r, \infty)$ (dashed brown). The right panel shows the ergodic coupons as a function of attention $\chi$ for (i) the approximate pooling MPE (solid blue) and its cross-sectional average (double-dashed red), (ii) the separating MPE (dashed black), and (iii) the auto-RM equilibrium (dot-dashed green).

lying above $m(\cdot, G)$ otherwise. The ergodic average difference in mortgage rates is approximately 120 bps pa, highlighting the nontrivial impact of cross-sectional heterogeneity in attention on the *level* of mortgage market rates.

In the same figure, we plot the corresponding separating MPE mortgage rates $m(r, \chi)$ for the slowest ($\chi = \chi_1 = 0\%$ pa) and fastest ($\chi = \chi_5 = 528\%$ p.a.) groups of households according to our clustering algorithm. The separating MPE mortgage rates for intermediate attention levels will be located in between these two curves, since these mortgage rates are increasing in attention (see Proposition 4). Once again, fast households benefit tremendously from the pooling environment; with price discrimination based on type, they would face mortgage rates with an ergodic average that is approximately 380 bps pa higher than that of the slowest households and 350 bps higher than those that they face in the approximate pooling MPE.
As noted before, fast (slow) households benefit from (are hurt by) cross-subsidies in connection not only with a given mortgage but also with all future refinancing transactions. To capture the full scope of this redistribution, we plot in the right panel of Figure 6 the ergodic average coupon $\mathbb{E}[c_t | \chi, \text{pooling}]$ as a function of attention $\chi$ in the approximate pooling MPE and its corresponding population cross-sectional average (under distribution $H$) $\mathbb{E}[c_t | \text{pooling}]$. The ergodic average coupons in the approximate pooling MPE decrease as $\chi$ increases, with the fastest group paying on average 90 bps less than the slowest group and 80 bps less than the “average” household. Importantly, this difference stems purely from the ex post difference in household refinancing rates rather than from equilibrium forces.

To capture the distributional effects stemming from the pooling—rather than separating—equilibrium, we also plot the ergodic average coupon $\mathbb{E}[c_t | \chi, \text{separating}]$ as a function of attention $\chi$ in the separating MPE. The ergodic coupon is upward sloping as a function of attention $\chi$, mainly due to the need for lenders to sell newly issued mortgages at a premium $\pi$ to recoup their origination costs—and thus the need to charge higher rates to generate such gains on sale. The degree of redistribution arising from the equilibrium effects is, in magnitude, as important as that arising from the ex post differences in attention; for instance, while the ergodic average coupon paid by the fastest households in the approximate pooling MPE is 80 bps below that paid by the “average” household in that equilibrium, it is also 180 bps lower than what such household would pay on average in the separating MPE.

The differences in the two panels of Figure 6 are substantial; however, as a household’s attention rate is not observable, these differences do not speak directly to the degree of redistribution among households with different observable characteristics. We address this issue in our online appendix by estimating the cross-subsidies for groups of households formed based on their FICO score or mortgage balance (measured at the household level) and a host of other covariates (measured at the zip code level). Overall, our computations suggest that the cross-subsidies that we document are regressive—in the sense that lower-income, lower-FICO, and younger households tend to be less attentive and thus make on average greater mortgage interest payments than higher-income, higher-FICO and older households.

---

40 Absent the gain on sale (when $\pi = 0$), the ergodic average coupon in the separating MPE is relatively insensitive to the level of attention; the only source of sensitivity stems from discount rate effects.

41 In particular, our online appendix shows that the ergodic average coupon for households with
7.4 Evaluating policy proposals via model counterfactuals

In this section, we quantitatively evaluate the policy proposals and counterfactuals discussed in Section 5. We first focus on automatically refinancing mortgages and then turn to the impact of the rise of nonbank and fintech mortgage lenders on market interest rates.

7.4.1 Automatically refinancing mortgages

As described in Section 5.1, consider the introduction of an auto-RM, i.e., a mortgage whose coupon rate automatically resets to the prevailing market rate if that rate is below the mortgage coupon. The dashed brown line in Figure 6 depicts the equilibrium rate of an auto-RM. As stated in Proposition 8, the auto-RM rate (net of fees) is always above the short rate, i.e., \( m(r, \infty) - f \geq r \). The auto-RM rate is also systematically higher than the mortgage rate in the approximate pooling MPE, i.e., \( m(r, \infty) \geq m(r, G) \); the ergodic average difference between the two rates is 130 bps, highlighting the substantial increase in initial rates with a move from the traditional mortgage to this new financial instrument. To assess the potential effect of such an increase in equilibrium mortgage interest rates on households’ housing and mortgage choice, we plot in the online appendix the debt-to-income (DTI) distribution at origination observed in our SFLP data vs. the counterfactual DTI distribution that would be prevalent in a world where households had access only to the auto-RM product. Focusing on the 43% DTI cutoff—the limit below which mortgages, until 2021, satisfied the “qualified mortgage” definition of the Consumer Financial Protection Bureau—approximately 20% of borrowers would be pushed above the DTI cutoff, potentially forcing them to downsize their house or increase their down payment upon purchase.

The right panel of Figure 6 shows that the ergodic average coupon in the auto-RM equilibrium (dot-dashed green line) is lower than the ergodic average coupon paid on average by households in the approximate pooling MPE—even though initial rates are higher, i.e., \( m(r, \infty) \geq m(r, G) \). In a partial equilibrium setting, in which we would hold \( m(r, G) \) constant, the ergodic average coupon for households able to entirely overcome their inattention friction would be \( \mathbb{E}[c_t | \chi = \infty, \text{pooling}] = 3.5\% \) (purple below-median FICO (below-median income, measured at the zip code level) is 6 bps (2 bps) higher than that for households with above-median FICO (above-median income).
dotted line)—i.e., 60 bps lower than when mortgage rates adjust to the equilibrium auto-RM rate. This calculation highlights the need to factor in the equilibrium response of mortgage rates in considering alternative contract designs: while some of these designs might appear to be substantially beneficial to households when prices are held fixed, their effect can be considerably dampened by equilibrium responses. In the approximate pooling MPE, all groups—besides the slowest one—have an ergodic average coupon below that in the auto-RM equilibrium. Thus, the fastest households are hurt by the introduction of the auto-RM given our unraveling argument from Proposition 9, while only the slowest households benefit.

7.4.2 The rise of nonbank mortgage lending

The mortgage industry has witnessed a structural shift over the past 20 years, with nonbank lenders responsible for a growing share of mortgage origination at the expense of traditional banks. Using SFLP data and the bank vs. nonbank classification of Buchak et al. (2018), we plot the share of mortgage origination by banks, nonbanks and “others” over time in Figure 7 (left panel). While banks had a market share of conforming mortgage originations greater than 75% in 2000, that share has consistently declined. At the same time, nonbank originators’ volumes have increased from less than 5% of total originations to more than 30% as of the end of 2021.

Using the SFLP data, we study the differential prepayment propensity for mortgages originated by banks vs. nonbanks as a function of rate gaps. We estimate the linear probability model

$$
\text{prepay}_{i,j,t} = \mathbb{1}_{\text{bank}} \beta_{\text{gapbin,bank}} \mathbb{1}(\text{gapbin})_{j,t} + \mathbb{1}_{\text{non-bank}} \beta_{\text{gapbin,non-bank}} \mathbb{1}(\text{gapbin})_{j,t} + \beta_X X_{i,j,t} + \epsilon_{i,j,t}
$$

for borrower $i$ with mortgage contract $j$ at time $t$, where $X$ is a vector of controls. Figure 7 shows the point estimates for $\beta_{\text{gapbin,bank}}$ and $\beta_{\text{gapbin,non-bank}}$ in our fully saturated specification, and where the bins used are constructed in 50 bps intervals. The “S-curves” estimated—and, in particular, the difference in levels between the negative and positive rate gaps—directly give us the average attention rate for bank- and nonbank-originated mortgage borrowers. On average, borrowers with bank-originated mortgages tend to be 100 bps per month less attentive than borrowers with

---

42The identity of the originator (or “seller”) is not available in the CRISM dataset. The SFLP data do not include the identity of the seller for each loan; instead, each month, sellers whose combined at-issuance unpaid principal balance is less than 1% of total issuances are classified as “others”.

41
Figure 7: **Rise in nonbank lending.** The left panel shows the fraction of new mortgages classified as originated by “banks”, “nonbanks” and “unknown” originators. The right panel shows our nonparametric estimate of the impact of the rate gap on prepayment rates in the SFLP loan sample.

nonbank-originated mortgages—a substantial difference in behavior. To reduce the risk that households borrowing from nonbank mortgage lenders are systematically different from those borrowing from banks, we saturate our regression with a battery of contract- and household-specific controls within the vector $X$.

Our results are consistent with those in Fuster, Lucca, and Vickery (2022), who conclude that faster prepayment speeds on fintech-originated mortgages stem from higher refinancing propensities rather than from selection of borrowers into fintech loans.

Using our estimate that nonbank lenders increase households’ effective attention rates by 12 pp pa, we can compute the extent to which the rise of nonbank mortgage lending puts upward pressure on mortgage rates. In a counterfactual approximate pooling MPE in which households’ attention rates are uniformly increased by 12%, ergodic average mortgage rates would increase by approximately 50 bps. This counterfactual thus highlights that the dynamics of the mortgage origination market, and in particular the aggressive nudging practices of nonbank lenders, can have a large

---

43 Those controls include (i) a fully nonparametric function of the borrower’s original FICO score, (ii) a fully nonparametric function of the borrower’s original combined LTV ratio, (iii) a first-time home buyer flag, and (iv) the borrower’s original real income. The SFLP data do not contain a borrower ID, only a loan ID; we are thus unable to include household fixed effects in our regression.
and systematic effect on mortgage market rates.

