Refinancing Frictions, Mortgage Pricing and Redistribution*

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

November 2022

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 which lead to more frequent refinancing lead to higher equilibrium mortgage rates and 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, as well as the seminar participants at USC, Northwestern University, UCLA, Copenhagen Business School, Cleveland Fed, Society of Economic Dynamics (2022), Chicago Fed and Racial Bias Workshop (2022), Rio FGV (2022) and TAU (2022), NYU Stern, and BU Finance for helpful discussions and feedback. Fabrice Tourre gratefully acknowledges financial support from the Danish Finance Institute as well as 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; vavra@uchicago.edu

1 Introduction

Many borrowers do not refinance their fixed-rate mortgage, even when it is financially beneficial to do so.[1] Heterogeneity in refinancing behavior can in turn lead to substantial inequality in the mortgage coupons borrowers actually pay over time (Gerardi, Willen, and Zhang, 2020; Zhang, 2022). Several mortgage market reforms have been discussed which might encourage more optimal refinancing and reduce the inequality resulting from this heterogeneity. However, assessing the full impact of any reform on borrower welfare requires also accounting for potential changes in lending behavior and resulting equilibrium mortgage rates (Campbell, 2006). For example, "automatically refinancing" mortgages would eliminate refinancing disparities across borrowers, but would also lead lenders to charge higher rates on newly originated mortgages. Despite their potential importance, a systematic analysis of these equilibrium forces has been limited by the complexity of the equilibrium environment.

In this paper, we develop a tractable framework to study equilibrium mortgage pricing in environments with borrower heterogeneity, which can be used to analyze the redistributive consequences of counterfactual mortgage market interventions. We begin by using this framework to characterize various general properties of equilibrium mortgage pricing. We then estimate our model using U.S. mortgage micro data and show that these estimates imply quantitatively important equilibrium effects in many applications.

We develop our framework in three key steps. In the first step, we provide a partial equilibrium characterization of optimal refinancing decisions for borrowers facing the two main frictions identified by the past literature (e.g. Andersen et al. (2020)). Specifically, we allow for both "inattention" or other non-monetary frictions (which generate time-dependent inaction) as well as fixed monetary costs of refinancing (which generate state-dependent inaction) and solve for the optimal behavior of borrowers, taking mortgage rates as given. Optimal behavior reduces to the influential Agarwal, Driscoll, and Laibson (2013) refinancing rule in certain special cases, but we derive more general results in an environment that contains a richer set of refinancing frictions.[2] Notably, we show that the introduction of the inattention friction dampens the effect of fixed costs on optimal refinancing decisions. When inattention frictions rise, rational borrowers act "as if" fixed costs decline and they choose to refinance at smaller "rate gaps" (the difference between the mortgage coupon and

[1] See e.g. Keys, Pope, and Pope (2016) and Andersen, Campbell, Nielsen, and Ramadorai (2020).

[2] Our analysis also allows for a less restrictive interest rate process, which is important for embedding the individual refinancing problem into a model with endogenous rates.

the current market interest rate on a similar mortgage). Intuitively, greater inattention reduces the option value of waiting to refinance. Inattention also provides an alternative explanation for empirical studies that conclude, using the framework of Agarwal, Driscoll, and Laibson (2013), that a large fraction of households make mistakes by refinancing too early.

In our second step, we embed the 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 (thereafter, "MBS"), which pool together the monthly payments made by borrowers (net of any intermediation fees), and we solve for new-issue MBS prices and resulting mortgage rates in the Markov perfect equilibrium. This environment is similar to that in Berger, Milbradt, Tourre, and Vavra (2021), but we develop new results characterizing the relative role of different frictions for equilibrium pricing.

In particular, we show that changing the size of monetary fixed costs has only small effects on equilibrium mortgage pricing while changing the size of non-monetary inattention frictions has instead much larger effects. Intuitively, borrowers with large rate gaps will choose to refinance even when facing sizable fixed costs. This means that fixed costs primarily induce inaction for borrowers with small rate gaps. However, whether a borrower refinances away a small rate gap or not has only a small effect on lender profits. In contrast, the presence of inattention frictions can lead to borrowers with much larger rate gaps that matter a lot for lender profits. This implies that inattention frictions have much larger effects on the profitability of mortgages under different rate paths and thus their pricing than fixed costs of refinancing.

The third step of our framework introduces heterogeneous refinancing frictions across borrowers into this equilibrium environment, which allows us to explore redistributive effects of various mortgage market interventions. Consistent with institutional features of the U.S. agency MBS market, we assume that mortgage investors cannot observe or price-discriminate based on households' individual attention level, leading us to focus on a "pooling" equilibrium. In general, this asymmetric information framework adds the cross-sectional distribution over households' coupons and attention rates as an additional state variable in the investors' pricing problem, therefore substantially increasing the complexity of our model. For tractability, we make two simplifying assumptions.

The first simplification affects the household side of our model: we assume that while borrowers face fixed costs of refinancing, these are not paid *upfront* and are instead capitalized into a higher interest rate for the new loan. Unlike upfront fixed costs, these capitalized costs do not lead to state-dependent refinancing decisions, and so this assumption dramatically increases tractability. This is

a strong assumption, but two observations suggest that the benefit in tractability comes at little cost in terms of quantitative realism. First, this assumption aligns well with actual mortgage markets: more than 80% of mortgage origination costs in the U.S. are rolled into higher rates, rather than paid up-front (Zhang, 2022). Second, the conclusion from the first two steps of our analysis that upfront fixed costs have modest effects on equilibrium pricing implies 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.

The second simplification affects the investors' side of our model: in order to reduce the dimensionality of the state space relevant for the pricing problem, we assume that investors exhibit a simple form of bounded rationality, valuing mortgages based on current short rates together with the *average* distribution of attention among refinancing borrowers. That is, the interest rate is a state variable in the lender problem but the distribution of attention in newly originated mortgages is not. Under this assumption, lenders account for the fact that the attention distribution amongst refinancing households differs in high and low interest rate environments but not for the fact that this distribution further depends on the entire past history of rates.[3] In conjunction with the no-fixed-costs assumption, investors' bounded rationality is key for tractability, and we provide some evidence that it is likely to have little quantitative impact on conclusions.[4]

When combined, these two simplifications allow us to solve for the pooling equilibrium of this (ex-ante) heterogeneous borrower model with aggregate shocks without having to introduce distributional state variables, and this is the key to our framework's tractability.[5] Indeed, our approximate pooling equilibrium framework is simple enough that we can write sufficient conditions to guarantee its existence and derive a number of important properties. For instance, we can precisely characterize the sense in which borrower heterogeneity matters for mortgage pricing, through a simple covariance adjustment term.

We next turn to the model's quantitative implications. We start by exploring the model's ability to fit observed mortgage outcomes and then turn to implications for counterfactual policy. Using a

[3]For example, holding current rates constant, the pool of refinancing mortgages will be more tilted towards high attention types if interest rates recently fell than if they did not.

[4]Specifically, we show that it makes little quantitative difference whether lenders base pricing on the average distribution of attention without conditioning on current rates or the distribution of attention conditional on current interest rates. The fact that this simpler form of state-dependence in the distribution of attention has little quantitative impact suggests that higher order forces which depend on the full history of past rates are likely even less important.

[5]An alternative approach would be to summarize distributional state variables, along the lines of Krusell and Smith (1998) (and its progeny). However, as we discuss in more detail in the text, these solution methods are ill-suited for our economic environment since the key logic underlying the viability of these methods is violated in our framework.

monthly borrower-level panel covering mortgages from 2005 to 2017, we estimate the cross sectional distribution of borrower attention via a maximum likelihood estimation that sorts borrowers into distinct groups. This econometric procedure identifies substantial cross-borrower heterogeneity in refinancing frictions. After identifying the level of heterogeneity observed in the data, we then test the model's ability to match the time-series of realized mortgage rates. Taking U.S. treasury yields from 2005 to 2017 as given, together with estimates of intermediation costs from the literature, we calculate implied equilibrium mortgage rates in the model and show that they align well with mortgage rates in the data, giving us confidence in the model's implications.

We then use the model to explore several counterfactual mortgage market interventions. In line with U.S. mortgage market institutions, we assume a pooling equilibrium in which lenders offer the same rate at origination to borrowers with different attention frictions. In this pooling equilibrium, low attention rate borrowers (which we will refer to, going forward, as "slow" borrowers) pay higher mortgage rates at origination than they would if they were not pooled with high attention types ("fast" borrowers), and high attention types pay lower rates. Thus, slow borrowers effectively subsidize fast borrowers through equilibrium price effects. We find that these effects are substantial. For example, in the counterfactual "separating" equilibrium, the fastest households would face mortgage rates at origination about 440bps higher than the slowest households.

Importantly, these redistributive forces arise from pooled mortgage pricing at origination and are thus distinct from the forces for inequality stemming from ex-post differences in refinancing rates and studied in Gerardi, Willen, and Zhang (2020). This means that accounting for equilibrium mortgage pricing amplifies the level of inequality measured in the prior literature. When we jointly account for both effects, we find significant redistribution across borrowers: the fastest ones, on average, make mortgage payments that are 14.8% lower than the slowest ones in life-time present value terms. Thus, moving from a pooling mortgage market equilibrium to a separating equilibrium could substantially reduce inequality arising from the liability side of households' balance-sheet.

Of course, the relevance of this particular counterfactual mortgage market equilibrium depends on the extent to which borrower attention is observable and thus potentially priced ex-ante. Such observable and priceable degree of attention may be smaller than the true borrower heterogeneity we measure ex-post. Empirical evidence, however, suggests that ex-ante priceable heterogeneity is substantial. For example, Gerardi, Willen, and Zhang (2020) shows important differences in refinancing rates by race. We have limited individual covariates in our mortgage data, but we find substantial differences in refinancing by credit score. Even when we aggregate our borrower

level data to zip codes to bring in additional covariates we find significant differences in refinancing by various observables. The extent of refinancing which can be predicted by these zip code level covariates is almost certainly a lower bound to what could be predicted using the same covariates at an individual level. Nevertheless, there is potential legal risk to fine-grained pricing of observable prepayment risk since it is correlated with various protected classes, and perhaps for this reason there is little active discussion about such reforms.

Moving from a pooling to a separating equilibrium is however not the only potential path to reducing mortgage market inequality induced by heterogeneous refinancing frictions. We thus explore the equilibrium effects of several counterfactual policies for reducing inequality which *have* been proposed. First, we explore the implications of moving to "automatically refinancing" mortgages, which would refinance automatically with no active borrower intervention when rates decline. While automatically refinancing mortgages would indeed reduce inequality, our model implies that they would have quantitatively significant equilibrium effects, substantially reducing their benefits for borrowers. Indeed, even though automatically refinancing mortgages lead to much more refinancing for inattentive borrowers, they also lead to an increase in average mortgage rates of about 130bps at origination, offsetting some of these gains. That is, automatically refinancing mortgages yield individual time-paths of mortgage coupons which decline more rapidly on average but that start from a higher initial value. Consequently, in equilibrium a movement to automatic refinancing mortgages changes the slope and the intercept of the mortgage coupon path in offsetting ways. The increase in the initial value of the mortgage coupon offsets some of the benefits from refinancing more rapidly: automatically refinancing mortgages ultimately still result in a reduction in average coupons of around 30bps for the slowest group, but this is smaller than the 90bps reduction that would arise assuming no change in mortgage pricing.

In addition to its distributional effects within the mortgage market, we note that this equilibrium increase in interest rates is likely to have important implications for access to housing markets. In particular, mortgage market reforms which lead to increases in equilibrium interest rates may exclude households who are at DTI limits from the housing market entirely. Borrowers only benefit from the more rapidly declining mortgage coupons induced by automatic refinancing if they are able to afford a mortgage at the initial higher rate in the first place. Our model does not analyze initial home purchases and instead focuses on cross-subsidies across borrowers from refinancing. However, a simple back of the envelope calculation suggests that the increase in interest rates arising from a move to automatically refinancing mortgages might force around 20% of borrowers to select smaller

homes requiring a smaller initial mortgage balance.

Next, rather than exploring changes in mortgage contracts (which would apply equally to all households) we instead explore policies that change the level of borrower frictions for a subset of the population. For example, financial literacy programs might lead some but not all households to refinance more optimally. If slow households increase their attention, they ultimately pay lower mortgage rates on average, although just like in the case of automatically refinancing mortgages, much of the benefits are dampened by equilibrium forces. However, slow borrowers who do not become more attentive are unambiguously made worse off by these policies: if other borrowers become more attentive and refinance more rapidly, lenders will increase mortgage rates in the pooling equilibrium. This means that the borrowers who remain slow to refinance now pay higher rates at origination and receive no compensating benefit arising from faster refinancing.

Lastly, we study the impact of the recent rise in FinTech and non-bank mortgage lending onto mortgage market interest rates. Our micro-data suggests that borrowing from non-banks leads households to be effectively more attentive – potentially via nudging and other devices used by non-bank lenders to encourage borrowers to refinance. The 100bps per month effective increase in attention rate is substantial; moving from a world in which banks are the dominant lenders in the mortgage market to one in which non-banks originate all mortgages leads equilibrium interest rates to rise by 50bps.

Overall these results show that accounting for equilibrium effects can matter in quantitatively important ways for evaluating the consequences of mortgage market reforms. Discussion of policies which interact with inequality should think carefully about these equilibrium interactions.

While our paper focuses on mortgage markets, our framework can be applied more broadly and be used to analyze redistribution induced by pooling equilibria in many other contexts with heterogeneity. Consider for example the classic labor market environment of Harris and Holmstrom (1982), in which risk neutral firms set wage contracts to insure risk averse workers who have stochastic productivity but cannot commit to turn down outside offers. We can use our framework in order to analyze a version of this environment extended to include various frictions faced by workers when moving from one job to another, as well as heterogeneity in the arrival rate of outside offers. This type of heterogeneity is pervasive in the world, and we can use our framework to explore its implications for wage determination.[6] In a pooling wage equilibrium, workers with infrequent

[6]There 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, other works might have constraints like children in school or spousal employment constraints that make them viewed as less "movable" by potential outside employers.

outside offers receive lower wages than they would in a separating equilibrium and thus effectively subsidize the wages of low loyalty workers with frequent outside offers.[7] Several other economic settings lend themselves to our modeling framework; while we leave their quantitative evaluation for future research, these environments all share the following features: on one side of the market, economic agents who are ex-ante heterogeneous make dynamic discrete choices about supplying or purchasing a particular good or service subject to some frictions, and the other side of the market is competitive but cannot, for informational or legal reasons, price-discriminate.

The remainder of the paper is structured as follows: Section 2 discusses the related literature. Section 3 treats household refinancing behavior in partial equilibrium, in which mortgage rates are exogeneously given. Section 4 introduces general equilibrium, both with homogeneous and also with ex-ante heterogeneous households. Section 5 leverages the general equilibrium nature of our model to assess policy experiments and counterfactuals. In Section 6, we present our data and discuss our estimation of household's heterogeneity. Section 7 then quantifies the pricing and welfare impact of this estimated heterogeneity through the lens of our model. Finally, Section 8 discusses the extent to which our framework can be used to study other economic environments that share certain key properties.

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 U.S. households fail to refinance even when it appears optimal to do so. They conduct a mail campaign targeted towards a sample of homeowners that could benefit from refinancing and find even this nudge induces little take-up, suggesting that the frictions inhibiting optimal refinancing are large. Agarwal, Rosen, and Yao (2016) provide empirical evidence that US households fail to refinance their mortgage optimally, and measure the cost of these financial mistakes. While their micro-data is not as detailed and rich as the Danish data in Andersen et al. (2020), they also study the heterogeneous nature of these inefficient refinancing decisions using zip code level and county level demographic characteristics, suggesting that household's suboptimal decision making is correlated with income and FICO scores, which are proxies for levels of financial sophistication (see also Amromin, Huang, Sialm, and Zhong (2018)).

[7] This type of pooling is likely to obtain both because "loyalty" cannot be directly observed and because it is illegal to set wages based on many characteristics, like marital status, which might correlate with the outside offer rate.

In complementary work to our paper, Fisher, Gavazza, Liu, Ramadorai, and Tripathy (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 time-limited teaser rate – say 3 years – which then resets to the market rate after this time. Differences in fixed costs induce heterogeneity in both the level of mortgage balances and propensity to refinancing. Using a partial equilibrium consumption model, they consider the consequences of a revenue equivalent fixed rate mortgage contract for the distribution of mortgage balances. 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 only provides general equilibrium solutions for the counterfactual separating equilibria by type. Like these two papers, our analysis is similarly motivated by heterogeneity in refinancing propensities, however, we develop a general equilibrium mortgage pricing framework which endogenizes MBS prices and mortgage rates, and show that these equilibrium forces can have important quantitative implications for many mortgage market interventions.

Two other closely related papers do study models with equilibrium mortgage pricing but in environments without permanent borrower heterogeneity. Guren, Krishnamurthy, and McQuade (2021) study the implications of mortgage market reforms in an equilibrium model with borrower refinancing and risk-neutral competitive mortgage investors. Their model features ex-post heterogeneity arising from idiosyncratic income realizations and moving shocks, but households are ex-ante identical. This means that their model cannot speak to the distributional issues that are the focus of our paper. In addition, frictions to refinancing in their model arise solely through fixed costs, which substantially complicate their analysis. The model environment in Berger et al. (2021) is most similar to our setup, but like Guren, Krishnamurthy, and McQuade (2021), their setup assumes homogeneous borrowers, and they focus on entirely different questions. Relative to Berger et al. (2021), our contribution is two-fold: first, we more fully analyze equilibrium and the importance of various frictions for pricing. Second, and more importantly, we extend their framework to an environment with permanent borrower heterogeneity and show that this heterogeneity generates important equilibrium effects on inequality.

Cross-sectional heterogeneity in household attention rates can be extracted from observed mortgage refinancing behavior using methods similar to those employed by Andersen et al. (2020), who estimate a refinancing hazard model consisting of a rate-incentive-dependent hazard and an at-

tention component. While our US household data is not as rich as the one available for Danish households, it is nonetheless sufficiently detailed to allow us to tease out the inattention distribution in the population, and understand household and mortgage characteristics that influence such inattention. In order to quantify the degree of cross-sectional attention heterogeneity in our data, we apply a clustering algorithm similar to that used by Lewis, Melcangi, and Pilossoph (2019) to estimate heterogeneity in marginal propensities to consume.

Many articles have studied, theoretically or empirically, the drivers of the wealth distribution, and various measures of wealth inequality. Many of these articles focus on one of the key sources of wealth inequality: heterogeneous capital returns. Benhabib, Bisin, and Zhu (2011) shows that the tails of the wealth distribution increase with capital income risk, and the persistence of capital returns. Bach, Calvet, and Sodini (2020) and Fagereng, Guiso, Malacrino, and Pistaferri (2016) provide empirical evidence about those heterogeneous capital returns from Swedish and Norwegian household data. While most of the existing studies focus on the return heterogeneity on the *asset* side of household's balance-sheet, financial literacy and inattention in the mortgage market contribute to wealth inequality via realized return heterogeneity on the *liability* side of the households' balance-sheet. While this heterogeneity, in magnitude, is likely to be more modest than the one documented on the asset side, it is however very persistent, and is thus likely to have a non negligible effect on wealth inequality.

3 Household refinancing behavior in partial equilibrium

In this section we present a simple 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 who face two types of potential refinancing frictions, which will lead to *state-dependent* and *time-dependent* inaction, as will be discussed in greater details shortly. For now, we assume a partial equilibrium setting in which mortgage rates are given; we will endogenize those rates when we study the general equilibrium properties of our model 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 the subjective rate ρ. Each household carries a long-term fixed-rate

prepayable mortgage, with coupon rate c_t, and constant unit notional balance. We denote m_t the prevailing mortgage market interest rate, i.e., the rate a household refinancing at time t can lock-in. Household's 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_χ. In addition to making refinancing decisions, households also move from one house to another at intensity ν; when doing so, they must reset their mortgage coupon to the prevailing mortgage rate, and bear a moving-related cost ψ_ν.[8]

Given our focus on a Markovian environment, we assume that the aggregate uncertainty is summarized by a latent state vector x_t, a possibly multi-dimensional, time-homogeneous, Itô process with drift $\mu(x)$, diffusion $\sigma(x)$ and infinitesimal generator $\mathcal{L}$.[9] The mortgage market interest rate is then a function $m_t = m(x_t)$ of this latent state vector. For now, we assume $m(\cdot)$ is continuous in x, and in Section 4 we prove that the equilibrium of our economy must satisfy this property.

Later on in the paper, we consider households that might differ in their attention intensity χ, and the consequences of this (permanent) 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 the parameter χ.

3.2 Interpreting the refinancing frictions

We first discuss various interpretations for the two frictions in our model. The inability for households to make decisions continuously is sometimes referred to as *time-dependent* inaction, and it has been featured in a large literature on rational inattention.[10] The attention parameter χ should be viewed as a stand-in for various non-monetary frictions. Some households for example cannot refinance even if it was beneficial for them to do so, due to insufficient home equity or non-verifiable income (see Beraja, Fuster, Hurst, and Vavra (2019)). Households have various degrees of financial literacy, and they might only partially understand the mechanics of refinancing a mortgage; they might for instance not be aware that monetary refinancing costs are modest and can be rolled into a higher mortgage interest rate. Thus, while we will refer to household's *inattention*, this friction

[8] Absent moving-related costs ($\psi_\nu = 0$), one can view ν as the sum of (a) a moving intensity, (b) an amortization intensity, and (c) a default intensity, under the assumption that contractual mortgage balances amortize exponentially (which is, to some extent, only an approximation of the actual amortization profile of a standard 30-year mortgage contract), and defaults are treated similarly to prepayments from the perspective of mortgage investors.

[9] The generator $\mathcal{L}$ is defined over all functions f of class $\mathcal{C}^2$ via $\mathcal{L}f(x) = \mu(x) \cdot \partial_x f(x) + \frac{1}{2}\text{trace}\left(\sigma'(x)\partial_{xx'}f(x)\sigma(x)\right)$.

[10] 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.

should be understood as encompassing a wide set of environmental and behavioral factors.

The upfront closing costs borne by households when refinancing leads to *state-dependent* inaction, as it arises when household's incentives causes it to *decide* to stay put, and such incentives vary as the economic environment evolves. These upfront closing costs include application fees as well as the "points" that are payable out of pocket by borrowers on the transaction closing date; they also represent a component of revenues collected by lenders upon mortgage origination.[11]

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

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

$$\begin{aligned} V(x,c) :&= \inf_{a\in\mathcal{A}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(a)}dt + \psi_\nu dN_t^{(\nu)} + a_t\psi_\chi dN_t^{(\chi)}\right)\right], \\ \text{s.t.} \qquad dc_t^{(a)} &= \left(m(x_t) - c_{t-}^{(a)}\right)\left(a_t dN_t^{(\chi)} + dN_t^{(\nu)}\right), \end{aligned} \tag{1}$$

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 χ (resp. ν), $c_t^{(a)}$ is the coupon rate on the mortgage for a household following strategy a, and the subscript on the expectation indicates 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 they choose to keep their existing mortgage. V captures all mortgage liabilities – related to household's *current* mortgage (at rate c), as well as 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 Appendix A.1, 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*

$$(\rho + \nu + \chi)\, V(x,c) = c + \mathcal{L}V(x,c) + \nu\left[V(x, m(x)) + \psi_\nu\right] + \chi \min\left[V(x,c), V(x, m(x)) + \psi_\chi\right] \tag{2}$$

[11] See Zhang (2022) for a description of the trade-off between "points" and the mortgage "rate" at origination.

The optimal refinancing choice satisfies

$$a^*(x,c) = \mathbb{1}_{\{c-m(x)\geq\theta(x)\}}, \tag{3}$$

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

$$V(x, m(x)) + \psi_\chi = V(x, m(x) + \theta(x)). \tag{4}$$

This proposition holds for any arbitrary (continuous) mortgage function $m(\cdot)$, i.e., not only the equilibrium one. Proposition 1 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 which we discuss now.

First, consider the environment where households do not bear any upfront closing costs. In that 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 ($\psi_\chi = 0$), the rate gap threshold is $\theta(x) = 0$, and the optimal refinancing choice is $a^*(x,c) = \mathbb{1}_{\{c\geq m(x)\}}$.*

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 σ. Introduce the constants*

$$\eta_\chi := \frac{\sqrt{2\left(\rho+\nu+\chi\right)}}{\sigma} \qquad \epsilon_\chi := \frac{\left(\rho+\nu\right)\left(\eta_0+\eta_\chi\right)}{\chi}$$

The household's value function satisfies $V(m,c) = \frac{c}{\rho} + v(z)$, with $z = c - m$ and v admitting functional form as in (A.1-A.2). The (state-independent) rate gap threshold $\theta > 0$ satisfies the implicit equation

$$e^{-\eta_0\theta} + \left(\eta_0+\epsilon_\chi\right)\theta = 1 + \left(\eta_0+\epsilon_\chi\right)\left(\rho+\nu\right)\psi_\chi. \tag{5}$$

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

$$\hat{\theta} = \sqrt{\frac{2}{\eta_0}\left(1+\frac{\epsilon_\chi}{\eta_0}\right)\left(\rho+\nu\right)\psi_\chi + \left(\frac{\epsilon_\chi}{\eta_0^2}\right)^2} - \frac{\epsilon_\chi}{\eta_0^2}. \tag{6}$$

Moreover θ is an increasing function of the attention rate χ, and asymptotically:

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

$$\lim_{\chi \to +\infty} \theta = \frac{1}{\eta_0} \left[1 + \eta_0 \psi_\chi (\rho + \nu) + W\left(-\exp\left\{ -1 - \eta_0 \psi_\chi (\rho + \nu) \right\} \right) \right], \tag{8}$$

where W is the Lambert W function.

Proposition 2 is proven in Appendix A.2. It generalizes the results of Agarwal, Driscoll, and Laibson (2013) (thereafter, "ADL") to the case where households are inattentive and face moving costs. 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 to mortgage rates, formula (7) implies that the rate gap threshold converges to the annuity value of the upfront closing cost ψ_χ, computed at the effective discount rate $\rho + \nu$. Most importantly, a decrease in χ reduces the rate gap threshold; intuitively, when the rate of attention a household pays to the mortgage market decreases, the household exercises its refinancing option "sooner" when the opportunity arises. Figure 1 illustrates the sensitivity of the cutoff θ to the household's attention rate. In our quantitative applications, we will show that the parameter χ in the data is around 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 by Agarwal, Driscoll, and Laibson (2013) in their model without inattention.

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, Plosser, Schnabl, and Vickery (2019) for instance conclude that borrowers in their data refinance at rate gaps that are on average too small, relative to the ADL threshold.[12] 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 too low rate gaps, as these empirical studies suggest), households may act rationally – refinancing aggressively when they have the chance – subject to their attention friction.

[12] Agarwal, Rosen, and Yao (2016) finds that households refinance at rate gaps with an average of 121bps, vs. an average ADL threshold of 158bps. Similarly, Fuster et al. (2019) finds that amongst refinancing households, more than half are executed at rates that are too small when assessed against the ADL threshold.

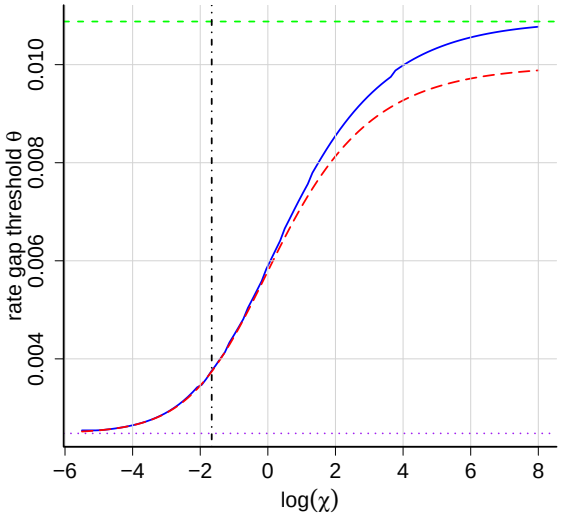

Figure 1: **Rate gap threshold θ vs. attention rate χ.** Solid blue (resp. long dash red) line represents rate gap threshold θ (resp. approximation $\hat{\theta}$ in (6)) in the case of m_t following a Brownian motion. Horizontal dash green line represents the limit for θ as $\chi \to +\infty$, as characterized in Agarwal, Driscoll, and Laibson (2013). Horizontal dotted purple line represents the limit for θ as $\chi \to 0$. Vertical dot-dashed black line shows $\chi = 19\%$, our estimate of the average attention rate (see Section 6). Figure computed for $\rho = 5\%$, $\nu = 7.4\%$, $\psi_\chi = 2\%$ and $\sigma = 1\%$.

Proposition 2 relies on the assumption that the mortgage rate m_t is a random walk. Instead, we want to focus on the general 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 will differ from those of the Brownian motion – in particular, equilibrium mortgage rates in our environment will be mean-reverting. In that 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? Appendix A.3 addresses this question by focusing on mortgage rates that follow an Orenstein-Uhlenbeck process, varying the degree of persistence. Moving from the pure random walk assumption to a mean-reverting process introduces a non-negligible amount of state dependence in $\theta(x)$, which then has 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 who manages to refinance can lock-in a low mortgage rate for a long time, without the need to refinance (and thus to incur upfront closing costs) in the near future. Instead, at high mortgage interest rates, rates are drifting downwards, and households might be reluctant to incur these costs today, given the projected future path of rates. Given the degree of persistence of mortgage

rates observed in the data, each p.p. increase in the current mortgage rate increases the rate gap threshold $\theta(x)$ by 4bps, so while this effect is non-negligible it is quantitatively modest.

4 Mortgage rates in general equilibrium

We now discuss the general equilibrium environment with mortgage investors and the resulting *equilibrium* mortgage rates. Reflecting a key feature of US mortgage markets, we assume that mortgages are pooled and traded by risk-neutral competitive investors that discount cash-flows at the short rate r_t, with r_t a smooth exogenous function $r(\cdot)$ of the latent state x. While households pay coupon c_t on their mortgage, investors only receive $c_t - f$, with a wedge f capturing fees charged by intermediaries for providing various services.[13] 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" π represents revenues generated by the original lender (in addition to those arising from upfront closing costs ψ_χ paid by households).[14] The sum $\pi + \psi_\chi$ 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, upfront closing costs ψ_χ have only small effects on equilibrium mortgage rates. Thus, the choice to include (or not) upfront closing costs will only have small effects on our conclusions. In practice, most households do not pay closing costs upfront and instead roll them into higher rates. Together these observations motivate us to then abstract from state-dependent frictions, which dramatically simplifies our subsequent analysis.

We initially focus on households who are ex-ante homogeneous in their attention rate χ. 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.

[13] In general, this fee rate f is meant to capture, amongst others, (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 thus enters into the broadly defined gain on sale π. 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 14.

[14] Total revenues – the upfront closing cost ψ_χ and the gain on sale π – compensate the lender for all costs incurred in connection with mortgage origination. Mortgage origination costs incurred by the initial lender 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 upfront by the original lender. See Fuster, Goodman, Lucca, Madar, Molloy, and Willen (2013) for a detailed description of mortgage lenders' costs of origination.