8 General framework

While we focus on the US residential mortgage market, our modeling approach is more general and can be applied to other environments where economic agents are ex ante heterogeneous and make dynamic discrete choices about entering into or renewing a long-term (non–state-contingent) contract subject to some frictions and the other side of the market is competitive but cannot price-discriminate for informational or legal reasons. We discuss below two alternative economic settings in which our framework can be applied: the labor market and the small business credit market.

8.1 Wage setting in labor markets

Consider a model of wage determination with stochastic productivity, risk-averse workers and one-sided commitment by the firm, as in Harris and Holmstrom (1982). Each worker has productivity $x_{it}$ that follows a time-homogeneous Itô process with drift $\mu(x)$ and diffusion $\sigma(x)$. For simplicity, assume that individual worker productivity shocks are purely idiosyncratic. Workers are heterogeneous in their job-hunting rate $\chi$—the rate at which they seek offers from competing firms. Since firms are risk neutral and workers are risk averse, the optimal labor contract is a fixed-wage contract, with a wage $w$ that is an endogenous function $W(x_{it})$ of the worker’s productivity at the time $t$ at which she is hired.\footnote{See, for instance, Harris and Holmstrom (1982) or Berk, Stanton, and Zechner (2010) for a discussion on the optimal labor contract in settings with risk-averse workers and a risk-neutral firm.} Workers stay in their job, earning their fixed wage, but might quit and move to another firm if and when they receive an outside offer. When a job offer is received at time $\tau$, the worker compares the proposed wage $W(x_{i\tau})$ to her current wage $w$ and decides to accept the offer if the lifetime utility $V(x_{i\tau}, W(x_{i\tau})) - \psi$ from moving, with $\psi$ being a switching cost.

Firms are risk neutral and competitive, with discount rate $r$. They value a worker with productivity $x$, wage $w$, and job-hunting rate $\chi$ according to

$$\Pi(x, w; \chi) = \mathbb{E}_x \left[ \int_0^\tau e^{-rt} (x_t - w) \, dt \right], \quad (22)$$

Workers stay in their job, earning their fixed wage, but might quit and move to another firm if and when they receive an outside offer. When a job offer is received at time $\tau$, the worker compares the proposed wage $W(x_{i\tau})$ to her current wage $w$ and decides to accept the offer if the lifetime utility $V(x_{i\tau}, w)$ from staying in the current job is below the lifetime utility $V(x_{i\tau}, W(x_{i\tau})) - \psi$ from moving, with $\psi$ being a switching cost.

Firms are risk neutral and competitive, with discount rate $r$. They value a worker with productivity $x$, wage $w$, and job-hunting rate $\chi$ according to

$$\Pi(x, w; \chi) = \mathbb{E}_x \left[ \int_0^\tau e^{-rt} (x_t - w) \, dt \right], \quad (22)$$
where $\tau$ is the quit time of the type-$\chi$ worker. With idiosyncratic productivity shocks only, there exists a well-defined stationary density $f_\infty(x, w, \chi)$ over workers’ productivity $x$, wage rate $w$ and type $\chi$,$^{45}$ and a corresponding stationary conditional type distribution of job transitioners $G(\chi|x)$. When pursuing a prospective employee, the firm acts competitively and offers a wage $W(x)$ that satisfies

$$E^G [\Pi(x, W(x); \chi)] = 0. \quad (23)$$

This break-even condition is the counterpart to (16) in the context of the mortgage market, and it pins down the equilibrium wage rate $W$. The expectation in (23) encodes firms’ inability to discriminate based on workers’ type $\chi$—due to either information asymmetry or discrimination laws covering protected class statuses that might be correlated with $\chi$.\(^{46}\) One can define a pooling MPE of this environment, in which (i) workers optimally switch firms subject to their search and job-hunting frictions, (ii) firms’ profits satisfy (22), and (iii) the equilibrium wage rate satisfies (23).

This environment allows us to study equilibrium wages; it allows us to analyze the impact of workers’ cross-sectional heterogeneity in job hunting rates on equilibrium wages and on the implicit cross-subsidies that aggressive at-work job hunters receive from loyal workers via the labor market. Rich micro data on job-to-job transitions and wages within industries can then be leveraged to discipline the model and discuss the quantitative implications of this cross-sectional heterogeneity, as well as policy counterfactuals. Qualitatively, this model would also feature an equilibrium cross-subsidy from more loyal workers to less loyal workers.

### 8.2 Other applications

While we discuss in some detail the labor market application of the framework developed in this article, we also emphasize that other environments could lend themselves to such an analysis. Subscription-based businesses, for instance, can be mapped into our model: cable or cellular phone services firms often offer promotional pricing at an

$^{45}$Our statement assumes that the equilibrium wage rate $W$ is monotonically increasing in productivity; this property can be verified ex post, when our postulated equilibrium has been computed numerically.

$^{46}$For example, marital status might predict mobility, but it is illegal to condition wages on marital status.
initially low rate for a short period of time before the rate is then increased. Various frictions (whether they are due to inattention, information gathering or pecuniary costs) lead to sluggish switching decisions by customers, and those with lower attention rates end up cross-subsidizing those who are more aggressive at switching products at the end of the promotional period. While a given economic application might come with specific assumptions and modeling devices that are unique to that environment, our approach’s tractability and ability to support systematic analysis of counterfactuals make it an attractive framework. We leave the precise evaluation and analysis of these different settings for future research.

9 Conclusion

In this paper, we have studied the equilibrium consequences of pooling ex ante heterogeneous agents facing various frictions and making dynamic discrete choices about entering into or renewing a long-term contract with a competitive sector that cannot price-discriminate based on type. We have applied our theoretical framework to the US conforming mortgage market—an ideal laboratory in which mortgage lenders, for various institutional reasons, end up offering mortgages without type-specific pricing, creating cross-subsidies from slow borrowers to fast ones. Our micro data suggests a large degree of cross-sectional heterogeneity in households’ attention rates, leading us to estimate significant cross-subsidies. Given that our measure of attention is correlated with income, the resulting redistribution is regressive, potentially to a much larger extent than that implied by the uniform credit guarantee scheme for agency mortgages. As policy discussions are regularly taking place in connection with a potential exit of the GSEs from conservatorship and the future of US housing finance, our paper provides a framework for exploring alternative mortgage market designs, taking into account the equilibrium effects of such counterfactuals.
References


Fuster, Andreas, David O Lucca, and James I Vickery. 2022. “Mortgage-Backed Securities.”


Appendix

1 Household’s refinancing behavior

1.1 Value function $V$

Proof of Proposition 1. First, the household decision problem can be recast as follows:

$$V(x, c) := \inf_{k \in K} \mathbb{E}_{x,c} \left[ \int_0^{+\infty} e^{-\rho t} \left( c^{(k)}_t dt + \psi dN^{(k)}_t \right) \right],$$

s.t. $$dc^{(k)}_t = \left( m(x_t) - c^{(k)}_t \right) \left( dN^{(k)}_t + dN^{(\nu)}_t \right),$$

where $K$ is a set of progressively measurable intensity processes $k = \{k_t\}_{t \geq 0}$ such that $k_t \in [0, \chi]$ at all times. Using this definition, we first show that $V$ must be increasing in $c$. Take $c' > c$ and an arbitrary intensity policy $k \in K$. The difference in payoffs for such intensity policy $k$ is:

$$\mathbb{E}_{x,c'} \left[ \int_0^{+\infty} e^{-\rho t} \left( c^{(k)}_t dt + \psi dN^{(k)}_t \right) \right] - \mathbb{E}_{x,c} \left[ \int_0^{+\infty} e^{-\rho t} \left( c^{(k)}_t dt + \psi dN^{(k)}_t \right) dt \right] 
\geq \mathbb{E}_x \left[ \int_0^\tau e^{-\rho t} (c' - c) dt \right] > 0,$$

where $\tau > 0$ a.s. is the first refinancing time under policy $k$. Taking the infimum over all admissible policies yields

$$V(x, c') = \inf_{k \in K} \mathbb{E}_{x,c'} \left[ \int_0^{+\infty} e^{-\rho t} \left( c^{(k)}_t dt + \psi dN^{(k)}_t \right) dt \right] 
\geq \inf_{k \in K} \mathbb{E}_{x,c} \left[ \int_0^{+\infty} e^{-\rho t} \left( c^{(k)}_t dt + \psi dN^{(k)}_t \right) dt \right] = V(x, c)$$

Thus $V$ is increasing in $c$. Moreover, since $\chi < +\infty$, a reasoning by contradiction can show that $V$ is in fact strictly increasing in $c$. Problem (1.1) is a standard stochastic control problem, for which standard results apply. For instance, for one-dimensional diffusions, and subject to some technical conditions on the operator $\mathcal{L}$, Strulovici and Szydlowski (2015)$^1$ provide for the value function $V$ being twice continuously differentiable in $x$, and satisfying the following HJB equation:

$$(\rho + \nu) V(x, c) = c + \mathcal{L}V(x, c) + \nu V(x, m(x)) + \min_{k \in [0,\chi]} \{ k (V(x, m(x)) + \psi - V(x, c)) \}$$

$^1$See also Fleming and Soner (2006); this latter article is not limited to one-dimensional diffusions, but includes additional – and more restrictive – conditions on the operator $\mathcal{L}$. 