4.1 Homogeneous households

In this section, all households share the same attention parameter χ. 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 is a household with attention intensity χ, when the latent state is x:

$$P(x, c; \chi) := \mathbb{E}_x \left[\int_0^\tau e^{-\left(\int_0^t r(x_s) ds\right)} (c - f)\, dt + e^{-\left(\int_0^\tau r(x_s) ds\right)} 1 \right], \tag{9}$$

where τ is the (random) prepayment time. Importantly, the pricing function P implicitly depends on an assumed mortgage rate function $m(x)$, via the prepayment time τ. Since discounted debt prices must be martingales, P must satisfy the following Feynman-Kac equation:

$$r(x)P(x, c; \chi) = c - f + \mathcal{L}P(x, c; \chi) + \left(\nu + \chi \mathbb{1}_{\{c - m(x) \geq \theta(x)\}}\right) [1 - P(x, c; \chi)]. \tag{10}$$

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 π at the time of loan origination in order to recoup their costs. This pins down the mortgage rate $m(x_t)$ via the equilibrium condition

$$P(x, m(x); \chi) = 1 + \pi. \tag{11}$$

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 satisfying (10), and (iv) a mortgage rate function $m(x)$ that satisfies (11).*

Since the definition of P in (9) implicitly depends on a mortgage rate function $m(x)$, and since the equilibrium condition (11) defines $m(x)$ implicitly via the function P, this is a potentially complex fixed point problem. Our equilibrium concept is then analogous to Markov perfect equilibria studied in the sovereign or dynamic corporate debt literature.[15] In these environments, the existence and uniqueness of the equilibrium frequently depends on various assumptions. In the context of mortgage prepayments, the special case without upfront closing costs allows us to derive several

[15] See Auclert and Rognlie (2016) for an example of MPE in the context of sovereign defaults with short term debt, or Chatterjee and Eyigungor (2012) for an example with long term debt. In the corporate debt literature, see DeMarzo and He (2021) for an example of MPE where a firm cannot commit to its capital structure, or Décamps and Villeneuve (2014) when instead the firm can commit to maintaining a constant debt balance.

sharp results. We will shortly argue that in equilibrium, the general setup with upfront 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.

Proposition 3. *Assume finite attention rate (i.e. $\chi < \infty$) and assume short term rates r_t are positive and bounded. Absent upfront closing costs (i.e. $\psi_\chi = 0$),*

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 uni-dimensional with $r(\cdot)$ monotone increasing, there exists a unique MPE in which the mortgage rate function $m(x)$ is increasing in x.

The proof can be found in Appendix B.1. In our model, households optimize over their refinancing decisions, subject to various frictions, taking as given the behavior of mortgage market interest rates m_t. Investors price mortgages competitively, taking as given households' refinancing behavior. Proposition 3 tells us that this fixed point problem, absent upfront closing costs, always admits a unique solution. In that special setup, we can disentangle the household decision problem from the investors' pricing problem: irrespective of how mortgage rates evolve, households always want to refinance when their coupon is above the current mortgage rate.[16] This simplifying assumption also allows us to deliver several comparative static results. The first of these results highlights the efficiency implications of the funding of origination costs and its split between (i) the gain on sale π and (ii) upfront closing costs ψ_χ.

Proposition 4. *Consider the environment without upfront closing costs (i.e. $\psi_\chi = 0$), and assume x is uni-dimensional, with $r(x)$ monotone increasing. In the (unique) MPE, the ergodic average prepayment rate is invariant to the gain on sale π.*

The proof can be found in Appendix B.2. Regardless of the gain on sale π, in the absence of upfront costs borne by households ($\psi_\chi = 0$), the number of households refinancing per period is invariant to π. As a consequence of Proposition 4, the way in which loan origination costs are financed – either by the lender (via higher mortgage rates) and recouped via gain on sale ($\pi > 0$), or by the borrower ($\psi_\chi > 0$) – has an impact on ergodic average prepayment rates. For the same loan origination costs $\pi + \psi_\chi$, if part is borne upfront by the borrower $\psi_\chi > 0$ (and consequently

[16] This observation holds irrespective of the gain on sale $\pi \geq 0$. When households must pay upfront closing costs ($\psi_\chi > 0$), 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 sufficient gain on sale π, in order to recover origination costs.

lower π), then refinancing incentives, ergodic prepayment rates, and thus dead-weight losses are all reduced compared to $\psi_\chi = 0$. Next, we analyze the impact of households' attention on equilibrium outcomes.

Proposition 5. *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 χ.*

Proposition 5 is proved in Appendix B.3. It 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 C.1 illustrates the sensitivity of the equilibrium mortgage rate function to a range of attention parameters χ that are below and above our estimated average value (see Section 6.2.2); when χ increases from 50% to 150% of this value, the ergodic average of equilibrium mortgage rates increases by 34bps.

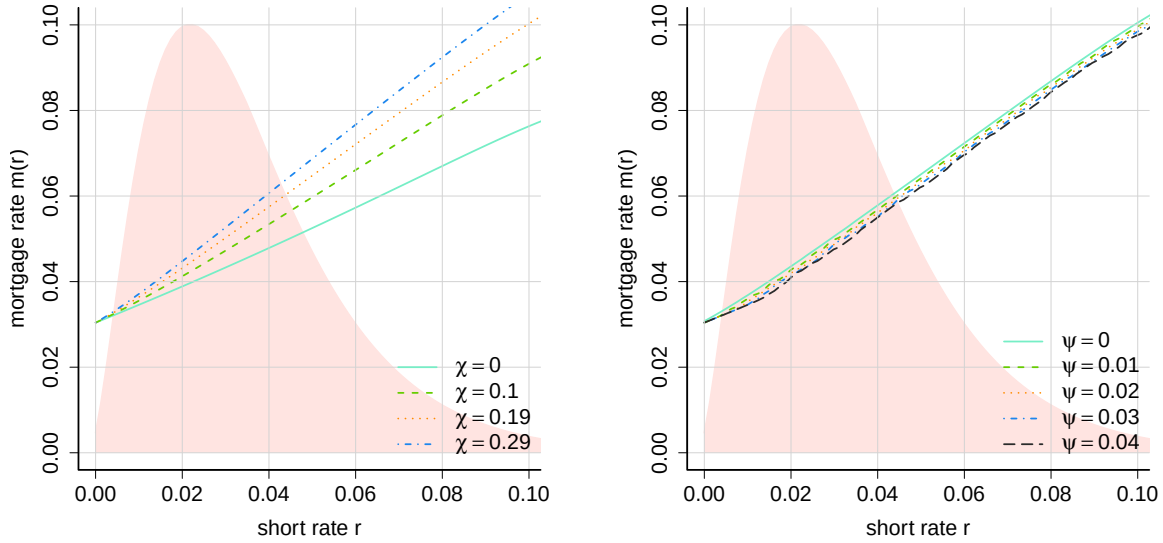

Figure 2: **Equilibrium homogenous-type mortgage rates vs. χ and ψ_χ.** Equilibrium mortgage rates m as a function of the short rate r. Left figure shows the sensitivity of the equilibrium mortgage rate to the attention rate χ when $\psi_\chi = 0$. Right figure shows the sensitivity of the equilibrium mortgage rate to the fixed-cost parameter ψ_χ when we set the attention rate χ to its estimated value 19% (see Section 6.2.2 and specifically Table 1). All other parameters are given in Table 2. The ergodic distribution of r is depicted as the pink area in the background.

When inattentive households bear upfront closing costs, we can study numerically the extent to which these costs influence equilibrium mortgage rates. The right hand side of Figure C.1 illustrates

this sensitivity for a range of realistic fixed-costs; going from $\psi_\chi = 0$ to $\psi_\chi = 2\%$ (of mortgage balance) causes equilibrium mortgage rates to decrease on average by 13bps.

Overall, Figure C.1 suggests that equilibrium mortgage rates have low sensitivity to upfront closing costs. Intuitively, while both fixed costs and inattention play a role, omitting fixed costs overestimates the occurrence of "small" refinancings (i.e. refinancings happening at small rate gaps), while omitting inattention underestimates the occurrence of "large" refinancings (i.e. refinancings happening at large rate gaps). In equilibrium, large refinancings have a much larger impact on investors' profits and losses – and thus on mortgage prices and rates – than the small ones. We have verified the robustness of this observation by studying various short rate processes and attention parameters.[17]

Thus, upfront costs have little quantitative effect on equilibrium interest rates in the environment with homogeneous attention. Perhaps even more importantly, upfront closing costs represent only a small fraction of total origination costs in practice: as documented in Zhang (2022), 80.5% of US households "roll" these costs into a higher mortgage coupon. Lenders' origination costs are then recouped via the gain on sale π. 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 upfront closing costs and then solving a much more complex option exercise problem.

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

Assumption 1. *Households do not face any upfront closing costs (i.e. $\psi_\chi = 0$).*

Assumption 1 will substantially simplify our numerical computations when we tackle the case with 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 we have studied until now with the heterogeneous environment we tackle next. When households are heterogeneous in their attention rate but investors can screen on χ, 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

[17] In more detail, keeping mortgage rates fixed, an increase in the upfront refinancing costs ψ_χ raises the rate gap threshold $\theta(x)$, dampening refinancing activity, the more so when interest rates are high, given the state-dependent nature of $\theta(x)$ documented in Appendix A.3. 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, the more so 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 are determined by the break-even condition (11). In other words, each type's pricing problem is isomorphic to the homogeneous pricing problem with that type's attention rate. Thus, we will sometimes refer to the MPE in the homogeneous case as the *Separating MPE*. Instead, when investors do not observe χ (i.e. in a "pooling" environment), significant complexities emerge.

4.2 Heterogeneous households

4.2.1 Infinite dimensional problem

Suppose now that there is a cross-section of attention types in the population. We assume that this ex-ante heterogeneity is permanent – in other words, the attention parameter χ is a constant attribute of a household. Let $H(\chi)$ denote the cumulative distribution over types (with associated density h). Crucially, we assume that investors – for whatever reason – do not or cannot screen on χ. We discuss this assumption, and relate it to institutional features of the US agency MBS market, in Section 4.3.2. One can (and we will) interpret this environment as if mortgages were traded in pools, formed at the time the mortgages are originated.

Let $F_t(c, \chi)$ be the joint cumulative distribution over outstanding coupon rates c and types χ 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-χ 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 exogeneous and follows a time-homogeneous Markov process. The density f_t, instead, evolves endogenously over time with households' refinancing decisions. The evolution of f_t is described by equations (B.9)-(B.10) in Appendix B.4.

Consider then the *shadow* price $P(S, c; \chi)$ of a mortgage with coupon c, conditional on knowing that the related household has attention rate χ. We will refer to $P(S, c; \chi)$ as a shadow price since investors do not observe χ, and thus cannot trade conditional on χ. $P(S, c; \chi)$ satisfies the

Feynman-Kac equation (B.11), described in Appendix B.4. This equation is the infinite dimensional counterpart to (10) when households are ex-ante heterogeneous and investors cannot observe χ.

Conditional on f_t and the current mortgage market interest rate m_t, the flow of households refinancing their mortgage between t and $t+dt$ has a type-distribution represented by the density

$$g_t(\chi) = \frac{\int_c \left(\nu + \chi \mathbb{1}_{\{c > m_t\}}\right) f_t\left(c, \chi\right) dc}{\int_\chi \int_c \left(\nu + \chi \mathbb{1}_{\{c > m_t\}}\right) f_t\left(c, \chi\right) dc d\chi}, \tag{12}$$

with corresponding cumulative distribution function $G_t(\chi)$. In order 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

$$g_t(\chi) = \frac{(\nu + \chi)h(\chi)}{\int_y (\nu + y)h(y)dy} = \left(\frac{\nu + \chi}{\nu + \bar{\chi}_H}\right) h(\chi), \tag{13}$$

with $\bar{\chi}_H := \mathbb{E}^H[\chi]$ the average degree of attention in the population. Next, consider the case where no-one's refinancing option is in the money at time t. The origination distribution g_t then coincides with the population distribution,

$$g_t(\chi) = h(\chi). \tag{14}$$

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, especially in low rate states.

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

$$\mathbb{E}^{G_t}\left[P\left(S_t, m(S_t); \chi\right)\right] := \int_\chi P\left(S_t, m(S_t); \chi\right) dG_t\left(\chi\right) = 1 + \pi, \tag{15}$$

subject to g_t given by (12), and where the superscript on the expectation indicates the distribution of household types χ 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 value function $V(S, c; \chi)$ defined via (1), (ii) the associated optimal refinancing policy satisfying (3), (iii) a shadow pricing function P defined via (9) and satisfying (B.11), (iv) a joint density*

f_t that satisfies (B.9)-(B.10)*, (v) a mortgage rate function $m(S)$ that satisfies* (15)*, with (vi) an origination distribution G that satisfies* (12)*.*

Such a Pooling MPE, which features an infinite dimensional state space, is reminiscent of problems in heterogeneous agents consumption-savings models in macroeconomics (see Krusell and Smith (1998), or more recently Ahn, Kaplan, Moll, Winberry, and Wolf (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 tackling 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. In order 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 either (i) a constant cross-section $G(\chi)$ or (ii) a state-dependent cross-section $G(\chi|x)$ at origination.*

Assumption 2 restricts the cross-sectional origination distribution used for pricing purposes to be at most dependent on the latent state x (in case (ii)), rather than on the full time-varying density f_t, and it will thus substantially reduce 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 prepayment, but do not fully consider how this prepayment behavior then affects the attention distribution of refinancers over time. The strength of Assumption 2 will depend 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 and the quantitative evaluation of our model, G will be either the (i) unconditional or (ii) conditional ergodic average origination distribution G_t,; this will insure that investors, while potentially making gains or losses upon their mortgage purchases at each point in time, break-even on average.[18]

[18] See Appendix B.9 for a discussion of why numerical methods such as those developed in Krusell and Smith (1998) are ill-suited for tackling our infinite dimensional problem.

4.2.3 Mortgage pricing in 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-χ mortgage. Rather than satisfying an infinite-dimensional Feynman-Kac equation, it now satisfies its finite-dimensional counterpart (10). Let $\bar{P}_G(x, c)$ be the expectation of $P(x, c; \chi)$ under the origination distribution G, and also the market price of a newly-issued mortgage pool:

$$\bar{P}_G(x, c) := \mathbb{E}^G[P(x, c; \chi)] \tag{16}$$

The counterpart to the market equilibrium condition (15), under Assumption 2, is then given by

$$\bar{P}_G(x, m(x)) = 1 + \pi \tag{17}$$

Finally, household's optimal refinancing behavior combined with the mortgage rate function $m(\cdot)$ imply 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

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

while the conditional origination distribution is given by

$$g(\chi|x) = \frac{h(\chi) \left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right)}{\int_\chi h(\chi) \left(\nu + \chi \int_{c \geq m(x)} f_\infty(c|x, \chi) dc \right) d\chi}. \tag{19}$$

Appendix B.8 provides additional details on our derivation. We use these distributions, in conjunction with Assumption 2, in order to build our approximate pooling Markov perfect equilibrium.

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) and satisfying (10), (iii) an ergodic joint density $f_\infty(x, c, \chi)$ and its corresponding ergodic marginal density over refinancers $g(\cdot)$ satisfying either consistency condition (18) (in the unconditional case), or (19) (in the conditional case), (iv) a newly-originated pool pricing function $\bar{P}_G$ defined via (16), and (v) a break-even condition (17).*

The Separating MPE and the Approximate Pooling MPE are similar in the fact 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_∞ over the latent state x, coupon c and inattention χ. Lastly, we introduce the concept of "monotonicty" of an equilibrium as follows.

Definition 4. *When x is uni-dimensional and $r(\cdot)$ is increasing, an Approximate Pooling MPE is said to be "monotone" if the related mortgage function $m(x, G)$ is increasing in x.*

The approximation imposed by Assumption 2 allows us to establish some useful theoretical results in addition to simplifying numerical calculations.

Proposition 6. *Let x be uni-dimensional and let $r(\cdot)$ be monotone increasing. Define the candidate mortgage rate*

$$m(x; G) := f + \frac{1 + \pi - \mathbb{E}^G [PO(x; \chi)]}{\mathbb{E}^G [IO(x; \chi)]}, \tag{20}$$

where G is a type distribution defined in (B.13) (unconditional) or (B.14) (conditional), where

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

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 $m(x; G)$ the equilibrium mortgage rate.

Our proof, detailed in Appendix B.5, 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 interest rate must satisfy (20). Conversely, if the object m, defined in Proposition 6, is monotone increasing in the latent state x, then an Approximate Pooling MPE exists and is unique. What are the properties of equilibrium mortgage rates in this environment with permanent attention heterogeneity? Our next proposition, proved in Appendix B.6, 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 7. *In a monotone unconditional Approximate Pooling MPE, the pool price $\bar{P}_G$ satisfies*

$$\bar{P}_G(x,c) = P(x,c;\bar{\chi}_G) - \mathbb{E}_x\left[\int_0^\tau e^{-\left(\int_0^t r(x_s)ds\right)} \mathbb{1}_{\{m(x_t)\leq c\}} \mathbb{C}\text{ov}^G(\chi, P(x_t,c;\chi))\, dt\right], \tag{21}$$

where τ is the (random) prepayment time for a household that has average attention $\bar{\chi}_G := \mathbb{E}^G[\chi]$.