The optimal Markov control is \( k^*(x, c) = \chi \mathbb{1}_{V(x, m(x)) + \psi \leq V(x, c)} \). Since \( V \) is strictly increasing in \( c \), this optimal policy can be re-written \( k^*(x, c) = \chi \mathbb{1}_{c - m(x) \geq \theta(x)} \), for a rate gap cutoff \( \theta(x) \) that satisfies

\[
V(x, m(x) + \theta(x)) = V(x, m(x)) + \psi
\]

\( \theta(x) \) is well defined since \( V \) is continuous and strictly increasing in \( c \). Reinjecting the optimal Markov control into the HJB equation satisfied by \( V \) yields the HJB equation in the main text.

1.2 Special case: \( m_t \) as a Brownian motion

Proof of Proposition 2. Assume that \( m_t = \sigma B_t + m_0 \), with \( B_t \) a Brownian motion. The household solves

\[
V(m, c) = \inf_{k \in \mathcal{K}} \mathbb{E}_{m, c} \left[ \int_0^{+\infty} e^{-\rho t} \left[ c_t^{(k)} dt + \psi dN_t^{(k)} \right] \right]
\]

where \( \mathcal{K} \) is defined in Section 1.1. Notice that the rate gap \( z_t^{(k)} := c_t^{(k)} - m_t \) evolves according to

\[
dz_t^{(k)} = -\sigma dB_t - z_t^{(k)} \left( dN_t^{(k)} + dN_t^{(\nu)} \right).
\]

\( V \) can be simplified as follows

\[
V(m, c) = \frac{c}{\rho} + \inf_{k \in \mathcal{K}} \mathbb{E}_{m, c} \left[ \int_0^{+\infty} e^{-\rho t} \left[ \left( c_t^{(k)} - c \right) dt + \psi dN_t^{(k)} \right] \right].
\]

In other words, \( V(m, c) = \frac{c}{\rho} + v(z) \), where

\[
v(z) = \inf_{k \in \mathcal{K}} \mathbb{E}_z \left[ \int_0^{+\infty} e^{-\rho t} \left[ \left( \psi - \frac{z_t^{(k)}}{\rho} \right) dN_t^{(k)} - \frac{z_t^{(k)}}{\rho} dN_t^{(\nu)} \right] \right]
\]

\[
dz_t^{(k)} = -\sigma dB_t - z_t^{(k)} \left( dN_t^{(k)} + dN_t^{(\nu)} \right).
\]

The value function \( v \) satisfies the following HJB:

\[
(\rho + \nu + \chi)v(z) = \frac{\sigma^2}{2} v''(z) + \chi \min \left( v(z), v(0) + \psi - \frac{z}{\rho} \right) + \nu \left( v(0) - \frac{z}{\rho} \right)
\]
Noting \( \theta \) the rate gap above which the household finds it optimal to refinance when given the opportunity to do so, we must have

\[
(\rho + \nu) v(z) = \frac{\sigma^2}{2} v''(z) + \nu \left( v(0) - \frac{z}{\rho} \right), \quad z \leq \theta
\]

\[
(\rho + \nu + \chi) v(z) = \frac{\sigma^2}{2} v''(z) + \nu \left( v(0) - \frac{z}{\rho} \right) + \chi \left( v(0) + \psi - \frac{z}{\rho} \right), \quad z \geq \theta
\]

Introduce the constant \( \eta_\chi := \sqrt{\frac{2(\rho + \nu + \chi)}{\sigma}} \). Note that \( v(z) = O(z) \) as \( z \to +\infty \) or as \( z \to -\infty \). Thus we must have

\[
v(z) = k_- e^{\eta_\theta(z-\theta)} + \frac{\nu \left( v(0) - \frac{z}{\rho} \right)}{\rho + \nu}, \quad z \leq \theta \tag{1.2}
\]

\[
v(z) = k_+ e^{-\eta_\chi(z-\theta)} + \frac{\nu \left( v(0) - \frac{z}{\rho} \right) + \chi \left( v(0) + \psi - \frac{z}{\rho} \right)}{\rho + \nu + \chi}, \quad z \geq \theta \tag{1.3}
\]

The constants \( k_-, k_+ \) must be such that \( v \) is continuously differentiable at \( z = \theta \). Moreover, since we must have \( \theta > 0 \), it must be the case that

\[
v(0) = \frac{\rho + \nu}{\rho} k_- e^{-\eta_\theta \theta}
\]

Taking \( \theta \) as given, the requirement that \( v \) be continuously differentiable at \( z = \theta \) yields a system of 2 equations in the 2 unknown \( k_-, k_+ \):

\[
k_- + \frac{\nu}{\rho + \nu} \left[ v(0) - \frac{\theta}{\rho} \right] = k_+ + \frac{\nu}{\rho + \nu + \chi} \left[ v(0) - \frac{\theta}{\rho} \right] + \frac{\chi}{\rho + \nu + \chi} \left[ v(0) + \psi - \frac{\theta}{\rho} \right]
\]

\[
\eta_0 k_- = -\eta k_+ - \frac{\nu + \chi}{\rho (\rho + \nu + \chi)}
\]

Using our formula for \( v(0) \), this system can be solved to yield:

\[
k_- = \frac{\chi}{(\rho + \nu + \chi) (\eta_0 + \eta_\chi) - \chi \eta_\chi e^{-\eta_0 \theta}} \left[ \eta_\chi \left( \psi - \frac{\theta}{\rho + \nu} \right) - \frac{1}{\rho + \nu} \right]
\]

\[
k_+ = \frac{\chi}{(\rho + \nu + \chi) (\eta_0 + \eta_\chi) - \chi \eta_\chi e^{-\eta_0 \theta}} \left[ \left( 1 - \frac{\chi e^{-\eta_0 \theta}}{\rho + \nu + \chi} \right) \left( \frac{1}{\rho + \nu} \right) + \eta_0 \left( \psi - \frac{\theta}{\rho + \nu} \right) \right]
\]

Finally, it must be the case that the borrower is exactly indifferent between (a) continuing with her current mortgage, or (b) paying the fixed cost and refinancing,
when \( z = \theta \). This necessarily means that
\[
v(\theta) = v(0) + \psi - \frac{\theta}{\rho} \Rightarrow k_- = \frac{\rho}{\rho + \nu} v(0) + \psi - \frac{\theta}{\rho + \nu}
\]

But since we know \( v(0) \) as a function of \( k_- \), this yields
\[
k_- = k_- e^{-\eta_0 \theta} + \psi - \frac{\theta}{\rho + \nu}
\]

Using our formula for \( k_- \), this yields, after some algebra, the implicit equation
\[
e^{-\eta_0 \theta} + \left[ \eta_0 + \frac{\rho + \nu}{\chi} (\eta_0 + \eta_\chi) \right] \theta = 1 + \left[ \eta_0 + \frac{\rho + \nu}{\chi} (\eta_0 + \eta_\chi) \right] (\rho + \nu) \psi
\]

Note \( F(\theta) \) the left-hand-side of the above equation. Notice that \( F \) is convex, and \( F'(0) > 0 \), which means that \( F \) is strictly increasing for \( \theta > 0 \). Moreover, \( F(0) = 1 < 1 + \left[ \eta_0 + \frac{\rho + \nu}{\chi} (\eta_0 + \eta_\chi) \right] (\rho + \nu) \psi \), and \( F(\theta) \to +\infty \) when \( \theta \to +\infty \). In other words, the above equation admits a unique positive solution \( \theta \). Re-write the implicit equation satisfied by \( \theta \) as follows:
\[
e^{-\eta_0 \theta} + (\eta_0 + \epsilon_\chi) \theta = 1 + (\rho + \nu) \psi (\eta_0 + \epsilon_\chi)
\]

\( \epsilon_\chi := \frac{(\rho + \nu)(\eta_0 + \eta_\chi)}{\chi} \)

Clearly, \( \epsilon_\chi \) is a positive and decreasing function of \( \chi \), converging to zero as \( \chi \to +\infty \). Differentiate the above equation w.r.t. \( \chi \) to obtain
\[
\frac{\partial \theta}{\partial \chi} = \frac{((\rho + \nu) \psi - \theta) \frac{\partial \epsilon_\chi}{\partial \chi}}{\epsilon_\chi + \eta_0 (1 - e^{-\eta_0 \theta})} > 0,
\]

with the last inequality following from \( \frac{\partial \epsilon_\chi}{\partial \chi} < 0 \) and \( (\rho + \nu) \psi - \theta = \frac{e^{-\eta_0 \theta} - 1}{\eta_0 + \epsilon_\chi} < 0 \). Thus the barrier \( \theta \) increases with \( \chi \). Finally, consider the asymptotic behavior of \( \theta \). When \( \chi \to +\infty \), the limiting value \( \theta_\infty \) solves
\[
e^{-\eta_0 \theta_\infty} + \eta_0 \theta_\infty = 1 + \eta_0 \psi (\rho + \nu)
\]

This delivers the ADL formula, using the Lambert function \( W \):
\[
\theta_\infty = \frac{1}{\eta_0} (1 + \eta_0 \psi (\rho + \nu) + W (-\exp (-1 - \eta_0 \psi (\rho + \nu)))))
\]
Instead, consider the case $\chi \to 0$. In that case, the optimal threshold converges to

$$\theta_0 = (\rho + \nu)\psi$$

We can also perform a Taylor expansion of (1.4) around $\theta = 0$, which allows us to obtain an approximation $\hat{\theta}$ of the value $\theta$:

$$\frac{\eta_0^2}{2} \hat{\theta}^2 + \epsilon \chi \hat{\theta} - (\rho + \nu)(\eta_0 + \epsilon \chi)\psi = 0$$

This allows us to conclude that the approximation $\hat{\theta}$ is equal to

$$\hat{\theta} = \sqrt{\frac{2}{\eta_0} \left( 1 + \frac{\epsilon \chi}{\eta_0} \right) (\rho + \nu) \psi + \left( \frac{\epsilon \chi}{\eta_0^2} \right)^2 - \frac{\epsilon \chi}{\eta_0}}$$

Finally, it is straightforward (but tedious) to verify that the optimal rate gap threshold $\theta$ is identical to that derived above if one were to assume a fixed cost upon moving. The intuition behind this result is straightforward: since there is equal probability that the household moves when the mortgage is in- or out-of-the-money (given the fact that the mortgage rate is a Brownian motion), the household’s refinancing strategy does not change in the presence of fixed moving costs.