Thus, the pool price $\bar{P}_G$ in the market dynamically behaves *as if* it were made up of homogeneous households with attention $\bar{\chi}_G$, but with an adjustment equal to the average (conditional on the rate gap being positive) discounted cross-sectional covariance $\mathbb{C}\text{ov}^G(\chi, P(x_t,c;\chi))$ between (a) shadow mortgage prices and (b) attention rates. If the shadow price P is "on average" decreasing in χ whenever the prepayment option is in the money, then 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 non-degenerate 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 (21) is indeed positive. Intuitively, keeping the average attention rate $\bar{\chi}_G$ constant, the faster households have a shorter effective maturity than the slower households. Investors are making money off the slower households while making losses off the 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 $\int_\chi P(x,c;\chi)dG(\chi)$. 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 8. *When a monotone Approximate Pooling MPE exists, denote $\underline{x}$ the lowest attainable latent state. The lowest mortgage rate $m(\underline{x})$ is invariant to the cross-sectional distribution G.*

Proposition 8 is proved in Appendix B.7. It relies on the observation that if a household has managed to lock-in the lowest attainable mortgage coupon $m(\underline{x})$, they will never refinance for strategic reasons, but 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 χ. The break-even condition at $x = \underline{x}$ allows us to conclude that $m(\underline{x})$ is invariant to G. Proposition 8 gives us some intuition for how the mortgage rate function $m(\cdot)$ might change, as the variance of the cross-sectional distribution G increases: we expect m to be relatively insensitive to the origination distribution G when rates are low, but substantially more in high interest rate states.

4.3 Discussion

In this section, we discuss several of our modeling assumptions, emphasize the externality that refinancing households impose onto inattentive households, and discuss the phenomenon of burnout and its relation to our model.

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. As such, 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 for the US mortgage market stems from empirical evidence and certain institutional features of the 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, Keys, Seru, and Vavra (2016) for additional evidence, for instance, 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) for instance document that between 2009 and 2014, only 20% of loans originated were kept on banks' balance-sheet. As of 2020, 70% of conforming mortgages were originated by speciality mortgage lenders, rather than deposit taking institutions; these specialty finance companies' sole objective is to originate conforming mortgages and immediately distribute them to investors via the agency MBS market ("originate to distribute"; see for instance Jiang (2019) or Buchak, Matvos, Piskorski, and Seru (2018)). In order to hedge their pipeline and sell

their origination book forward, these specialty finance companies make use of a specific derivatives market: the To-Be-Announced ("TBA") market. In that market, buyers do not know the exact mortgage pool they will receive at settlement, but instead only 5 characteristics of the pool: the agency (Fannie Mae or Freddie Mac), the average coupon rate, the maturity (typically 30-year or 15-year), the face value (i.e. the aggregate notional balance of the mortgages in the pool), 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.[19] These considerations rationalize the pooling equilibrium that we argue is relevant in the context of the US conforming mortgage market.

4.3.3 Externality

In equilibrium, the refinancing decision of a borrower paying attention today not only affects that borrower's payoff, but also that of currently inattentive borrowers via the equilibrium rate, creating an externality. In particular, an attentive borrower's decision to prepay and to exercise its refinancing option optimally (causing the lender to incur origination costs) pushes mortgage rates higher, thereby imposing a cost – higher payments – borne mostly by borrowers not currently able to act. Since borrowers do not internalize this cost, it leads to an externality that is purely "redistributive" when $\pi = 0$, but that instead leads to dead-weight losses when $\pi > 0$.[20] We illustrate this point precisely in Appendix B.10.

4.3.4 Burnout

Prepayment models used by the financial industry typically include, above and beyond the rate incentive, various factors that capture seasonality effects, seasoning (the fact that newly originated mortgages tend to have slower prepayment speeds than older but otherwise comparable mortgages), as well as burnout (the fact that pools with higher past cumulative prepayments tend to have slower

[19] See 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. One may wonder why lenders, when originating mortgages with "superior" prepayment characteristics, would want to use the TBA market – in which the cheapest securities are typically delivered – rather than the "specified pool" market when selling their production. The answer is two-fold. First, the TBA market is a forward market – in other words, it allows lenders to hedge their existing production without immediately selling their loans, whereas the specified pool market requires almost immediate settlement of a trade. This allows lenders to accumulate a portfolio of mortgages before selling into the agency MBS market. Second, the TBA market benefits from substantially more liquidity than the specified pool market, as discussed in Bessembinder, Maxwell, and Venkataraman (2013) or Gao, Schultz, and Song (2017).

[20] This type of externality also arises in dynamic debt run models, as in He and Xiong (2012).

prepayment speeds than otherwise comparable pools with lower past cumulative prepayments).[21] Our model with heterogeneous attention rates endogeneously produces the burnout effect: in a given mortgage pool, faster households will tend to refinance earlier and thus leave the pool faster than slower households, causing a decrease in prepayment speeds as the pool factor declines. This effect was also noted by Stanton (1995) in a model where households have heterogeneous fixed costs of prepaying.

4.4 Redistribution and cross-subsidies in the mortgage market

In this section, we consider the distributional effects of the pooling equilibrium on different types of households. In what follows, we will denote $m(x, G)$ the equilibrium mortgage rate in the Approximate Pooling MPE with origination distribution G, and (with a slight abuse of notation) $m(x, \chi)$ the corresponding mortgage rate in the Separating MPE for type χ households.

Competition implies that investors break-even when they purchase new mortgages, so in a separating equilibrium (or equivalently in an environment with no heterogeneity in attention) investors do not expect to make or lose any money on any of the mortgages they fund.[22] Such an environment thus rules out cross-subsidies amongst households. However, as discussed in Section 4.3.2, institutional features of the US conforming mortgage market imply that there is pooling for a large set of households. Pooling in turn results in cross-subsidies, as investors only break-even on average, as $\bar{P}_G(x, m(x)) = \int P(x, m(x); \chi) dG(\chi) = 1 + \pi$ does not imply $P(x, m(x); \chi) = 1 + \pi$ unless the short rate is at its lower bound, as discussed in Proposition 8.

We focus on three different measures of redistribution. First, we consider the difference in the annualized rate households of type χ would pay in the Separating MPE versus the rate they pay in the Approximate Pooling MPE:

$$\Delta m(x; \chi) := m(x, \chi) - m(x, G). \tag{22}$$

When this difference is positive, type-χ households effectively receive a subsidy from the market, and it coincides with investors losing money on this specific mortgage, $P(x, m(x, G); \chi) < 1 + \pi$.[23]

[21] See for instance Spahr and Sunderman (1992).

[22] Of course, they may realize ex-post profits or losses on any individual mortgage depending on the realization of idiosyncratic attention and movements in r.

[23] Alternatively, assume that household type χ 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 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.

However, the investor's shadow price for a type-χ mortgage provides only a partial picture of cross-subsidies in the model, since attention is a permanent attribute of households. Indeed, households with different attention rates have different effective maturities of their current mortgages, and they will benefit (if they are a "fast" type) or get hurt (if they are a "slow" type) by the difference between $P(x, m(x); \chi)$ and the initial mortgage price $1 + \pi$ not just in the current mortgage but also every time they refinance again in the future. We thus consider, as an alternative measure of redistribution, the valuation of all future liabilities for a type-χ household at origination relative to the population average:

$$\Delta V(x; \chi) := \int V(x, m(x); \chi) h(\chi) d\chi - V(x, m(x); \chi). \tag{23}$$

When $\Delta V(x; \chi) > 0$, the households' personal valuation of its future mortgage liability is lower than that of the average household.

Third, and most simply, we compute the ergodic average coupon paid by each household type:

$$c_\infty(\chi) := \int_x \int_c c \cdot f_\infty(x, c|\chi) dc dx, \tag{24}$$

where $f_\infty(\cdot, \cdot|\chi)$ is the joint ergodic density over coupons and the latent state vector x for households of type χ, given the prevailing mortgage market interest rate $m(x)$ determined in our Approximate Pooling MPE. These three measures of redistribution will be considered when we study various policy proposals and perform counterfactual calculations, as described next.

5 Policy evaluations and counterfactual calculations

We now turn to various policy proposals put forth in the academic literature and in policy circles, primarily aimed at improving mortgage borrowers' welfare via the reduction of 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 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 (thereafter, "auto-RM"), as suggested for instance by Keys, Pope, and Pope (2016) or Campbell, Jackson, Madrian, and Tufano (2011). With an auto-RM, the household pays the minimum realized mortgage rate since the mortgage's inception at time τ:

$$\underline{m}_t \equiv \min_{\tau \leq s \leq t} \{m_s\}. \tag{25}$$

The contractual rate of such a product is thus tied to the minimum process of the mortgage market interest rate. On the face of it, the auto-RM seems like a great idea for inattentive or financially unsophisticated households, since it reduces the impact of financial mistakes on these households' lifetime mortgage liabilities' cost.[24] However, discussions around this proposal are usually cast in *partial equilibrium*, and thus fail to take into account the *general equilibrium* response of mortgage rates. Our model can speak to this response.

In order 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 as in adjustable rate mortgages. However, unlike in 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 only incurred at the time households move, repay 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] We relegate all technical details to Appendix C.1.

We make four observations about this environment. First, even though households may still have heterogeneous χ, 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 that features no cross-subsidies.

[24] See for instance Agarwal, Rosen, and Yao (2016) for a quantification of the life-time cost of these mistakes.

[25] Under Assumption 3, $m(x, \infty)$ is the limit of the Separating MPE's mortgage market interest rate $m(x, \chi)$ as $\chi \to +\infty$ when we set the gain on sale $\pi = 0$, and thus we use this notation also 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 Appendix B.10.

Second, the effective mortgage duration ceases to be state dependent, since the moving intensity in our model is assumed orthogonal to household type and rate and coupon environment.[26] All mortgages therefore have the same expected duration $1/\nu$.

Third, traditional fixed-rate prepayable mortgages we have thus far considered trigger origination costs when refinancing that are recovered by lenders via a combination of (i) upfront closing costs ψ_χ paid by borrowers, and (ii) the gain on sale π 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.

Fourth, since mortgage market investors continually receive coupons according to the minimum realized rate, borrowers almost always "underpay" for mortgage financing, relative to the case where they are getting funds via a more traditional adjustable rate mortgage. Lenders optimally charge a rate $m(x,\infty) \geq r(x)+f$ in order 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 Appendix C.2:

Proposition 9. *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 χ. Initially, all households have a traditional fixed-rate prepayable mortgage. In the resulting Pooling MPE, at the time of a refinancing, the slowest households over-pay 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, pushing further up traditional mortgage rates and skewing the attention distribution for 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:

[26] Our assumption that moving rates are independent of coupon and rate environments is not entirely supported by the data; Berger et al. (2021) for instance shows that prepayment hazards related to moves are increasing functions of the rate gap – the difference between the contractual coupon on a mortgage and the current market interest rate.

[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.

Proposition 10. *With heterogeneous attention rates and the ability for households to either use (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 their expected net present value 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 that are too high for their target home. In other words, the disutility of having 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 its high cash-flow volatility, compared to a more traditional mortgage. The welfare impact of this type of contract, in the presence of risk-averse households, will depend on the co-movement of income with rates.

5.2 Improving financial literacy

Next, consider the cross-sectional attention distribution H. This distribution is a stand-in for a range of frictions, one of which is financial literacy, as discussed in Section 3.2. The large cross-subsidies from financially unsophisticated to sophisticated households suggests that policies raising financial literacy may appear attractive to policy makers interested in reducing inequality.

Through the lens of our model, raising financial literacy can be viewed as a shift in the cross-sectional distribution from H to some $\tilde{H}$ with $\tilde{H} \succeq H$, where $\succeq$ indicates *first-order stochastic dominance* (thereafter, "FOSD"). Once again, while such goal seems like a welfare-improving initiative, it might have unintended negative consequences for some households, as we argue next.

Consider an attention distribution H on some support $\mathcal{X} = [\chi_\ell, \chi_h]$, and consider educating a set of households $X \subset \mathcal{X}$ so that their type monotonically increases to some subset $X' \subset \mathcal{X}$ for a

[28] Absent any required gain on sale ($\pi = 0$), $m(r, G) < m(r, \infty)$ is a natural outcome. Instead, 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.

new distribution that satisfies $\tilde{H} \succeq H$. Then, everyone not in the improved set, i.e., $\mathcal{X} \setminus X$, will be *unambiguously* made worse off, as the pool is more attentive on average and thus $m(x, \tilde{H})$ increases for each latent state x, while their type stayed constant. Further, while those in the improved set have higher financial literacy and exercise their refinancing options more efficiently, they now face greater mortgage interest rates; and so it is a quantitative questions whether or not they are uniformly better off, which we turn to below.

5.3 Fintech and non-bank lenders' nudging

A shift in the distribution H can also be viewed as the model counterpart to recent trends in mortgage origination and servicing that have been observed over the past decade in the US mortgage market. Buchak et al. (2018) for instance document a significant increase in the share of all mortgages originated by speciality finance companies, reaching 50% in 2015. Financial technology lenders' (thereafter, "FinTech firms") market share has also been rising during this time period, accounting for 8% of total US mortgage issuance as of 2016, as documented by Fuster et al. (2019). Non-bank and FinTech lenders tend to nudge borrowers to encourage refinancings, potentially significantly improving households' effective attention rate, with an impact on equilibrium mortgage rates that can be quantified, through the lens of our model, via a rightward shift in the cross-sectional distribution H.

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 mortgage borrowers, which is necessary for quantifying the equilibrium consequences of the mortgage reforms discussed in the previous section.

6.1 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.[29] 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 will 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.

6.2 Disentangling sample noise from heterogeneity

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 θ bps above the effective mortgage rate that month, i.e., $gap > \theta$.[30] 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 probability of attention and thus refinancing in a given effective month equal to

$$p_i = p(\chi_i) := 1 - \exp\left(-\chi_i/12\right)$$

Given the "Calvo" assumption implicit in our modeling of household inattention, for a given t_i and χ_i, the realization of s_i is i.i.d.[31] The maximum likelihood estimate (thereafter, "MLE") for an

[29] Details of our data sample and our calculations are disclosed in greater details in the Online Appendix.

[30] We reintroduce the threshold θ 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 over-estimate the number of "small" rate refinancings, and these refinancings do not alter the equilibrium pricing function substantially. However, when *measuring* the distribution of attention rates, getting small refinancings wrong does impact the estimated attention distribution. Under our assumption of state-independent attention rate, in the absence of upfront closing costs, any threshold $\theta > 0$ results in a consistent estimation of attention rates. However, in the presence of upfront closing costs $\psi_\chi > 0$, a balance has to be struck to estimate attention rates – high θ yields more consistent attention rates but rapidly reduces sample size, while low θ underestimates attention rates by attributing inaction in periods of small positive rate gaps to inattention instead of fixed cost.

[31] 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 $\{m_t\}_{t\geq 0}$, households with higher attention rate χ will have on average a lower number of "effective periods" t_i than households with lower χ. In other words, even

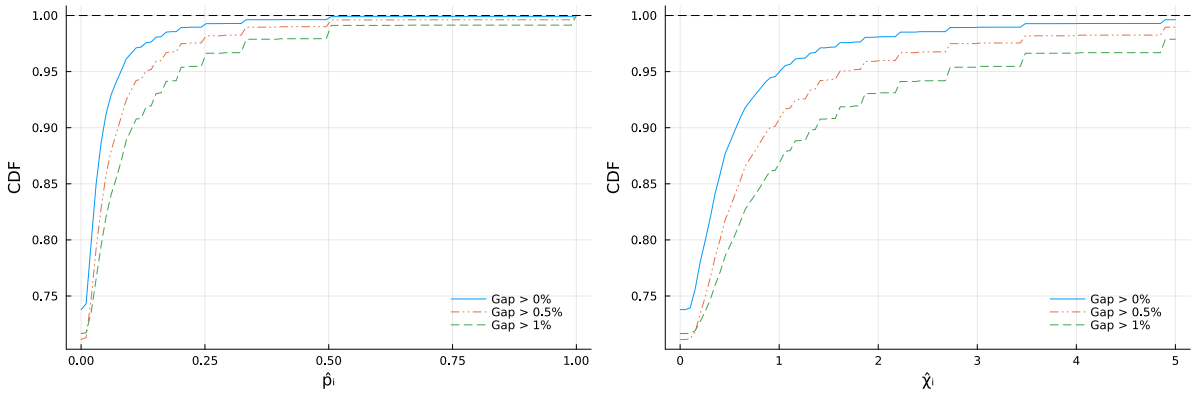

Figure 3: **Raw cross-sectional attention distribution.** Left panel shows the distribution of $\hat{p}_i = [\text{nr refi}]_i/[\text{nr effective periods}]_i$ for loan-amount-weighted observations; right panel shows the corresponding distribution for weighted χ_i

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 (thereafter, "CDF") of refinancing propensities $\hat{p}_i = s_i/t_i$ for different rate gap cutoffs, all weighted by household's 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 who 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.