### 1.3 Optimal threshold $\theta(x)$ vs. speed of mean-reversion

We focus on the partial equilibrium of our model, and assume that the mortgage rate follows an Ornstein-Uhlenbeck process:

$$dm_t = -\kappa(m_t - \bar{m})dt + \sigma dB_t$$

This parametrization for the mortgage rate process nests the special case of the Brownian motion studied in the main text, by setting $\kappa = 0$. We solve the stochastic control problem numerically with a finite difference scheme, and plot in Figure 1.1 the ergodic average threshold $E[\theta(x_t)]$ and the ergodic average slope of the threshold $E[\theta'(x_t)]$ as a function of the speed of mean-reversion $\kappa$. Our parameter choice includes (i) a subjective discount rate $\rho = 5\%$, (ii) a mortgage rate volatility $\sigma = 1\%$ (both (i) and (ii) being consistent with ADL), (iii) a moving plus amortization rate $\nu = 7.4\%$ (consistent with the 4.1\% yearly moving rate of Berger, Milbradt, Tourre, and Vavra (2021) plus a yearly mortgage amortization rate of 3.3\%), (iv) upfront closing costs $\psi = 2\%$ (consistent with ADL for their base case calibration under the assumption that the household has a mortgage balance of around $200,000$), and (v) an ergodic average mortgage rate $\bar{m} = 5\%$, consistent with the time-series average of mortgage rates from beginning 2000 until end 2021.
Figure 1.1: Rate gap threshold $\theta$ vs. mortgage process half-life. Rate gap threshold $\theta$ in the case of $m_t$ following an OU process for various speeds of mean-reversion $\kappa$. Left plot shows the ergodic average threshold $E[\theta(x_t)]$, while right plot shows the ergodic average slope of the threshold $E[\theta'(x_t)]$. Horizontal dash green line represents the limit $\kappa \to 0$—i.e. when $m_t$ is a Brownian motion. Figure computed for $\rho = 5\%$, $\chi = 19\%$, $\nu = 7.4\%$, $\psi = 2\%$, $\bar{m} = E[m_t] = 5\%$ and $\sigma = 1\%$.

Figure 1.1 suggests that the average rate gap threshold $E[\theta(x)]$ is mildly dependent on the half-life of the mortgage rate process – for our chosen parameters, it ranges from 33bps to 36bps. Most importantly though, moving from the pure random walk assumption to a mean-reverting process introduces a non-negligible amount of state dependence in the barrier $\theta(x)$.

2 Mortgage market equilibrium

2.1 MPE existence and uniqueness in homogeneous case

Proof of Proposition 3. Discounted debt prices must be martingales, thus

$$ r(x)P(x, c; \chi) = c - f + LP(x, c; \chi) + (\nu + \chi 1_{c - m(x) \geq \theta(x)}) (1 - P(x, c; \chi)) . \quad (2.1) $$

The function $P$, solution of (2.1), is implicitly dependent on a mortgage rate function $m(x)$, via the decision rule $\theta(x)$, which comes out of the household refinancing problem. It thus means that the equilibrium mortgage rate, implicitly defined via $P(x, m(x); \chi) = 1 + \pi$, is the outcome of a potentially complex fixed-point problem.
Our proof has two steps; we first tackle the case \( \pi = 0 \), and then generalize to the case \( \pi > 0 \). In both cases, we assume no upfront closing costs (i.e. \( \psi = 0 \)), and we assume that \( r_t \in [\underline{r}, \bar{r}] \), with \( 0 \leq \underline{r} < \bar{r} < +\infty \), and \( \chi < +\infty \). In that environment without upfront closing costs paid by households, the decision rule simplifies to \( \theta(x) = 0 \), in other words the optimal intensity solves \( k^*(x, c) = \chi 1_{\{c \geq m(x)\}} \).

i. In this section, we restrict ourselves to the case where \( \pi = 0 \). To make further progress, we study the auxiliary problem

\[
\tilde{P}(x, c; \chi) := \inf_{k \in \mathcal{K}} \mathbb{E}_x \left[ \int_0^{+\infty} e^{-\int_0^t (r(x_s) + k_s + \nu) ds} \left( c - f + k_t + \nu \right) dt \right],
\]

where \( \mathcal{K} \) is defined in Section 1.1. The function \( \tilde{P} \) does not depend, directly or indirectly, on any equilibrium object; in other words, one can view \( \tilde{P} \) as the solution to a single-agent stochastic control problem. Arguments similar to those developed in Section 1.1 allow us to argue that \( \tilde{P} \) is twice continuously differentiable in \( x \), continuous and increasing in \( c \), satisfying the HJB equation

\[
(r(x) + \nu) \tilde{P}(x, c; \chi) = c - f + \nu + \mathcal{L} \tilde{P}(x, c; \chi) + \min_{k \in [0, \chi]} \left\{ k \left( 1 - \tilde{P}(x, c; \chi) \right) \right\}.
\]

The optimal Markov control is \( \tilde{k}(x, c) = \chi 1_{\{\tilde{P}(x, c; \chi) \geq 1\}} \). Since \( r_t \) is restricted to be on \( \mathbb{R}_+ \), we must have \( \tilde{P}(x, 0; \chi) < 1 \). Similarly, since \( r_t \) is bounded above by \( \bar{r} \), for \( c \) sufficiently high we must have \( \tilde{P}(x, c; \chi) > 1 \). Since \( \tilde{P} \) is continuous and increasing in \( c \), by the intermediate value theorem there must exist a unique real value \( c = m(x) \) that satisfies

\[
\tilde{P}(x, m(x); \chi) = 1
\]

Given this construction, and given that \( \tilde{P} \) is monotone in \( c \), the set of events \( \{\tilde{P}(x_t, c; \chi) \geq 1\} \) is identical to the set of events \( \{m(x_t) \leq c\} \). We can then verify that the auxiliary function \( \tilde{P} \) is none other than the pricing function \( P \), and the mortgage rate function \( m(x) \) defined via (2.3) is unique and satisfies the equilibrium condition \( P(x, m(x); \chi) = 1 \).

ii. We now consider the case \( \pi > 0 \) – i.e. the case where mortgage origination triggers costs, borne by lenders and recouped via higher mortgage rates. In this section, we also assume that the latent state \( x \) is one-dimensional and \( r(\cdot) \) is increasing. We will prove that there exists a unique monotone equilibrium in that case – i.e. a unique MPE in which the mortgage rate function is monotone increasing in \( x \). Take an arbitrary \( x^* \), and define \( \tau_{x^*, \chi} \) as a stopping time with arrival intensity \( \nu + \chi 1_{\{x_t \leq x^*\}} \). As will be seen shortly, \( x^* \) represents the latent

\[2\text{Formally, if } \omega \text{ is a (unit mean) exponentially distributed random variable and if we introduce} \]
state that was prevalent the last time a household refinanced. Consider the interest-only “IO” and principal-only “PO” net present values, defined via

\[ IO(x; \chi) := \mathbb{E}_x \left[ \int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds} dt \right] \]

\[ PO(x; \chi) := \mathbb{E}_x \left[ e^{-\int_0^{\tau_{x,\chi}} r_s ds} \right] \]

These objects represents, respectively, the valuation of an IO and a PO whenever the latent state variable is \( x \), and whenever the prepayment time is driven by a point process with (time-varying) intensity \( \nu + \chi \mathbb{1}_{\{x \leq x^* \}} \). Introduce the function \( m \), defined via

\[ m(x) := f + \frac{1 - PO(x; \chi)}{IO(x; \chi)} + \frac{\pi}{IO(x; \chi)}. \]  \hspace{2cm} (2.4)