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

6.2.1 Testing for a unique homogeneous group

To do so, we first check whether our refinancing data contains *some* heterogeneity in attention, by testing whether it could have been generated by a homogeneous group of households with common attention rate χ. We perform a 2-sided Kolmogorov-Smirnov test between the empirical distribution of $\hat{p}_i$ and a simulated distribution of $\hat{p}_i^{sim}$. This simulated distribution utilizes the maximum

though we treat the number of "effective periods" t_i as exogenous for our estimation of χ, within the model those two are linked. Later on, our GMM estimation of the attention distribution will address this shortcoming rigorously.

likelihood estimator for χ under the assumption that the refinancing data is 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 it by repeatedly sampling the binomial distribution with probability $\hat{p}_{N=1}$ for a given vector of t_i's.[32] The Kolmogorov-Smirnoff 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.

6.2.2 Measuring the distribution over permanent heterogeneity $H(\chi)$

Next, we use two distinct approaches relying on MLE in order to quantify the degree of cross-sectional attention heterogeneity in the data. Importantly, both approaches allow us to explicitly account for sampling noise by using all the information in an observation (s_i, t_i), rather than only relying on the ratio $\hat{p}_i = s_i/t_i$. In particular, our estimation treats an observation of $s_i = 1$ refinancing success out of $t_i = 2$ trials differently than an observation of $s_i = 10$ refinancing successes out of $t_i = 20$ trials, even though they both result in the same observed $\hat{p} = 1/2$.[33]

Our first approach relies on a clustering algorithm. For a given number N of homogeneous groups, we use ML to estimate the group-specific rates χ_k for $k \in \{1, ..., N\}$, and we allocate each individual i into a group k to derive the estimated weight of each group.[34] We choose $N = 5$ and $gap > 0.5\%$, and in Appendix E.2 we verify the robustness of our conclusions to alternative choices. Table 1 displays the results of our estimation. 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 – roughly 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.

[32] We numerically derive the theoretical distribution of $\hat{p} = s/t$ given t by 10,000 times sampling successes 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 $\hat{p}^{sim}$.

[33] Note that our modeling of the inattention friction does not allow for "gap-dependent" attention rates $\chi(y)$ where y is the mortgage rate gap. Though convenient for computational purposes, the main impetus for modeling attention this way is that our CRISM panel data is not as good as the headline number of households in the sample suggests. Estimating "gap-dependent" attention rates would be extremely noisy since (i) most households have relatively low refinancing rates and (ii) most households do not have a high number of "effective periods".

[34] The optimal number N of groups is an open question, and for the moment we abstract from it. Importantly, as $N \to N_{HH}$, our approach converges to the realized CDF of attention probabilities $\hat{p}_i$.

χ	$p(\chi)$	Std Error p	$H(\chi)$	$G(\chi)$
0.0	0.0	0.0007	0.811	0.619
0.2343	0.0193	0.0002	0.061	0.087
0.5627	0.0458	0.0004	0.07	0.131
1.3882	0.1092	0.001	0.043	0.109
5.2775	0.3558	0.005	0.014	0.054

Table 1: **Estimation of population distribution $H(\chi)$ and origination distribution $G(\chi)$.** Estimation assumes $N = 5$ homogeneous groups, and focuses on households and months with $gap > 0.5\%$, weighted by average loan amount. The average attention rate in our sample is $\bar{\chi}_H = 19\%$, while the average attention rate *for refinancers* is $\bar{\chi}_G = 53\%$.

As an alternative to the clustering approach above, we introduce a flexible parametric distribution $f(\cdot;\theta)$ for χ or $p(\chi)$, and maximize the resulting likelihood w.r.t. θ. To facilitate our numerical computations, we use a distribution for p (instead of χ) that results in an analytical likelihood function: the Beta distribution, with parameter vector $\theta = (\alpha, \beta)$ and density

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}. \tag{26}$$

As shown in Appendix E, this specification nests the exponential distribution (with parameter $\hat{\theta}$) over the attention parameter χ, when $(\alpha,\beta) = \left(1, \hat{\theta}/dt\right)$ with dt being the period length (i.e. $dt = 1/12$). Our unconstrained MLE yields a point estimate $(\hat{\alpha},\hat{\beta}) = (1.229, 79.550)$, with a corresponding mean monthly refinancing probability of $\mathbb{E}[p] = 1.52\%$. When imposing an exponential distribution for χ, i.e., $\alpha = 1$, we instead obtain a point estimate of $(\alpha,\hat{\beta}) = (1, 120.025)$ which yields $\hat{\theta} = 10.002$, and a mean monthly refinancing probability of $\mathbb{E}[p] = .83\%$.

6.3 Deriving the origination distribution $G(\cdot)$

Since CRISM follows borrowers throughout the sample period regardless of their refinancing activity, the attention distribution estimated in Table 1 corresponds to the distribution $H(\chi)$ of permanent heterogeneity in the population. However, only the attention distribution $G_t(\chi)$ of *refinancing households* is relevant for pricing purposes. To reduce the complexity of the infinite dimensional problem discussed in Section 4.2.1, we use Assumption 2 and approximate the origination distribution $G_t(\chi)$ by its ergodic values, effectively averaging out the path-dependence of G_t. Our algorithm for constructing G is described in details in Appendix B.8.

With H estimated using our CRISM data and displayed in Table 1, we derive G under the assumption that interest rates and other model parameters are those used in Section 7 and described in Table 2, with the interest rate being uni-dimensional $r(x) = x$. 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). Figure 4 summarizes our results. The left panel shows the ergodic unconditional origination distribution $G(\chi)$ that we will be using for our quantitative analysis in Section 7, while the right panel shows the corresponding *conditional* distribution $G(\chi|r)$, which we will be using to test the robustness of our conclusions to Assumption 2. Both origination distributions over-represent high-χ types, and under-represent low-χ types, relative to the population distribution H, as discussed in Section 4.2.1. This observation holds for the conditional distribution $G(\chi|r)$ for all but the highest interest rate state, at which point no one voluntarily refinances. The state $r = \bar{r}$ corresponds to the situation described by (14); the ergodic average conditional distribution $G(\chi|\bar{r})$ is thus equal to the distribution $H(\chi)$ of permanent heterogeneity. 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 distortion for the unconditional origination distribution $G(\chi)$, while not as significant, remains sizeable, as shown in the left panel of Figure 4.

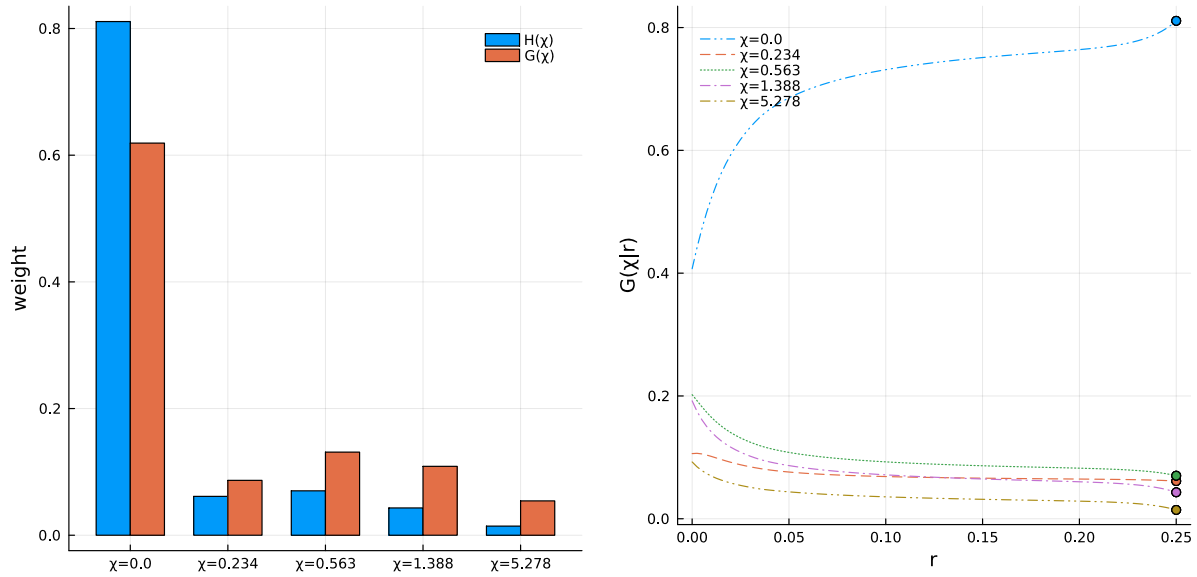

Figure 4: **Ergodic origination distribution $G(\cdot)$ implied by $H(\cdot)$.** Left panel shows the unconditional population distribution $H(\chi)$ (left blue bars) and the unconditional origination distribution $G(\chi)$ (right orange bars). Right panel shows the conditional origination distribution $G(\chi|r)$ (solid lines) and the unconditional population distribution $H(\chi)$ (thick dots).

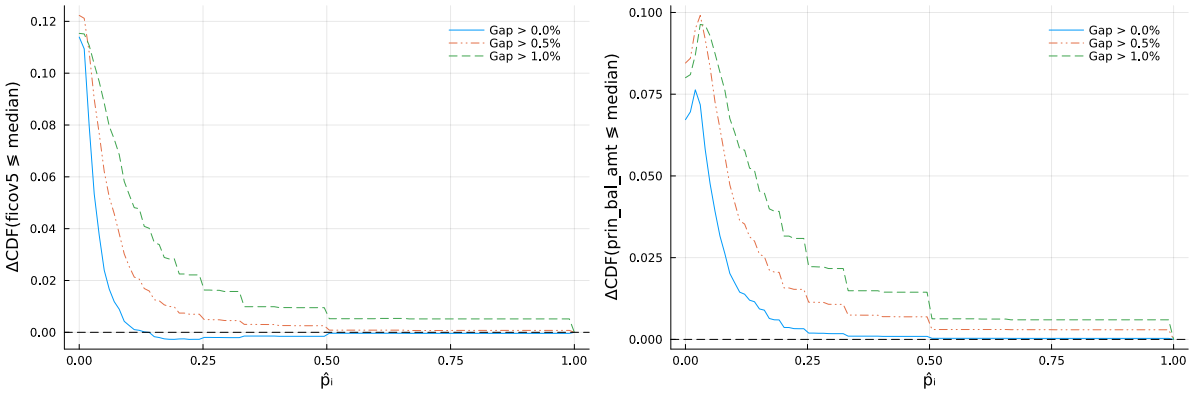

Figure 5: **Weighted cross-sectional distribution splits.** CDF differences of $\hat{p}_i$ weighted by average loan balance according to sample splits. Left panel splits by median FICO score, while right panel splits by median loan amount.

6.4 Household sub-samples

So far we have documented substantial heterogeneity in the degree of household attention in mortgage refinancing decisions in the population. We now investigate which observable household characteristics are related to attention.

To begin with, consider Figure 5, 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.[35] This potentially reflects a greater degree of financial frictions faced by these latter households.

In the right panel of Figure 5, 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 χ_i.

Using an approach identical to that described in Section 6.2.2 except that we constrain the estimation to the group means $\boldsymbol{\chi} = \{\chi_k\}_{k=1,\ldots,N}$ from the baseline procedure, we estimate the attention distributions $H_{split}(\chi)$ for each of our sub-samples, compute the corresponding origination distribution $G_{split}(\chi)$, and summarize our results in Table E.3.

7 Quantitative implications

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

7.1 Estimation and calibration of remaining model parameters

We use the cross-sectional attention distribution estimated non-parametrically in Section 6.2.2 with $N = 5$ groups, as displayed in Table 1, and subsamples according to above- and below-median FICO and loan amounts, as displayed in Table E.3. The short term interest rate r_t follows a one-factor, square root diffusion process as in Cox, Ingersoll Jr, and Ross (1985). In other words, $r(x) = x$, $\mu(x) = \kappa(\mu - x)$ and $\sigma(x) = \sigma\sqrt{x}$. We take as the relevant short-rate the 3-months treasury rate, and estimate the parameters of our model (the long run mean μ, the speed of mean reversion κ,

[35] 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 in slightly negative.

and the volatility parameter σ) via MLE on a sample from 1971 to 2021.

The parameter ν can be interpreted as the sum of a moving intensity, default intensity, and amortization 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 $\nu_{mat} = 3.3\%$. Given the negligible default experience in US agency mortgages post financial crisis, we set the default intensity $\nu_{def} = 0\%$. Thus, the "forced" prepayment intensity is assumed to be $\nu = \nu_{mov} + \nu_{mat} + \nu_{def} = 7.4\%$.

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 2019 FHFA report on guarantee fees).[36] Lastly, since we assume no closing costs borne by the household $\psi_\chi = 0$, since 80% of US households finance such closing costs via higher rates, and since the average cost of mortgage intermediation is 4.6% (see Zhang (2022)), we set the gain on sale to $\pi = 80\% \times 4.6\% = 3.68\%$. We solve our model using a standard finite difference method, as documented in greater detail in Appendix D.

Parameter	Value	Interpretation
μ	0.035	Long-run short rate mean
κ	0.13	Mean-reversion coefficient
σ	0.06	Volatility
ν	0.074	Total unconditional prepay rate
f	0.0045	Ongoing portion of G-fees
π	0.0368	Gain on sale

Table 2: Baseline parameter values

7.2 Validating of our equilibrium mortgage pricing

We first confront our model-implied mortgage rate function $m(x, G)$ to 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

[36] In principle, the agency MBS coupon is lower than the average loan pool interest rate not only due to the ongoing portion of the G-fees, but also due to the 25bps base servicing fee. However, this servicing fee belongs to the originator, who can monetize it by selling the "mortgage servicing rights" (or "MSR"). In our quantitative application, the market value of the MSR is included in the gain on sale π realized by mortgage lenders when selling a loan 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 similar discussion on the decomposition of the gain on sale into a TBA premium and the MSR value.

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, retrieve the implied short term interest rate $r(x_t)$, which we then use in order to compute the relevant model-implied mortgage rate $m_t = m(x_t, G)$.[37] In Figure 6, we plot the time series of model-implied mortgage rates, as well as 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 is not only highly-correlated with its data counterpart, but it also has a time-series average that is only slightly below what we measure in the data – with a difference of 17bps p.a., which can be attributed to our assumption of a zero "option adjusted spread".[38] The only notable difference between our model-implied rates and their empirical counterpart occurs during the financial crisis period of 2008 and early 2009. Stress in financial markets and questions surrounding the implicit government backing of the GSEs resulted in very wide mortgage spreads at that time. 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 is a good fit to actual mortgage data given the simple one-factor term structure interest rate process we assume.

[37] We choose our benchmark interest rate to be the Libor swap curve since agency MBS trade at an option adjust spread ("OAS") 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 35bps to the 10 year constant maturity treasury rate reported by the St Louis Fed, where the 35bps corresponds to the average 10 year swap spread over the sample period 2000-2017.

[38] See Boyarchenko, Fuster, and Lucca (2019) for an extensive discussion of the concept of "option adjusted spread" in the agency MBS market.