\( m \) is continuous in \( x \). We argue that \( m \) is a monotone increasing function of \( x \), and that a monotone equilibrium exists, in which \( m(x) \) is the equilibrium mortgage market interest rate. Consider first the special case \( \pi = 0 \). In that case, we know from the previous section (i) that an equilibrium exists and is unique. Since the objective in problem (2.2) is decreasing in \( x \), it must be the case that the function \( \tilde{P} \) defined in (2.2) is decreasing in \( x \), which must mean that the equilibrium mortgage rate, when \( \pi = 0 \), is monotone increasing in \( x \). In that case, the mortgage rate function must correspond to that defined in (2.4) (with \( \pi = 0 \)—this is the case since

\[ \tilde{P}(x, m(x); \chi) = 1 = \mathbb{E}_x \left[ \int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x) - f) dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds} \right] \]

\[ = (m(x) - f)IO(x; \chi) + PO(x; \chi), \]

which directly implies (2.4) for \( \pi = 0 \). As \( m(x) \) is increasing when \( \pi = 0 \), it must be the case that \((1 - PO(x; \chi))/IO(x; \chi)\) is increasing in \( x \). For \( \pi > 0 \), we additionally need to show that \( 1/IO(x; \chi) \) is increasing in \( x \). To this end, note that for \( x_1 < x_2 \), we must always have

\[ \mathbb{E}_{x_2} \left[ \int_0^{\tau_{x_2,\chi}} e^{-\int_0^t r_s ds} dt \right] \leq \mathbb{E}_{x_1} \left[ \int_0^{\tau_{x_2,\chi}} e^{-\int_0^t r_s ds} dt \right] \leq \mathbb{E}_{x_1} \left[ \int_0^{\tau_{x_1,\chi}} e^{-\int_0^t r_s ds} dt \right]. \]

The first inequality above stems from the fact that if the initial interest rate is \( r(x_1) \), the full time path of future interest rates is below that which would be relevant if the initial interest rate was \( r(x_2) \). The second inequality stems from the fact that, for a given starting level of the latent state \( x_1 \), we must have the stopping time inequality \( \tau_{x_2,\chi} \leq \tau_{x_1,\chi} \) almost surely. In other words, \( IO(x; \chi) \) must be decreasing in \( x \). This allows us to conclude that \( m \), defined in (2.4), is

---

The compensator \( \Lambda_t = \int_0^t (\nu + \chi \mathbb{1}_{\{x_s \leq x^* \}}) ds \), then the stopping time \( \tau_{x^*,\chi} \) is the (random) time that satisfies \( \Lambda_{\tau_{x^*,\chi}} = \omega \).
monotone increasing in $x$. Given this observation, we must have an equilibrium in which $m$ is the mortgage rate, since $m$ must satisfy

$$1 + \pi = \mathbb{E}_x \left[ \int_0^{\tau_{x,X}} e^{-\int_0^t r_s ds} (m(x) - f) dt + e^{-\int_0^{\tau_{x,X}} r_s ds} \right]$$

That equilibrium is unique, since we showed its existence by construction. In other words, in any monotone equilibrium, it must be the case that the mortgage rate function satisfies (2.4).

\[\square\]

**2.2 Comparative statics**

*Proof of Proposition 4.* Consider first the case $\pi = 0$. Since $P = \tilde{P}$ can be defined via equation (2.2), it must be the case that $P$ is decreasing in $\chi$. Thus, the mortgage rate function $m(x)$ is an increasing function of $\chi$, whenever $\pi = 0$. Consider then the case where $\pi > 0$, and where the latent state $x$ is one-dimensional and $r(\cdot)$ is increasing. Given our conclusion for the case $\pi = 0$, it must be the case that $(1 - PO(x; \chi))/IO(x; \chi)$ is increasing in $\chi$. Define (with a slight abuse of notation)

$$IO(x, x^*; \chi) := \mathbb{E}_x \left[ \int_0^{\tau_{x,X}} e^{-\int_0^t r_s ds} \right],$$

which solves the PDE

$$\left( r(x) + \nu + \chi 1_{\{x \leq x^*\}} \right) IO(x, x^*; \chi) = 1 + L IO(x, x^*; \chi)$$

Differentiate this equation w.r.t. $\chi$ to obtain

$$\left( r(x) + \nu + \chi 1_{\{x \leq x^*\}} \right) \partial_\chi IO(x, x^*; \chi) = -1_{\{x \leq x^*\}} IO(x, x^*; \chi) + L \partial_\chi IO(x, x^*; \chi)$$

Thus, $\partial_\chi IO(x, x^*; \chi)$ admits the integral representation

$$\partial_\chi IO(x, x^*; \chi) = -\mathbb{E}_x \left[ \int_0^{\tau_{x,X}} e^{-\int_0^t r_s ds} 1_{\{x \leq x^*\}} IO(x_t, x^*; \chi) dt \right] < 0$$

Thus, $IO(x; \chi)$ is monotone decreasing in $\chi$. This must mean that the mortgage rate function, defined via (2.4), is increasing in $\chi$, whenever $\pi > 0$.

\[\square\]

**2.3 Infinite dimensional problem with heterogeneity**

In this section, we discuss the key mathematical equations characterizing the Pooling MPE. As a reminder, $H(\chi)$ denotes the cumulative distribution over types (with associated density $h$), while $F_t$ denotes the joint cumulative distribution over outstanding coupon rates $c$ and types $\chi$ in the population at time $t$ (with associated joint density
Since types are a permanent household attribute, we must have
\[
\int_c f_t(c, \chi) dc = h(\chi). \tag{2.5}
\]
Consider then the density \( f_t \). It evolves endogenously over time with idiosyncratic mortgage refinancing decisions, which, aggregated using a weak law of large numbers, lead to locally deterministic movements in \( f_t \). The Kolmogorov Forward Equation ("KFE") that describes these changes is then, for \( c \neq m(S) \):
\[
df_t(c, \chi) = -\nu + \chi 1_{\{c \geq m(S_t)\}} f_t(c, \chi) dt, \quad c \neq m(S). \tag{2.6}
\]
The density \( f_t \), between \( t \) and \( t + dt \), looses mass at rate \( \nu \) for \( c < m(S_t) \), and at the higher rate rate \( \nu + \chi \) for \( c \geq m(S_t) \), as households strategically refinance. This equation holds everywhere except at \( c = m(S_t) \), a state at which refinancing and moving households are being "reinjected"; the relevant equation in that case is
\[
\lim_{c \uparrow m(S_t)} \partial_c f_t(c, \chi) = \lim_{c \downarrow m(S_t)} \partial_c f_t(c, \chi) = \nu h(\chi) + \chi Z \int_t^{t + \infty} f_t(c, \chi) dc. \tag{2.7}
\]
The right-hand-side of this equation is the flux of households exogenously moving at rate \( \nu \) and the flux of type-\( \chi \) households refinancing in the time interval \([t, t + dt]\), while the left-hand-side is the kink in the density at \( c = m(S) \) induced by the reinjection of such households at that particular point of the state space.

Let \( P(S, c; \chi) \) be the shadow price of a mortgage with coupon \( c \), conditional on knowing that the related household has attention rate \( \chi \). The shadow price solves the following infinite dimensional Feynman-Kac equation, which takes into account (i) changes in the distribution \( f_t \), and (ii) the behavior of type-\( \chi \) households:
\[
\frac{d}{dt} P(S, c; \chi) = -\nu + \chi 1_{\{c \geq m(S_t)\}} + \left( \nu + \chi 1_{\{c \geq m(S_t)\}} \right) \left[ 1 - P(S, c; \chi) \right]
+ \int T[f](c, \chi) \frac{\delta P}{\delta f(c, \chi)} dc d\chi, \tag{2.8}
\]
with \( \delta P/\delta f \) the functional derivative of \( P \) w.r.t. \( f \) at \( (c, \chi) \) and the operator \( T \) defined via
\[
T[f](c, \chi) = -\left( \nu + \chi 1_{\{c > m(S_t)\}} \right) f(c, \chi). \tag{2.9}
\]
See Achdou, Buera, Lasry, Lions, and Moll (2014) for another example of such infinite-dimensional PDE in the context of consumption-savings models in incomplete markets with aggregate shocks.
2.4 Approximate Pooling MPE existence and uniqueness

Proof of Proposition 5. We establish the existence and uniqueness of the Approximate Pooling MPE using a method similar to Section 2.1 for the case \( \pi > 0 \). To that effect, consider the dynamic system \((x_t, x^*_t, \chi_t)\), where

\[
\begin{align*}
\frac{dx_t}{dt} &= \mu(x_t)dt + \sigma(x_t)dB_t \\
\frac{dx^*_t}{dt} &= (x_t - x^*_{t-\chi})dN^\chi_t,
\end{align*}
\]

where \( N^\chi_t \) is a point process with arrival intensity \( \nu + \chi \mathbb{1}_{\{x_t \leq x^*_{t-\chi}\}} \). This dynamic system admits a generator \( \mathcal{H}_\chi \) defined for any smooth function \( \phi(x, x^*) \) via

\[
\mathcal{H}_\chi \phi(x, x^*) = \mathcal{L}_\chi \phi(x, x^*) + (\nu + \mathbb{1}_{\{x \leq x^*\}})(\phi(x, x) - \phi(x, x^*))
\]

The eigen-function (associated with the eigen-value zero) of the adjoint of the operator \( \mathcal{H}_\chi \) gives us the stationary density \( f_\infty(x, x^*|\chi) \) of the dynamic system \((x_t, x^*_t, \chi_t)\). Introduce then the distribution \( g(\chi) \), either the unconditional one defined via

\[
g(\chi) = \frac{h(\chi) \int_x \left( \nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi) \right) f_\infty(x)dx}{\int_x h(\chi) \int_x \left( \nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi) \right) f_\infty(x)dx d\chi},
\]

or the conditional one defined via

\[
g(\chi|x) = \frac{h(\chi) \left( \nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi) \right) dx}{\int_x h(\chi) \left( \nu + \chi \int_{x^* \geq x} f_\infty(x^*|x, \chi) \right) d\chi}.
\]

Define the candidate mortgage rate \( m(x; G) \) via

\[
m(x; G) := f + \frac{1 + \pi - \mathbb{E}^G[PO(x; \chi)]}{\mathbb{E}^G[IO(x; \chi)]},
\]

If the function \( m(x; G) \) is increasing in \( x \), a monotone Approximate Pooling MPE must exist, and this equilibrium is unique amongst all monotone equilibria. Note that \( m \) must satisfy

\[
1 + \pi = \mathbb{E}^G \left[ \mathbb{E}_x \left[ \int_0^{\tau_{x, \chi}} e^{-\int_0^t r_s ds} (m(x) - f)dt + e^{-\int_0^{\tau_{x, \chi}} r_s ds} \right] \right]
\]

Consider then the price \( \hat{P}_G(x, m(x^*)) \) of a mortgage with coupon \( m(x^*) \),

\[
\hat{P}_G(x, m(x^*)) := \mathbb{E}^G \left[ \mathbb{E}_x \left[ \int_0^{\tau_{x, \chi}} e^{-\int_0^t r_s ds} (m(x^*) - f)dt + e^{-\int_0^{\tau_{x, \chi}} r_s ds} \right] \right],
\]
then clearly if \( m \) is increasing, \( \bar{P}_G \) must be increasing in \( x^* \), with \( \bar{P}_G (x^*, m(x^*)) = 1 + \pi \) – in other words the equilibrium conditions are satisfied.

2.5 Integral representation of \( \bar{P}_G \) for unconditional \( G(\chi) \)

*Proof of Proposition 6.* The HJB equation (2.1) holds for all \( \chi \), and thus, taking expectations w.r.t. the unconditional issuance type distribution \( G(\chi) \), we have

\[
\begin{align*}
 r(x) \bar{P}_G (x, c) &= c - f - \mathbb{1}_{\{m(x) \leq c\}} \text{Cov}^G (\chi, P(x, c; \chi)) + \mathcal{L}\bar{P}_G (x, c) \\
 &\quad + (\nu + \bar{\chi}G \mathbb{1}_{\{m(x) \leq c\}}) \left( 1 - \bar{P}_G (x, c) \right) 
\end{align*}
\] (2.13)

One can then use Feynman-Kac to conclude that \( \bar{P}_G \) admits the integral representation in Proposition 6.

2.6 Invariance of lowest attainable mortgage rate

*Proof of Proposition 7.* Under the assumption that \( x \) is uni-dimensional and that \( r(\cdot) \) is monotone increasing, call \( \bar{x} \) the lowest bound for \( x \). The monotone Approximate Pooling MPE implies \( m \) is increasing in \( x \), and thus \( m(\bar{x}) \) must be the lowest attainable mortgage rate. Then we have

\[
P(x, m(\bar{x}); \chi) = P(x, m(\bar{x}); \chi'), \quad \forall \chi, \chi',
\]

as \( \chi \) only influences the refinancing channel, and whenever households have locked in the lowest possible rate \( c = m(\bar{x}) \), we have \( c \leq m(x_t) \) at all future date \( t \) regardless of type \( \chi \). Thus, from the break-even condition \( \bar{P}(x, m(x)) = 1 + \pi \), we have

\[
P(x, m(\bar{x}); \chi) = 1 + \pi, \quad \forall \chi.
\]

\( m(\bar{x}) \) is thus invariant to the distribution \( G \), and thus the distribution \( H \).

3 Policy evaluations and counterfactuals

3.1 Construction of the auto-RM

We first argue that the auto-RM market rate is a reference rate computed by looking at debt instruments traded in the market and prepayable at any time, with a call premium \( \pi \). Indeed, note \( P^*(x, c) \) the price of such a prepayable instrument with coupon \( c \) when the latent aggregate state is \( x \):

\[
P^*(x, c) := \inf \mathbb{E}_x \left[ \int_0^{\tau \wedge \tau_\nu} e^{-\int_0^s r(x_s) \, ds} \,(c - f) \, ds + \mathbb{1}_{\{\tau \leq \tau_\nu\}} (1 + \pi) e^{-\int_0^\tau r(x_s) \, ds} + \mathbb{1}_{\{\tau \geq \tau_\nu\}} e^{-\int_0^{\tau_\nu} r(x_s) \, ds} \right],
\]
where \( \tau_v \) is a Poisson time with arrival rate \( \nu \). This optimal stopping problem is a free-boundary problem, with an endogeneous boundary \( m^*(x) \) to be determined. The variational inequality, valid for any \( x \), is

\[
\max \{ - \left( \nu + r(x) \right) P^*(x, c) + c - f + \nu + \mathcal{L}P^*(x, c), P^*(x, c) - (1 + \pi) \} = 0
\]

The HJB (i.e. the left hand side of the above inequality) holds in the continuation region \( c \leq m^*(x) \), while the equality \( P^*(x, c) = 1 + \pi \) holds at the boundary of the continuation region \( c = m^*(x) \), where \( m^*(x) \) is the “auto-RM rate”. The optimality condition for the stopping time \( \tau \) takes the form of the smooth pasting condition

\[
\partial_x P^*(x, m^*(x)) = 0.
\]

The price function \( P^*(x, c) \) satisfies \( P^*(x, c) \leq 1 + \pi \) for any coupon \( c \) and latent state \( x \). \( P^* \) is increasing in \( c \), and of course \( P^*(x, m^*(x)) = 1 + \pi \).

Note then that \( P^* \) is the limit, as \( \chi \to +\infty \), of the following problem

\[
\hat{P}(x, c; \chi) := \inf_{k \in \mathcal{K}_\chi} \mathbb{E}_x \left[ \int_0^{\tau_k/\tau_v} e^{-\int_0^t r(s) ds} (c - f) ds + 1_{\{\tau_k \leq \tau_v\}} (1 + \pi) e^{-\int_0^{\tau_k} r(s) ds} + 1_{\{\tau_k \geq \tau_v\}} e^{-\int_0^{\tau_v} r(s) ds} \right]
\]

\[
= \inf_{k \in \mathcal{K}_\chi} \mathbb{E}_x \left[ \int_0^{\tau_v} e^{-\int_0^t (r(s) + \nu + k_t) ds} (c - f + \nu + k_t (1 + \pi)) ds \right],
\]

where \( \mathcal{K}_\chi \) is the set of progressively measurable processes \( \{k_t\}_{t \geq 0} \) so that \( k_t \in [0, \chi] \) for all \( t \), and \( \tau_k \), in the first equation, is a Poisson time with jump intensity \( k_t \). The auto-RM rate is thus a reference rate that can be computed by looking at debt instruments traded in the market, and that are prepayable at any time at \( 1 + \pi \). These prepayable debt instruments, when issued, have a price and market value of \( 1 + \pi \), and a fair coupon equal to \( m^*(x) \), the reference rate for the auto-RM.\(^3\) Households are then locked into that auto-RM instrument, pay the floating rate \( m^*(x_t) \) at all times, up to the point where they move. At such time, they prepay the mortgage balance \$1, and are forced to refinance into a new mortgage. Upon taking a new mortgage, households receive proceeds \$1 from lenders, but given that the loan pays a reference rate \( m^*(x) \), the market value of such loan is equal to \( 1 + \pi \), meaning that lenders can recoup their origination costs. Households then pay the floating rate \( m^*(x_t) \) until the time they move and sell their house. By construction, the reference rate \( m^*(x_t) \) satisfies

\[
m^*(x_t) = \inf_{t \geq s \geq 0} m^*(x_s)
\]

### 3.2 Auto-RM vs. short rates

**Proof of Proposition 8.** We consider the case \( \pi \geq 0 \) – i.e. the case where mortgage origination costs are potentially incurred, and recouped by lenders via higher mort-

\(^3\)Given the nature of Brownian motions, these prepayable instruments have, at the time of issuance, zero duration.
gage rates. As discussed in Appendix 3.1, the price $P^*$ of the auto-RM solves

$$P^*(x, c) = \inf_{\tau} \mathbb{E}_x \left[ \int_0^\tau e^{-\int_0^t (r(x_s)+\nu)ds} (c - f + \nu) ds + (1 + \pi) e^{-\int_0^\tau (r(x_s)+\nu)ds} \right].$$

Now assume for a second that there exists a latent state $\hat{x}$ so that $r(\hat{x}) > m(\hat{x}) - f$. Assume at time $t = 0$, $x_0 = \hat{x}$, and consider a stopping strategy $T = \inf\{t \geq 0 : r(x_t) = m(\hat{x}) - f\}$. Clearly, since $r(x_0) = r(\hat{x}) > m(\hat{x}) - f$ and since $x$ has continuous sample path, $T > 0$ a.s. We then have the following set of inequalities

$$1 + \pi = P^*(\hat{x}, m(\hat{x})) = \inf_{\tau} \mathbb{E}_{\hat{x}} \left[ \int_0^\tau e^{-\int_0^t (r(x_s)+\nu)ds} (m(\hat{x}) - f + \nu) dt + (1 + \pi) e^{-\int_0^\tau (r(x_s)+\nu)ds} \right]$$

$$\leq \mathbb{E}_{\hat{x}} \left[ \int_0^T e^{-\int_0^t (r(x_s)+\nu)ds} (m(\hat{x}) - f + \nu) dt + (1 + \pi) e^{-\int_0^T (r(x_s)+\nu)ds} \right]$$

$$< 1 + \pi,$$

where the last inequality follows since for $t < T$, we must have $r(x_t) > m(\hat{x}) - f$. This is the contradiction we were looking for.