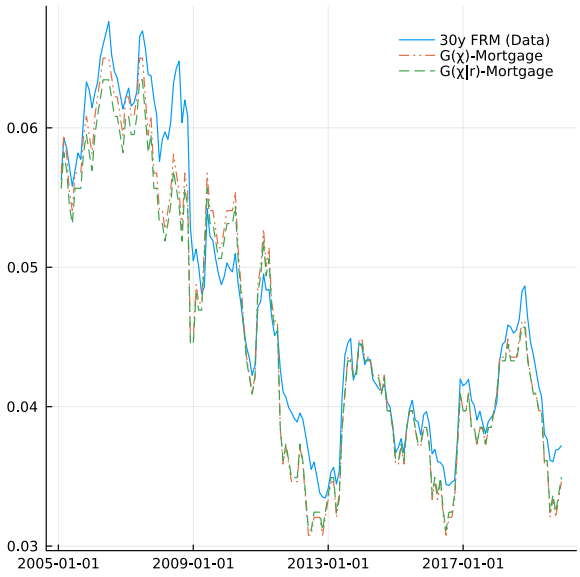

Figure 6: **Pricing time-series 2005-02 to 2020-01.** The blue line represents Freddie Mac's 30y primary mortgage market survey rate, while the red dot-dashed and green dashed line represent the model-implied mortgage rate for unconditional and conditional pricing respectively.

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 7 shows the mortgage function $m(r, G)$ given our estimated heterogeneity in Table 1 (solid black line), and the counterfactual mortgage function $m(r, \bar{\chi}_G)$ in the hypothetical environment where all households are homogeneous in their attention rate (dashed red line). Both mortgage functions are strictly increasing in the short rate, with the same value at $r = 0$, but with $m(\cdot, \bar{\chi}_G)$ lying above $m(\cdot, G)$ otherwise. The ergodic average difference in mortgage rates is approximately 120bps p.a. in our baseline, highlighting the non-trivial impact of cross-sectional heterogeneity in attention on the *level* of mortgage market rates.

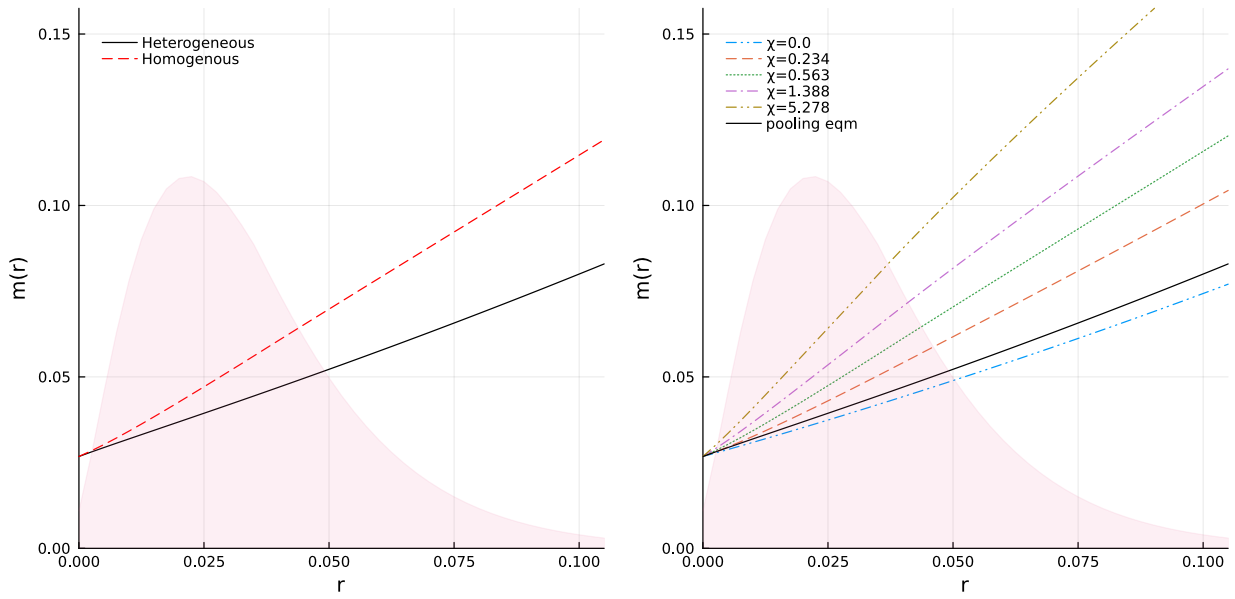

Figure 7: **Equilibrium mortgage rates.** Left panel shows equilibrium mortgage rate for (i) the Approximate Pooling MPE (black line) and (ii) the MPE with a homogeneous group of households with attention rate $\bar{\chi}_G = 53.14\%$ (red dashed line); right panel shows the mortgage rates for each type χ in a Separating MPE with ex-ante heterogeneous households, and the solid black line as the Approximate Pooling MPE. The pink curve in the background depicts the ergodic distribution of the short rate r_t.

Next, the right panel of Figure 7 shows the corresponding Separating MPE mortgage functions $m(r, \chi_i)$ for all groups $i \in \{1, ..., N\}$ (non-solid lines), alongside the Approximate Pooling MPE mortgage function $m(r, G)$ (black solid line). Faster households receive worse mortgage rates in a Separating MPE, i.e., $m(r, \chi) > m(r, \chi')$ for $\chi > \chi'$, as proved in Proposition 5. In the Separating MPE, the fastest households would face mortgage rates with an ergodic average that is around 380bps p.a. higher than that of the slowest households.

The right panel of Figure 7 can also be used to back out the rate subsidies in the Pooling MPE vs. the Separating MPE, i.e., $m(r, \chi) - m(r, G)$. The fastest households for instance would face mortgage rates in a Separating MPE that are 350bps p.a. higher on average than those they actually face in the Pooling MPE. These subsidies are asymmetric: everyone but the slowest group tends to receive a subsidy, while the slowest group is paying those subsidies. Investors tend to lose money at origination on high-χ households – i.e. $P(r, m(r); \chi) < 1 + \pi$ for those household types – and make money on low-χ households. However, as discussed in Section 4.4, these investors' profits/losses at origination are an incomplete measure of redistribution, given that households

refinance their mortgage at different frequencies, depending on their type. To account for this, we show in the left panel of Figure 8 the household value functions for the different attention rates χ relative to the average household value, $\Delta V(r;\chi)$, defined in (23). This latter measure suggests a large degree of redistribution across households: those in the fastest group, on average, make future payments that amount to a valuation that is 14.5% (of the mortgage balance) lower than households in the slowest group.

Lastly, the right panel of Figure 8 depicts the ergodic average mortgage coupon by type in the Approximate Pooling MPE (blue dots), in the Separating MPE (cross), and in the corresponding MPE with homogeneous households with attention rate $\bar{\chi}_G$ (red dot). Ergodic average coupons in the Approximate Pooling MPE are decreasing as attention χ increases, with the fastest group paying on average 90bps less than the slowest group. Importantly, this difference purely stems from the ex-post difference in household refinancing rates, rather than from general equilibrium forces. We also notice that for all groups, the ergodic average coupon paid in the Approximate Pooling MPE remains lower than the corresponding average if households were homogeneous with attention rate $\bar{\chi}_G$ – in that latter equilibrium, mortgage rates are substantially higher, as previously discussed. In the corresponding Separating MPE, the ergodic coupon is upward sloping as a function of the attention rate χ, mainly due to the need for mortgage lenders to recoup origination costs by selling new mortgages so as to realize the gain on sale π. Indeed, since faster households impose more frequent origination costs and since mortgage originators must break-even, they need to charge higher rates. This outcome is in stark contrast with the hypothetical scenario of no origination costs (i.e. $\psi_\chi = \pi = 0$), in which case the ergodic average coupon in the Separating MPE is relatively insensitive to the level of attention as shown in Appendix F.[39]

[39]The only source of sensitivity of the ergodic average coupon to the level of attention in the case $\pi = 0$ stems from discount rate effects.

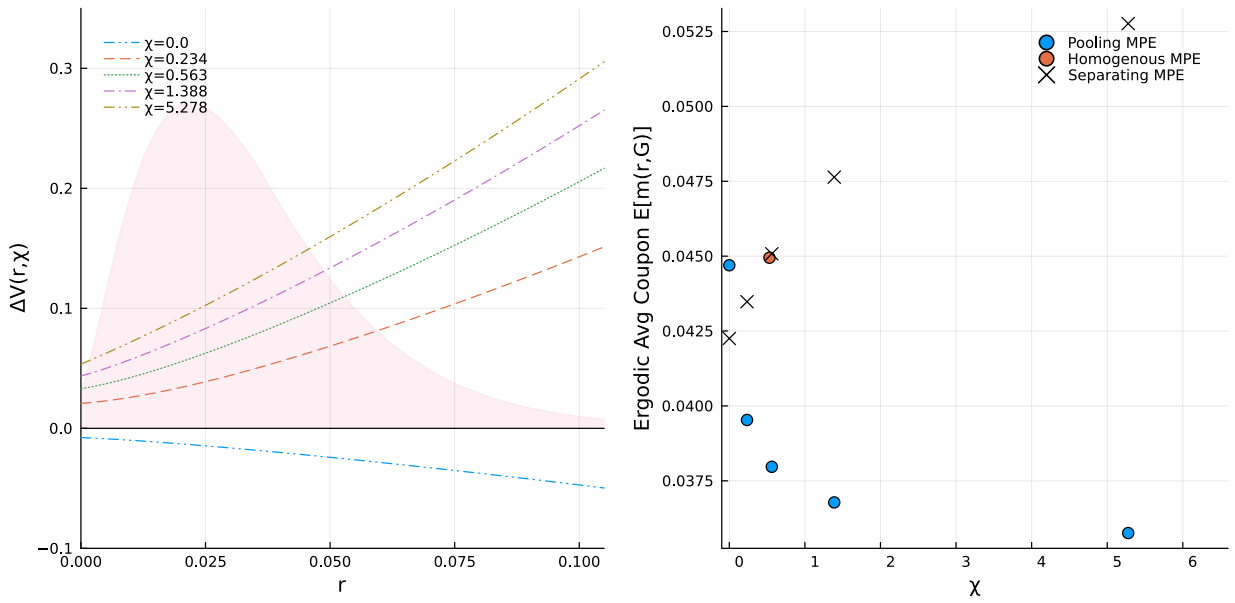

Figure 8: **Cross-subsidies and ergodic average rates by type.** Left panel shows $\Delta V(r;\chi)$ – i.e. the difference between (a) households' cross-sectional average value function and (b) type-χ household's value function. Right panel shows the ergodic average coupon paid by each type-χ household in the Approximate Pooling MPE by type (blue dots), in the homogenous MPE (red dot), and in the Separating MPEs by type (cross).

The differences in both panels of Figure 8 are substantial; however, as a household's attention rate is not observable, these differences do not speak directly to the degree of redistribution amongst households of different *observable* characteristics. Figure 9 helps us address this question, by showing (blue line) the average value function difference $\Delta V(r;\chi)$ between households with FICO scores above and below the population median, leveraging our MLE procedure of Section 6.2.2 and the allocation of each individual in our sample into one specific attention type χ_i according to Table E.3. 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 3. To judge the robustness of the ZIP code level variables, we calculate the ZIP code level average FICO score from the household level observations.[40] We see that household level observations give about 20% stronger results than the ZIP code level observations. Table 3 suggests that the cross-

[40]Household ZIP code level covariate is the average of the household's time series over the relevant zip code value.

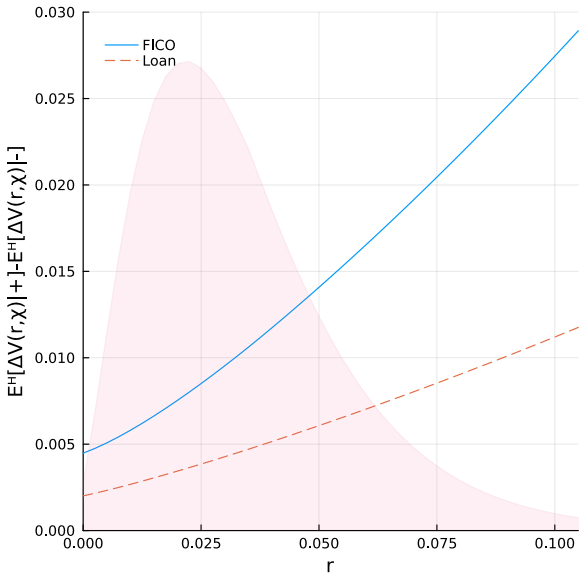

Figure 9: **Cross-subsidies by above/below median sample splits:** The blue line shows the household's lifetime value function difference between being an above median FICO household (FICO+) vs. a below median FICO household (FICO-); the red dashed line shows the difference between an above median vs. 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 3: Ergodic differences by above/below median sample splits (in bps)

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.

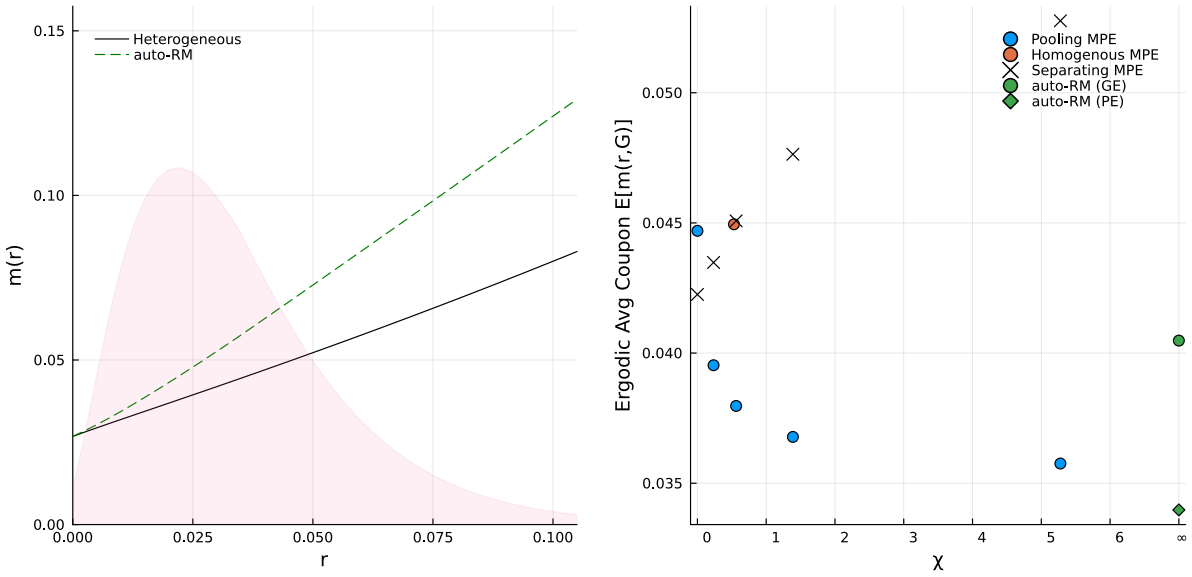

Figure 10: **Automatic refinancing mortgages.** Left panel shows pricing of the auto-RM (teal), and the Approximate Pooling MPE (black) graphed against the ergodic density of r. Right panel shows the ergodic average coupon in the Approximate Pooling MPE (blue dots) and the Separating MPE (cross) per type, as well as the MPE with homogeneous households (red dot), and automatic refinancing mortgage in GE (green dot) and in PE with unchanged pricing from baseline (green diamond).

7.4 Evaluating policy proposals via model counterfactuals

In this section, we quantitatively evaluate the policy proposals discussed in Section 5. We first focus on automatically refinancing mortgages, then turn towards the benefits of financial literacy interventions, and conclude with a quantification of the impact of the rise of non-bank and FinTech mortgage lenders onto 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 left panel of Figure 10 depicts the equilibrium rate of an Auto-RM as the dashed green line. As stated in Proposition 9, 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 prevalent equilibrium rate in the Approximate Pooling MPE (black line in Figure 10), i.e. $m(r, \infty) \geq m(r, G)$; the ergodic average difference between the two rates is 130bps, highlighting the substantial increase in initial

rates when moving 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 Appendix C.3 the debt-to-income (thereafter, "DTI") distribution at origination observed in our SFLP data-set, vs. the counterfactual DTI distribution that would be prevalent in a world where households only had access to the Auto-RM product. Focusing on the 43% DTI cutoff – the limit below which mortgages, until 2021, satisfy the "Qualified Mortgage" definition of the Consumer Financial Protection Bureau – approximately 20% of borrowers would be pushed above such DTI cutoff, potentially forcing them to downsize their house or increase their downpayment upon purchase.