3.3 Auto-RM impact on initial debt-to-income ratio
Figure 3.1: **DTI distribution and counterfactual DTI distribution.** Left figure shows the DTI distribution in the SFLP data. Right figure shows the counterfactual DTI distribution if mortgage rates were higher than those actually realized, with a difference corresponding to the ergodic average difference between (a) mortgage rates in the Approximate Pooling MPE and (b) mortgage rates in the auto-RM equilibrium.

4 Household attention in mortgage refinancing data

4.1 Data

We rely on information from Equifax Credit Risk Insight Servicing McDash (“CRISM”). This monthly-frequency data-set covers the period from May 2005 until December 2017. It contains unique borrower IDs, mortgage IDs, a prepayment indicator if a loan prepaid in a given month, the original coupon rate on the loan, its current principal balance, as well as the current FICO score of the related borrower. We build an indicator describing the type of prepayment (rate refinancing, cash-out refinancing or moves), and a measure of the current combined loan-to-value ratio (thereafter, “CTLV”) using house price data from Corelogic. We construct, for each household and each month, the effective mortgage market rate available to a household, by regressing observed contractual rates on characteristics. This allows us to construct the rate gap – i.e. the difference between (i) the mortgage coupon and (ii) the household effective mortgage market rate. Our data-set allows us to track a borrower and their different mortgages through time. It contains 20,094,230 loan-month-borrower observations, with 246,330 unique borrower IDs.

For some of our econometric work, we also leverage the single-family loan performance (“SFLP”) data-set from Fannie-Mae. While CRISM allows us to track
individual households across loans, SFLP only allows us to track monthly mortgage performance data for a sample of conforming loans originated between January 2000 and December 2021. This means that the SFLP data cannot distinguish refinancing from other types of prepayment. However, this data is nevertheless useful since it contains covariates which are absent in CRISM – for instance the identity of the original lender and of the mortgage servicer.

4.2 Estimating the attention distribution

4.2.1 Individual attention rates

Given \( \chi \) and \( dt \), the probability of refinancing in an “effective” period is \( p = 1 - \exp(-\chi dt) \). We estimate the probability per period by using an observation for each individual

\[
x_i = \frac{[\text{Nr of refi}]}{[\text{Nr of effective periods}]}
\]

Assuming that attention times are independent and identically distributed,

\[
\mathbb{E}[x_i] = \frac{\mathbb{E}[\text{Nr of refi}]}{[\text{Nr of effective periods}]} = \frac{p \cdot [\text{Nr of effective periods}]}{[\text{Nr of effective periods}]} = p
\]

Thus, using \( N_{\text{sim}} \) individuals observed over the panel from 0 to \( T \), we have \( N_{\text{sim}} x_i \) observations.

4.2.2 Clustering algorithm

For our clustering algorithm, we fix \( N \), the number of groups. Whether a household belongs to one group or another is determined via maximum likelihood. It is easier to state this optimization in terms of attention probability per month \( p \), rather than in terms of attention rate \( \chi \).

To each household \( i \), we associate a binomial random variable where the number of successes is “Nr of refi”, and the number of trials is “Nr of effective periods”. The logarithm of the probability mass function is then

\[
\log (pmf_i) = \log \left( \frac{[\text{Nr of effective periods}_i]}{[\text{Nr of refi}_i]} \right) + [\text{Nr of refi}_i] \log p_i
\]

\[
+ ([\text{Nr of effective periods}_i] - [\text{Nr of refi}_i]) \log (1 - p_i)
\]

Suppose that we have a sample of \( N_{HH} \) households, each with \([\text{Nr of effective periods}]_i\) and \([\text{Nr of refi}]_i\). Suppose further that we impose that there is a vector \( p \) with possible values in \( P \) of length \( N \leq N_{HH} \). We want to use MLE to estimate the optimal vector \( P \) (which implies an optimal vector \( p \)) for a given number \( N \) of groups and a given sample of households \( N_{HH} \). The log-likelihood function of observing \( s \) successes in \( t \)
trials for a given probability $p$ by a single household is given by

$$ L(t, s; p) = \log \binom{t}{s} + s \log p + (t - s) \log (1 - p) $$

Observe that for the log-likelihood, the log of the binomial coefficient $\log \binom{t}{s}$ for given vectors of trials and successes is independent of $p$, and can thus be ignored in the maximization. Given vectors $s$ and $t$, we want to estimate the maximum likelihood that those observations were generated by a vector $p = [p_1, ..., p_N]$ where $p_i \in \mathbb{P}$ with $\mathbb{P} \in [0, 1]^N$, i.e., having $N < N_{HH}$ entries. For a given $P$, the log-likelihood (omitting the binomial coefficient term) is given by

$$ L(P) = \sum_{i=1}^{N_{HH}} \max_{p \in \mathbb{P}} \{s_i \log p + (t_i - s_i) \log (1 - p)\} \quad (4.1) $$

so that for each $i$, we are assigning an element $p \in \mathbb{P}$ that maximizes the $i$'s observation likelihood. To get the MLE, we maximize $L(P)$ over the set $\mathbb{P}$.

### 4.3 Attention rate distribution

Table 4.1 reports our point estimate and standard errors for the discrete distribution $H$ of attention types in our sample of borrowers, using our main group-based specification with $gap > 0.5\%$. It also shows the corresponding unconditional ergodic average origination distribution in our Approximate Pooling MPE. For alternative specifications, Table 4.2 shows the MLE specification for $gap > 0\%$, and Table 4.3 shows the MLE specification for $gap > 1\%$.

<table>
<thead>
<tr>
<th>$\chi$</th>
<th>$p(\chi)$</th>
<th>Std Error $p$</th>
<th>$H(\chi)$</th>
<th>$G(\chi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0007</td>
<td>0.811</td>
<td>0.619</td>
</tr>
<tr>
<td>0.2343</td>
<td>0.0193</td>
<td>0.0002</td>
<td>0.061</td>
<td>0.087</td>
</tr>
<tr>
<td>0.5627</td>
<td>0.0458</td>
<td>0.0004</td>
<td>0.07</td>
<td>0.131</td>
</tr>
<tr>
<td>1.3882</td>
<td>0.1092</td>
<td>0.001</td>
<td>0.043</td>
<td>0.109</td>
</tr>
<tr>
<td>5.2775</td>
<td>0.3558</td>
<td>0.005</td>
<td>0.014</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Table 4.1: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on households and months with $gap > 0.5\%$, weighted by average loan amount. The average attention rate is $\bar{\chi}_H = 19.0\%$. 


Table 4.2: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on households and months with $gap > 0\%$, weighted by average loan amount. The average attention rate is $\bar{\chi}_H = 13.5\%$.

<table>
<thead>
<tr>
<th>$\chi$</th>
<th>$p(\chi)$</th>
<th>Std Error $p$</th>
<th>$H(\chi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0006</td>
<td>0.788</td>
</tr>
<tr>
<td>0.1799</td>
<td>0.0147</td>
<td>0.0001</td>
<td>0.064</td>
</tr>
<tr>
<td>0.3953</td>
<td>0.0323</td>
<td>0.0002</td>
<td>0.08</td>
</tr>
<tr>
<td>0.8517</td>
<td>0.685</td>
<td>0.0006</td>
<td>0.046</td>
</tr>
<tr>
<td>2.4855</td>
<td>0.1871</td>
<td>0.0024</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table 4.3: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on households and months with $gap > 1\%$, weighted by average loan amount. The average attention rate is $\bar{\chi}_H = 22.8\%$.

<table>
<thead>
<tr>
<th>$\chi$</th>
<th>$p(\chi)$</th>
<th>Std Error $p$</th>
<th>$H(\chi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0009</td>
<td>0.857</td>
</tr>
<tr>
<td>0.2845</td>
<td>0.0234</td>
<td>0.0003</td>
<td>0.043</td>
</tr>
<tr>
<td>0.753</td>
<td>0.0608</td>
<td>0.0006</td>
<td>0.054</td>
</tr>
<tr>
<td>2.0478</td>
<td>0.1569</td>
<td>0.0018</td>
<td>0.031</td>
</tr>
<tr>
<td>7.6738</td>
<td>0.4724</td>
<td>0.0063</td>
<td>0.015</td>
</tr>
</tbody>
</table>

4.4 Household sub-samples

In the main text, we document substantial heterogeneity in the degree of household attention in mortgage refinancing decisions in the population. In this section, we investigate which observable household characteristics are related to attention.
To begin with, consider Figure 4.1, which shows the differences of the loan-amount weighted cumulative distributions in $\hat{p}_i$ for different sample splits, based on household characteristics. The left panel depicts the difference in cumulative distribution functions between the “below median FICO” sample and the “above median FICO” sample. For that sample split, the “above median FICO” distribution first-order stochastic dominates the “below median FICO” distribution when restricting to rate gaps greater than 0.5% or 1%, suggesting that borrowers with higher credit scores exhibit more attention than borrowers with lower credit score.\(^4\) This potentially reflects a greater degree of financial frictions faced by these latter households.

In the right panel of Figure 4.1, we split the sample along the median average loan balance outstanding. This time, the “above median loan amount” distribution first-order stochastic dominates the “below median loan amount” distribution for all considered gap restrictions. This reflects greater inattention from households with smaller mortgages; for such households, the cost of making mistakes in refinancing is lower than that for households with larger mortgages, potentially rationalizing this observed difference in the distribution of the estimated parameter $\chi_i$.