Next, the right panel of Figure 10 depicts the ergodic average coupons. Note that even though initial rates are higher, i.e., $m(r, \infty) \geq m(r, G)$, the ergodic average coupon paid by households in an Auto-RM (green dot) is 50bps lower than that in the MPE with homogeneous household with average attention rate $\bar{\chi}_G$ (red dot). However, it is not necessarily lower than the ergodic average coupon across types in the Approximate Pooling MPE; in other words, the cross-subsidy enjoyed by fast households in the Approximate Pooling MPE more than overcomes their suboptimal refinancing behavior. In a partial equilibrium setting, in which we would hold $m(r, G)$ constant at its baseline level, the ergodic average coupon for households able to entirely overcome their inattention friction would be 3.4% (green diamond) – i.e. 60bps 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 when considering alternative contract designs: while some of these designs might appear to be substantially beneficial to households in partial equilibrium, their effect can be considerably dampened by general 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 of Proposition 10, while only the slowest households benefit. Lastly, the ergodic average mortgage rate prevalent in the homogeneous MPE is 60bps p.a. above that in the auto-RM equilibrium, reflecting the efficiency of the smart-contract in removing dead-weight costs from refinancing.

7.4.2 Financial literacy interventions

We next consider the general equilibrium effects of interventions that improve the degree of attention in the population. In practice, various policies – ranging from systematic financial literacy

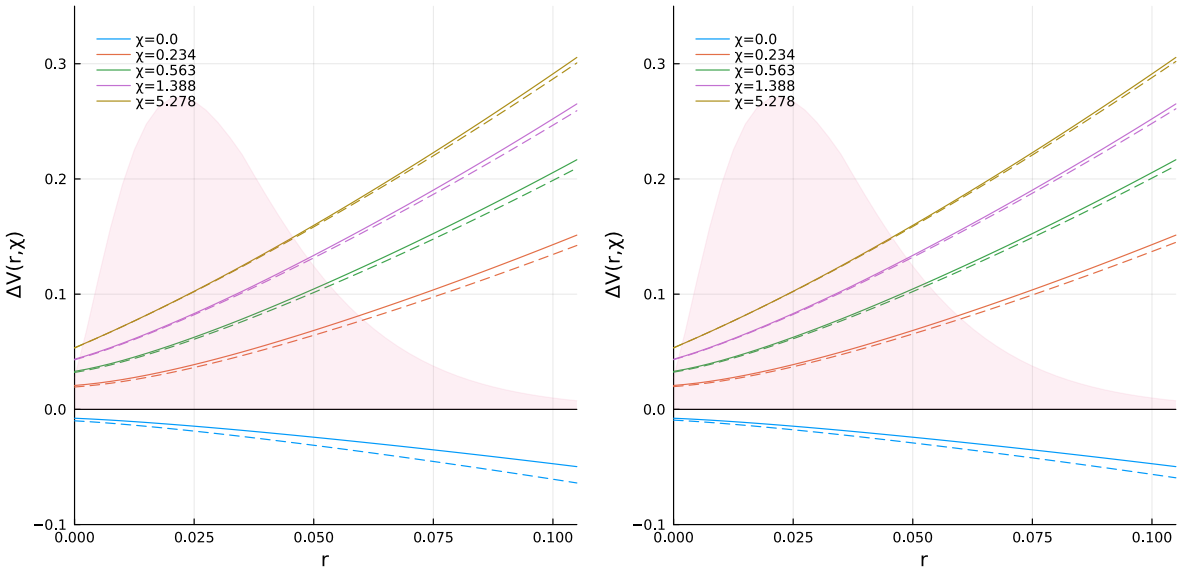

Figure 11: **Financial literacy interventions.** Left panel depicts the cross-subsidy ΔV by group before the shift (solid) and after the attention shift of all below-median FICO households to the distribution of above-median FICO households (dashed). Right panel depicts similar measures of cross-subsidies when the sub-sample of focus are split by above vs. below median initial loan balance.

education in school, to advertising campaigns making people more comfortable interacting with the banking sector – can lead to such improvements. To discipline our counterfactual calculations, we consider (i) improving the below-median FICO attention distribution to that of the above-median FICO attention distribution, and (ii) improving the below-median loan amount attention distribution to that of the above-median loan amount attention distribution. The applicable subsample distributions are given in Table E.3.

The solid lines in the left and right panels of Figure 11 provide the lifetime value at origination for the households, a reproduction of the left panel of Figure 8. The dashed lines depict the shift in this lifetime value at origination that arises from an improvement in attention rates to FICO+ (left panel) and Loan+ (right panel). We see that any distributional attention improvement shifts the lifetime value of a household down, conditional on a household's type. This is caused by mortgage rates increasing, so conditional on remaining the same type a household is worse off. Of course, with the distributional shift, a household ex-ante has a higher likelihood of being a higher attention type. But conditional on not changing their type, a household is unambiguously worse off. However, those losses are unevenly distributed: For the FICO shift, the ergodic losses are 56bps for the slowest

group, and 5bps for the fastest group.

A key take-away from these examples is that poorly targeted educational programs – or an uneven uptake of nudges to refinance – can worsen the redistribution embedded in the mortgage market in the context of a pooling equilibrium. In other words, if financial literacy interventions ultimately fail to reach the bottom of the attention distribution, they can actually worsen the cross-subsidies paid by the slowest households.

7.4.3 The rise of non-bank mortgage lending

Lastly, the mortgage lending industry has witnessed a structural shift over the past 20 years, with non-bank lenders responsible for a growing share of mortgage origination, at the expense of the more traditional banking sector. Using SFLP data and the bank vs. non-bank classification of Buchak et al. (2018), we plot the share of bank, non-bank and "others" mortgage origination over time in Figure 12 (left panel).[41] While banks had a greater than 75% market share of conforming mortgage originations in 2000, that share has consistently declined. At the same time, non-bank originators' volumes have increased, from less than 5% of total originations to more than 30% as of end of 2021.

Using SFLP data, we then study the differential prepayment propensity for mortgages originated by banks vs. non-banks as a function of rate gaps. In other words, we estimate the linear probability model:

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

for borrower i with mortgage contract j at time t, where X is a vector of controls. Figure 12 shows the point estimates for $\beta_{gapbin,bank}$ and $\beta_{gapbin,non-bank}$ in our fully saturated specification, and where the bins used are constructed in 50bps intervals. The "S-curves" estimated – and in particular the difference in level between negative and positive rate gaps – directly give us the average attention rate for bank and non-bank originated mortgage borrowers. On average, borrowers for bank-originated mortgages tend to be 100bps per month less attentive than borrowers for non-bank-originated mortgages – a substantial difference in behavior. To reduce the risk that households borrowing from non-bank mortgage lenders are systematically different from those borrowing from

[41] The identity of the originator (or "seller") is not available in the CRISM dataset. The SFLP data does 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".

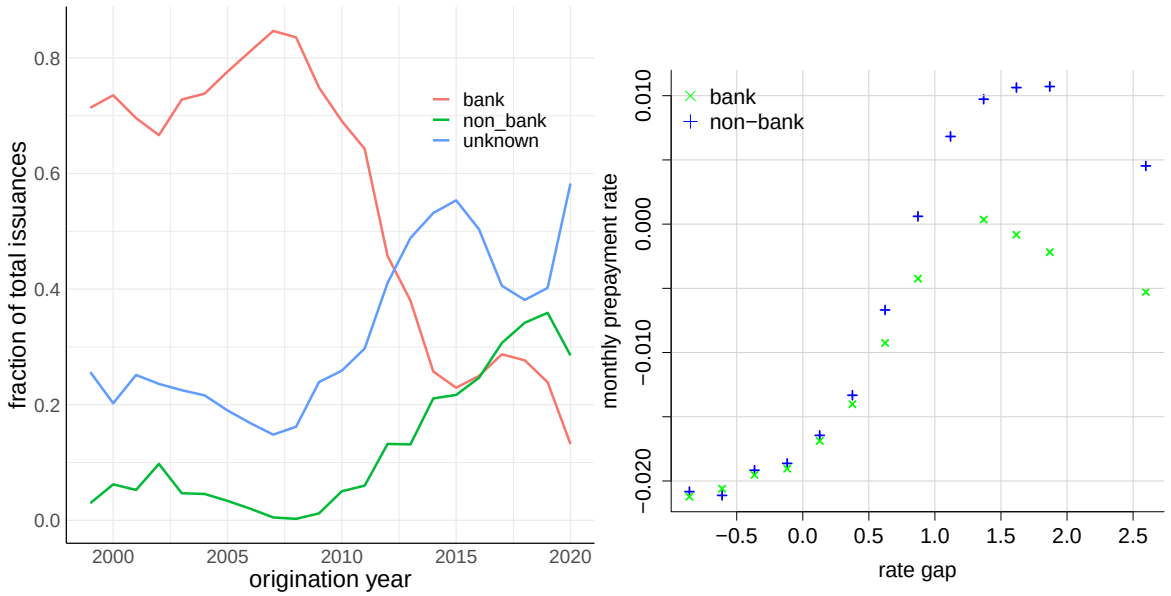

Figure 12: **Rise in non-bank lending.** Left panel shows the fraction of new mortgages that are classified as originated by "banks", "non-banks" vs. "unknown". Right panel shows our non-parametric estimate of the impact of the rate gap onto prepayment rates in the SFLP loan sample.

banks, we saturate our regression with a battery of contract-specific and household-specific controls within the vector X.[42] 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 a selection effects of borrowers into fintech loans.

Using our estimate that non-bank lenders increase household's effective attention rates by 12 p.p. per annum, we can compute the extent to which the rise of non-bank mortgage lending puts upward pressure on mortgage interest rates. In a counterfactual Approximate Pooling MPE in which households' attention rates are uniformly increased by 12%, ergodic average new-issue mortgage rates would increase by approximately 50bps. This counterfactual thus highlights that the dynamics of the mortgage origination market, and in particular the aggressive nudging practices of non-bank lenders, can have a large and systematic effect on mortgage market interest rates.

[42] Those controls include (i) a fully non-parametric function of borrower's original FICO score, (ii) a fully non-parametric function of borrower's original original combined LTV ratio, (iii) a first-time home buyer flag, and (iv) the borrower's original real income. The SFLP data does not contain a borrower ID, only a loan ID; we are thus unable to include household fixed effects into our regression.

8 General framework

While we focus on the US residential mortgage market, our model framework is more general and can be applied to several other environments where economic agents are ex-ante heterogeneous, make dynamic discrete choices about supplying or purchasing a particular good or service subject to some frictions, and the other side of the market is competitive but cannot price-discriminate for informational or legal reasons. In the context of mortgages, households face time- and state-dependent frictions when making the discrete choice about whether (or not) to refinance their existing loan; loan suppliers are competitive mortgage market investors, facing time-varying, aggregate shocks (via movements in their cost of funds r_t). The inability for mortgage investors to price-discriminate based on type stems from the structure of the US agency MBS market. Mortgage refinancing is not the only economic environment where our modeling approach is relevant.

8.1 Wage setting in labor markets

The labor market is a second important laboratory in which our framework can be applied. Consider for instance a model of wage determination with stochastic productivity, risk-averse workers and one-sided commitment by the firm, as in Harris and Holmstrom (1982). More specifically, 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* χ – the rate at which they seek offers from competing firms. Let H be the ex-ante distribution over workers' job hunting rate. 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 $\mathcal{W}(x_{it})$ of the worker's productivity at the time t at which they were hired.[43] 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 τ, the worker compares the proposed wage $\mathcal{W}(x_{i\tau})$ to their current wage w, and decides to accept the offer if the life-time utility $V(x_{i\tau}, w)$ from staying in the current job is below the life-time utility $V(x_{i\tau}, \mathcal{W}(x_{i\tau})) - \psi$ from moving, potentially incurring some welfare costs ψ when switching firms.

Firms are risk-neutral, competitive, and discount profits at the constant rate r. From the firm's

[43] 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.

point of view, the value of a worker with productivity x, wage w, and job hunting rate χ is

$$\Pi(x, w; \chi) = \mathbb{E}_x \left[\int_0^\tau e^{-rt} (x_t - w) \, dt \right], \tag{27}$$

where τ is the quit time of the type-χ worker. Since we have assumed that workers' productivity is only exposed to idiosyncratic shocks, there exists a well defined stationary density $f_\infty(x, w, \chi)$ over workers' productivity x, wage rate w and type χ.[44] From this stationary density f_∞, we can back out the corresponding type (stationary) distribution of *job transitioners* $G(\chi|x)$ conditional on productivity x. At the time a firm makes an offer to a prospective employee, the firm acts competitively, offering a wage $\mathcal{W}(x)$ that satisfies

$$\mathbb{E}^{G(\chi|x)} \left[\Pi(x, \mathcal{W}(x); \chi) \right] = 0. \tag{28}$$

This break-even condition is the counterpart to (17) in the context of the mortgage market, and it pins down the equilibrium wage rate $\mathcal{W}$. The expectation on the left hand side of (28) encapsulates the idea that firms cannot discriminate based on workers' type χ – either because this type cannot be observed by the firm, or because of discrimination laws covering protected classes which might be correlated with χ.[45] Finally, one can easily define a Pooling MPE of this environment, in which (i) workers optimally switch firms subject to their search and job hunting friction, (ii) firms' profits satisfy (27), and (iii) the equilibrium wage rate satisfies the break-even condition (28).

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 in order 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 details the labor market application of the framework developed in this article, we also emphasize that other environments could lend themselves to such an analysis.

[44] Our statement assumes that the equilibrium wage rate $\mathcal{W}$ is monotonic increasing in productivity; this property can be verified ex-post, when our postulated equilibrium has been computed numerically.

[45] For example, marital status might predict mobility but it is illegal to condition wages on marital status.

Subscription-based businesses for instance can be mapped into our model: cable or cellular phone services firms often offer promotional pricing at an initially low rate for a short period of time, before this 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 product 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, the tractability of our approach, and the ability to perform systematic analysis of counterfactuals, remain attractive features of our framework. We leave the precise evaluation and analysis of these different settings for future research.

9 Conclusion

In this paper, we have studied the general equilibrium consequences of pooling ex-ante heterogeneous agents facing various frictions, making dynamic discrete choices about a product or service provided by 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 general equilibrium effects of such counterfactuals.

References

Abel, Andrew B, Janice C Eberly, and Stavros Panageas. 2007. "Optimal inattention to the stock market." *American economic review* 97 (2):244–249.

Achdou, Yves, Francisco J Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2014. "Partial differential equation models in macroeconomics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372 (2028):20130397.

Achdou, Yves, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2021. "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach." *The Review of Economic Studies* 89 (1):45–86.

Agarwal, Sumit, John C Driscoll, and David I Laibson. 2013. "Optimal mortgage refinancing: a closed-form solution." *Journal of Money, Credit and Banking* 45 (4):591–622.

Agarwal, Sumit, Richard J Rosen, and Vincent Yao. 2016. "Why do borrowers make mortgage refinancing mistakes?" *Management Science* 62 (12):3494–3509.

Ahn, SeHyoun, Greg Kaplan, Benjamin Moll, Thomas Winberry, and Christian Wolf. 2018. "When inequality matters for macro and macro matters for inequality." *NBER macroeconomics annual* 32 (1):1–75.

Amromin, Gene, Jennifer Huang, Clemens Sialm, and Edward Zhong. 2018. "Complex mortgages." *Review of Finance* 22 (6):1975–2007.

Andersen, Steffen, John Y Campbell, Kasper Meisner Nielsen, and Tarun Ramadorai. 2020. "Sources of inaction in household finance: Evidence from the Danish mortgage market." *American Economic Review* 110 (10):3184–3230.

Auclert, Adrien and Matthew Rognlie. 2016. "Unique equilibrium in the Eaton–Gersovitz model of sovereign debt." *Journal of Monetary Economics* 84:134–146.

Bach, Laurent, Laurent E Calvet, and Paolo Sodini. 2020. "Rich pickings? Risk, return, and skill in household wealth." *American Economic Review* 110 (9):2703–47.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The distribution of wealth and fiscal policy in economies with finitely lived agents." *Econometrica* 79 (1):123–157.

Beraja, Martin, Andreas Fuster, Erik Hurst, and Joseph Vavra. 2019. "Regional heterogeneity and the refinancing channel of monetary policy." *The Quarterly Journal of Economics* 134 (1):109–183.