Using a clustering approach identical to that described in the main text (with the notable exception that we constrain the estimation to the group means $\chi = \{\chi_k\}_{k=1,...,N}$ from the baseline procedure), we estimate the attention distributions $H_{\text{split}}(\chi)$ for each of our sub-samples, compute the corresponding origination distribution $G_{\text{split}}(\chi)$, and summarize our results in Table 4.4.

\(^4\)When we focus on gaps greater than 0%, the “above median FICO” distribution almost first-order stochastic dominates the “below median FICO” distribution – with the only exception of $\hat{p}_i$ between 0.15 and 0.5, for which the difference in cumulative distributions is slightly negative.
Table 4.4: Group-based estimation of the attention distribution, assuming $N = 5$ homogeneous groups, focusing on households and months with gap $> 0.5\%$, weighted by average loan amount, holding the $\chi$ vector constant at its baseline estimated level. The average attention rate for FICO+ is $\bar{\chi}_H = 23.0\%$ (population) and $\bar{\chi}_G = 60.3\%$ (refinancers); for Loan+, the corresponding averages are $\bar{\chi}_H = 24.1\%$ and $\bar{\chi}_G = 65.6\%$.

<table>
<thead>
<tr>
<th></th>
<th>FICO+</th>
<th>Loan+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi$</td>
<td>$H(\chi)$</td>
<td>$G(\chi)$</td>
</tr>
<tr>
<td>0.0</td>
<td>0.763</td>
<td>0.551</td>
</tr>
<tr>
<td>0.2343</td>
<td>0.078</td>
<td>0.104</td>
</tr>
<tr>
<td>0.5627</td>
<td>0.089</td>
<td>0.157</td>
</tr>
<tr>
<td>1.3882</td>
<td>0.054</td>
<td>0.129</td>
</tr>
<tr>
<td>5.2775</td>
<td>0.017</td>
<td>0.059</td>
</tr>
</tbody>
</table>

5 Quantitative implications

5.1 Redistribution using covariates

In this section, we study the degree of redistribution amongst households of different observable characteristics. Figure 5.1 shows (blue line) the average value function difference $\Delta V(r; \chi)$ between households with FICO scores above and below the population median, leveraging our clustering procedure and the allocation of each individual in our sample into one specific attention type $\chi_i$. The ergodic average difference in value function between these two groups is around 90bps p.a., with high-FICO households better off than low-FICO households. The red line shows the same result for above and below median loan outstanding. The ergodic difference amounts to around 70bps p.a. The ergodic differences for both the lifetime cost and for the ergodic average coupons for a host of covariates, most of them measured at the ZIP code level, are given in Table 5.1. To judge the robustness of the ZIP code level variables, we calculate the ZIP code level average FICO score from the household level observations.\(^5\) We see that household level observations give about 20% stronger results than the ZIP code level observations for FICO results. Table 5.1 suggests that the cross-subsidies we are documenting are regressive – in the sense that lower income households tend to be less attentive, and thus pay on average greater mortgage interest payments than higher-income households.

---

\(^5\)Household ZIP code level covariate is the average of the household’s time series over the relevant zip code value.
Figure 5.1: Cross-subsidies by above/below median sample splits: The blue line shows the household’s lifetime value function difference between being an above median FICO household (FICO+) vs. a below median FICO household (FICO-); the red dashed line shows the difference between an above median vs. a below median loan amount household (loan+ vs loan-).

|                                | $\mathbb{E}[\mathbb{E}^H[\Delta V(r, \chi)|+]-\mathbb{E}^H[\Delta V(r, \chi)|-]]$ | $\mathbb{E}^H[c_\infty(\chi)|+]-\mathbb{E}^H[c_\infty(\chi)|-]$ |
|--------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| FICO                          | 94                                                                                | -6                                                                                |
| FICO (ZIP)                    | 77                                                                                | -5                                                                                |
| principal balance             | 68                                                                                | -4                                                                                |
| combined LTV                  | -27                                                                               | 2                                                                                 |
| home-ownership rate (ZIP)     | 9                                                                                 | -1                                                                                |
| less than high-school education (ZIP) | -15                                               | 1                                                                                 |
| high-school education (ZIP)   | -2                                                                                | 0                                                                                 |
| some college education (ZIP)  | 2                                                                                 | 0                                                                                 |
| bachelor and above education (ZIP) | 18                                                | -1                                                                                |
| median income (ZIP)           | 27                                                                                | -2                                                                                |
| population share below 35 (ZIP) | -15                                               | 1                                                                                 |
| median age (ZIP)              | 27                                                                                | -2                                                                                |
| population (ZIP)              | -38                                                                               | 3                                                                                 |

Table 5.1: Ergodic differences by above/below median sample splits (in bps)
Consider a continuum of risk-neutral small firms of measure 1. Each firm generates income normalized to 1, has debt with notional balance normalized to $b$, and technology that fails with intensity $\lambda(x_t, \chi)$, with $\lambda(\cdot, \cdot)$ a known positive function that is increasing in $x_t$ and increasing in $\chi$. $x_t$ is an observable, aggregate variable that represents the state of the economy; it follows a diffusion with drift $\mu(x)$ and volatility coefficient $\sigma(x)$. $\chi$ is a firm-specific, time-invariant object that represents the intrinsic quality of the firm’s project. Importantly, $\chi$ is not observable by the banking sector. The firm quality distribution in the economy is $H$. When a firm’s project fails, the firm ends up defaulting on its existing debt. At such time, a new firm immediately enters the economy, with the same quality, so as to preserve the distribution $H$ over permanent quality heterogeneity.

The banking sector is risk-neutral and competitive; banks provide funding to small firms via loan contracts that mature at Poisson arrival rate $\nu$ and that carry an interest rate equal to the sum of (i) the risk-free rate $r$ and (ii) a credit spread $s_t$, fixed and determined at the time the loan is originated, and meant to compensate banks for expected future credit losses. Firms have the option to refinance their bank debt early, subject to potential refinancing frictions – they must bear fixed debt issuance costs $\psi$, and only make decisions at discrete points in time, arriving with intensity $\alpha$.

A firm currently financed with a loan at interest rate spread $s$ has equity value

$$V(x, s) := \sup_{a \in A} E_{x, s} \left[ \int_0^{\tau_x} e^{-rt} \left( 1 - \left( r + s_t(a) \right) b \right) dt - a_t \psi dN_t^{(\alpha)} \right],$$

s.t. $d s_t(a) = \left( S(x_t) - s_t(a) \right) \left( a_t dN_t^{(\alpha)} + dN_t^{(\nu)} \right)$,

where $A$ is a set of progressively measurable binary actions $a = \{a_t\}_{t \geq 0}$ such that $a_t \in \{0, 1\}$ at all times, $S(x_t)$ is the equilibrium credit spread charged by banks on new loans when the aggregate state of the economy is $x_t$, $\tau_x$ is the firm’s default time (with time-varying intensity $\lambda(x_t, \chi)$), $N_t^{(\nu)}$ (resp. $N_t^{(\alpha)}$) is a counting process for maturity events (resp. refinancing decisions).

Firms refinance whenever the economy is improving “sufficiently”; their decision depends on the spread $s$ over the risk-free rate currently paid on their loan. Specifically, a firm optimally refinances when $s - S(x) \geq \theta(x)$, where the state-dependent spread threshold $\theta(\cdot)$ satisfies

$$V(x, S(x)) - \psi = V(x, S(x) + \theta(x))$$

$^6$The parameter $b$ can thus be interpreted as the debt-to-income ratio of a given firm.

$^7$This assumption can easily be relaxed, by assuming for example that investors can also rely upon a public and noisy signal of firm quality; in that case, the equilibrium would result in separation based on the public signal, and pooling for the private signal.
Banks are competitive when offering new loans to a new customer firm. The shadow price of a given $1$ notional loan to a borrower with known quality $\chi$ is given by

$$P(x, s; \chi) = \mathbb{E}_x \left[ \int_0^{\tau_{\chi} \wedge \tau_\theta} e^{-(r + \nu)t} (r + s + \nu) \, dt + \mathbb{1}_{\tau_{\chi} < \tau_\theta} e^{-(r + \nu)\tau_{\chi}} \rho + \mathbb{1}_{\tau_{\chi} > \tau_\theta} e^{-(r + \nu)\tau_\theta} \right],$$

where $\tau_{\chi}$ (resp. $\tau_\theta$) is the firm’s default time (resp. loan prepayment time), and $\rho$ is the recovery rate realized by creditors upon a default. Banks break-even when a new loan is issued; they price loans under a (potentially state-dependent) firm quality distribution $G$, so that their no-profit loan origination condition can be written

$$\mathbb{E}^G [P(x, S(x); \chi)] = 1$$

The origination distribution $G$ is distinct from the distribution over firm quality $H$, with $G$ skewed towards riskier firms, since (i) high risk firms will have a refinancing spread threshold $\theta$ that is lower than that of low risk firms, and (ii) riskier firms default at higher intensity, and are replaced by firms with identical quality that will immediately seek loan funding. In this model, low quality firms are subsidized by high quality firms, since they finance themselves at credit spreads more advantageous than if banks could observe the firm quality $\chi$. An improvement in aggregate credit market conditions triggers a wave of loan refinancing events, consistent with the data. This model emphasizes the capital misallocation taking place in the banking sector due to the unobserved firm quality and the competitive banking sector.
References