Berger, David, Konstantin Milbradt, Fabrice Tourre, and Joseph Vavra. 2021. "Mortgage prepayment and path-dependent effects of monetary policy." *American Economic Review* 111 (9):2829–78.

Berk, Jonathan B, Richard Stanton, and Josef Zechner. 2010. "Human capital, bankruptcy, and capital structure." *The Journal of Finance* 65 (3):891–926.

Bessembinder, Hendrik, William F Maxwell, and Kumar Venkataraman. 2013. "Trading activity and transaction costs in structured credit products." *Financial Analysts Journal* 69 (6):55–67.

Boyarchenko, Nina, Andreas Fuster, and David O Lucca. 2019. "Understanding mortgage spreads." *The Review of Financial Studies* 32 (10):3799–3850.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru. 2018. "Fintech, regulatory arbitrage, and the rise of shadow banks." *Journal of financial economics* 130 (3):453–483.

Calvo, Guillermo A. 1983. "Staggered prices in a utility-maximizing framework." *Journal of monetary Economics* 12 (3):383–398.

Campbell, John. 2006. "Household Finance." *Journal of Finance Studies* 61 (1):1553–1604.

Campbell, John Y, Howell E Jackson, Brigitte C Madrian, and Peter Tufano. 2011. "Consumer financial protection." *Journal of Economic Perspectives* 25 (1):91–114.

Chatterjee, Satyajit and Burcu Eyigungor. 2012. "Maturity, indebtedness, and default risk." *American Economic Review* 102 (6):2674–99.

Chernov, Mikhail, Brett R Dunn, and Francis A Longstaff. 2018. "Macroeconomic-driven prepayment risk and the valuation of mortgage-backed securities." *The Review of Financial Studies* 31 (3):1132–1183.

Cox, John C, Jonathan E Ingersoll Jr, and Stephen A Ross. 1985. "A theory of the term structure of interest rates." *Econometrica* .

Décamps, Jean-Paul and Stéphane Villeneuve. 2014. "Rethinking Dynamic Capital Structure Models with Roll-Over Debt." *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 24 (1):66–96.

DeMarzo, Peter M and Zhiguo He. 2021. "Leverage dynamics without commitment." *The Journal of Finance* 76 (3):1195–1250.

Fagereng, Andreas, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. 2016. "Heterogeneity in returns to wealth and the measurement of wealth inequality." *American Economic Review* 106 (5):651–55.

Fisher, Jack, Alessandro Gavazza, Lu Liu, Tarun Ramadorai, and Jagdish Tripathy. 2021. "Refinancing cross-subsidies in the mortgage market." *Bank of England Staff Working Paper* (948).

Fleming, Wendell H and Halil Mete Soner. 2006. *Controlled Markov processes and viscosity solutions*, vol. 25. Springer Science & Business Media.

Fuster, Andreas, Laurie S Goodman, David O Lucca, Laurel Madar, Linsey Molloy, and Paul Willen. 2013. "The rising gap between primary and secondary mortgage rates." *Available at SSRN 2378439* .

Fuster, Andreas, Stephanie H Lo, and Paul S Willen. 2017. "The time-varying price of financial intermediation in the mortgage market." Tech. rep., National Bureau of Economic Research.

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

Fuster, Andreas, Matthew Plosser, Philipp Schnabl, and James Vickery. 2019. "The role of technology in mortgage lending." *The Review of Financial Studies* 32 (5):1854–1899.

Gao, Pengjie, Paul Schultz, and Zhaogang Song. 2017. "Liquidity in a Market for Unique Assets: Specified Pool and To-Be-Announced Trading in the Mortgage-Backed Securities Market." *The Journal of Finance* 72 (3):1119–1170.

Gerardi, Kristopher, Paul Willen, and David Hao Zhang. 2020. "Mortgage Prepayment, Race, and Monetary Policy." Working paper.

Golosov, Mikhail and Robert E Lucas Jr. 2007. "Menu costs and Phillips curves." *Journal of Political Economy* 115 (2):171–199.

Guren, Adam M., Arvind Krishnamurthy, and Timothy J. McQuade. 2021. "Mortgage Design in an Equilibrium Model of the Housing Market." *The Journal of Finance* 76 (1):113–168.

Harris, Milton and Bengt Holmstrom. 1982. "A Theory of Wage Dynamics." *The Review of Economic Studies* 49 (3):315–333.

He, Zhiguo and Wei Xiong. 2012. "Dynamic debt runs." *The Review of Financial Studies* 25 (6):1799–1843.

Hurst, Erik, Benjamin J Keys, Amit Seru, and Joseph Vavra. 2016. "Regional redistribution through the US mortgage market." *American Economic Review* 106 (10):2982–3028.

Jiang, Erica Xuewei. 2019. "Financing competitors: Shadow banks' funding and mortgage market competition." *USC Marshall School of Business Research Paper Sponsored by iORB, No. Forthcoming* .

Keys, Benjamin J, Devin G Pope, and Jaren C Pope. 2016. "Failure to refinance." *Journal of Financial Economics* 122 (3):482–499.

Krusell, Per and Anthony A Smith, Jr. 1998. "Income and wealth heterogeneity in the macroeconomy." *Journal of political Economy* 106 (5):867–896.

Lewis, Daniel J, Davide Melcangi, and Laura Pilossoph. 2019. "Latent heterogeneity in the marginal propensity to consume." *FRB of New York Staff Report* (902).

Maskin, Eric and Jean Tirole. 2001. "Markov perfect equilibrium: I. Observable actions." *Journal of Economic Theory* 100 (2):191–219.

Reis, Ricardo. 2006. "Inattentive consumers." *Journal of monetary Economics* 53 (8):1761–1800.

Spahr, Ronald W and Mark A Sunderman. 1992. "The effect of prepayment modeling in pricing mortgage-backed securities." *Journal of Housing Research* :381–400.

Stanton, Richard. 1995. "Rational Prepayment and the Valuation of Mortgage-Backed Securities." *The Review of Financial Studies* 8 (3):677–708.

Strulovici, Bruno and Martin Szydlowski. 2015. "On the smoothness of value functions and the existence of optimal strategies in diffusion models." *Journal of Economic Theory* 159:1016–1055.

Zhang, David Hao. 2022. "Closing Costs, Refinancing, and Inefficiencies in the Mortgage Market." Working paper.

Appendix

A Proofs: partial equilibrium

A.1 Value function V

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

$$\begin{aligned} V(x,c) :&= \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi_\nu dN_t^{(\nu)} + \psi_\chi dN_t^{(k)}\right)\right], \\ \text{s.t.} \qquad dc_t^{(k)} &= \left(m_t - c_{t-}^{(k)}\right)\left(dN_t^{(k)} + dN_t^{(\nu)}\right), \end{aligned}$$

where $\mathcal{K}$ is a set of progressively measurable intensity processes $k = \{k_t\}_{t\geq 0}$ such that $k_t \in [0,\chi]$ at all times, and $N_t^{(k)}$ (resp. $N_t^{(\nu)}$) is a counting process with jump intensity k_t (resp. ν). Using this definition, we first show that V must be increasing in c. Take $c' > c$, and take an arbitrary intensity policy $k \in \mathcal{K}$. Compute the difference in payoffs for such intensity policy k:

$$\begin{aligned} \mathbb{E}_{x,c'}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi_\nu dN_t^{(\nu)} + \psi_\chi dN_t^{(k)}\right)\right] - \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi_\nu dN_t^{(\nu)} + \psi_\chi dN_t^{(k)}\right)dt\right] \\ \geq \mathbb{E}_x\left[\int_0^{\tau} e^{-\rho t}(c'-c)dt\right] > 0, \end{aligned}$$

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

$$\begin{aligned} V(x,c') &= \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c'}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi_\nu dN_t^{(\nu)} + \psi_\chi dN_t^{(k)}\right)dt\right] \\ &\geq \inf_{k\in\mathcal{K}} \mathbb{E}_{x,c}\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)}dt + \psi_\nu dN_t^{(\nu)} + \psi_\chi dN_t^{(k)}\right)dt\right] = V(x,c) \end{aligned}$$

Thus V is increasing in c. Moreover, given that $\chi < +\infty$, a simple reasoning by contradiction can show that V is in fact strictly increasing in c. Problem (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)[46] provide for the value function V being twice continuously differentiable in x, and satisfying the following HJB equation:

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

The minimization problem yields the optimal Markov control $k^*(x,c) = \chi \mathbb{1}_{\{V(x,m(x))+\psi_\chi \leq V(x,c)\}}$. Moreover, 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

$$V(x, m(x)+\theta(x)) := V(x,m(x)) + \psi_\chi$$

Such rate gap cutoff θ 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 (2). □

[46] 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 $\mathcal{L}$.

A.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 standard Brownian motion. The household solves

$$\begin{aligned} V(m,c) :&= \inf_{k\in\mathcal{K}} \mathbb{E}_{m,c}\left[\int_0^{+\infty} e^{-\rho t}\left[c_t dt + \psi_\chi dN_t^{(k)} + \psi_\nu dN_t^{(\nu)}\right]\right] \\ dm_t &= \sigma dB_t \\ dc_t &= (m_t - c_{t-})\left(dN_t^{(k)} + dN_t^{(\nu)}\right). \end{aligned}$$

The rate gap $z_t := c_t - m_t$ evolves according to

$$dz_t = -\sigma dB_t - z_{t-}\left(dN_t^{(k)} + dN_t^{(\nu)}\right).$$

V can be simplified as follows

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

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

$$\begin{aligned} v(z) :&= \inf_{k\in\mathcal{K}} \mathbb{E}_z\left[\int_0^{+\infty} e^{-\rho t}\left[\left(\psi_\chi - \frac{z_{t-}}{\rho}\right)dN_t^{(k)} + \left(\psi_\nu - \frac{z_{t-}}{\rho}\right)dN_t^{(\nu)}\right]\right] \\ dz_t &= -\sigma dB_t - z_{t-}\left(dN_t^{(k)} + dN_t^{(\nu)}\right). \end{aligned}$$

The value function v satisfies the following HJB:

$$(\rho+\nu+\chi)v(z) = \frac{\sigma^2}{2}v''(z) + \chi\min\left(v(z), v(0) + \psi_\chi - \frac{z}{\rho}\right) + \nu\left(v(0) + \psi_\nu - \frac{z}{\rho}\right)$$

Noting θ the rate gap above which the household finds it optimal to refinance when given the opportunity to do so, we must have

$$\begin{aligned} (\rho+\nu)v(z) &= \frac{\sigma^2}{2}v''(z) + \nu\left(v(0) + \psi_\nu - \frac{z}{\rho}\right) && z \le \theta \\ (\rho+\nu+\chi)v(z) &= \frac{\sigma^2}{2}v''(z) + \nu\left(v(0) + \psi_\nu - \frac{z}{\rho}\right) + \chi\left(v(0) + \psi_\chi - \frac{z}{\rho}\right) && z \ge \theta \end{aligned}$$

Introduce the constant $\eta_\chi > 0$:

$$\eta_\chi := \frac{\sqrt{2(\rho+\nu+\chi)}}{\sigma}$$

Note that $v(z) = O(z)$ as $z \to +\infty$ or as $z \to -\infty$. Thus we must have

$$v(z) = k_- e^{\eta_0(z-\theta)} + \frac{\nu}{\rho+\nu}\left[v(0) + \psi_\nu - \frac{z}{\rho}\right] \qquad z \leq \theta \tag{A.1}$$

$$v(z) = k_+ e^{-\eta_\chi(z-\theta)} + \frac{\nu}{\rho+\nu+\chi}\left[v(0) + \psi_\nu - \frac{z}{\rho}\right] + \frac{\chi}{\rho+\nu+\chi}\left[v(0) + \psi_\chi - \frac{z}{\rho}\right] \qquad z \geq \theta \tag{A.2}$$

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

$$v(0) = \frac{\rho+\nu}{\rho}\left[k_- e^{-\eta_0\theta} + \frac{\nu\psi_\nu}{\rho+\nu}\right]$$

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

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

$$\eta_0 k_- - \frac{\nu}{\rho(\rho+\nu)} = -\eta_\chi k_+ - \frac{\nu+\chi}{\rho(\rho+\nu+\chi)}$$

Using our formula for $v(0)$, this simplifies to

$$k_-\left(1 - \frac{\chi e^{-\eta_0\theta}}{\rho+\chi+\nu}\right) - k_+ = \left(\frac{\chi}{\rho+\chi+\nu}\right)\left(\psi_\chi - \frac{\theta}{\rho+\nu}\right)$$

$$\eta_0 k_- + \eta_\chi k_+ = \frac{-\chi}{(\rho+\nu)(\rho+\nu+\chi)}$$

We can solve this linear system for (k_-, k_+) to obtain:

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

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

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

$$v(\theta) = v(0) + \psi_\chi - \frac{\theta}{\rho}$$

$$\Rightarrow k_- = \frac{\rho}{\rho+\nu}v(0) + \psi_\chi - \frac{\nu}{\rho+\nu}\psi_\nu - \frac{\theta}{\rho+\nu}$$

But since we know $v(0)$ as a function of k_-, this yields

$$k_- = k_- e^{-\eta_0\theta} + \psi_\chi - \frac{\theta}{\rho+\nu}$$

Using the formula we obtained for k_-, this yields, after some algebra, the implicit equation

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

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, $F(0) = 1 < 1 + \left[\eta_0 + \frac{\rho+\nu}{\chi}(\eta_0+\eta_\chi)\right](\rho+\nu)\psi_\chi$, and $F(\theta) \to +\infty$ when $\theta \to +\infty$. In other words, the above equation admits a unique positive solution θ. Re-write the implicit equation satisfied by θ as follows:

$$\begin{aligned} e^{-\eta_0\theta} + (\eta_0+\epsilon_\chi)\,\theta &= 1 + (\rho+\nu)\psi_\chi\,(\eta_0+\epsilon_\chi) \\ \epsilon_\chi &:= \frac{(\rho+\nu)(\eta_0+\eta_\chi)}{\chi} \end{aligned} \tag{A.3}$$

Clearly, ϵ_χ is a positive and decreasing function of χ, converging to zero as $\chi \to +\infty$. Differentiate the above equation w.r.t. χ to obtain

$$\frac{\partial\theta}{\partial\chi} = \frac{((\rho+\nu)\psi_\chi - \theta)\,\frac{\partial\epsilon_\chi}{\partial\chi}}{\epsilon_\chi + \eta_0\,(1-e^{-\eta_0\theta})} > 0,$$

with the last inequality following from $\frac{\partial\epsilon_\chi}{\partial\chi} < 0$ and $(\rho+\nu)\psi_\chi - \theta = \frac{e^{-\eta_0\theta}-1}{\eta_0+\epsilon_\chi} < 0$. Thus we have proven that the barrier θ increases as households become more attentive. Finally, consider the asymptotic behavior of θ. When $\chi \to +\infty$, the limiting value θ_∞ solves

$$e^{-\eta_0\theta_\infty} + \eta_0\theta_\infty = 1 + \eta_0\psi_\chi(\rho+\nu)$$

This delivers the ADL formula, using the Lambert function W:

$$\theta_\infty = \frac{1}{\eta_0}\left(1 + \eta_0\psi_\chi(\rho+\nu) + W\left(-\exp\left(-1-\eta_0\psi_\chi(\rho+\nu)\right)\right)\right)$$

Instead, consider the case $\chi \to 0$. In that case, the optimal threshold converges to

$$\theta_0 = (\rho+\nu)\psi_\chi$$

Finally, we can perform a Taylor expansion of (A.3) around $\theta = 0$, which allows us to obtain an approximation $\hat{\theta}$ of the value θ:

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

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

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

□

A.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, i.e. it satisfies the stochastic differential equation

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

This parametrization for the mortgage rate process nests the special case of the Brownian motion we studied in Proposition 2, by setting $\kappa = 0$. We solve the stochastic control problem (1) numerically with a finite difference scheme, and plot in Figure A.1 the ergodic average threshold $\mathbb{E}[\theta(x_t)]$ and the ergodic average slope of the threshold $\mathbb{E}[\theta'(x_t)]$ as a function of the speed of mean-reversion κ. 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 rate $\nu = 4.1\%$ (consistent with Berger et al. (2021)), (iv) fixed moving and refinancing costs $\psi_\nu = \psi_\chi = 2\%$ (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.

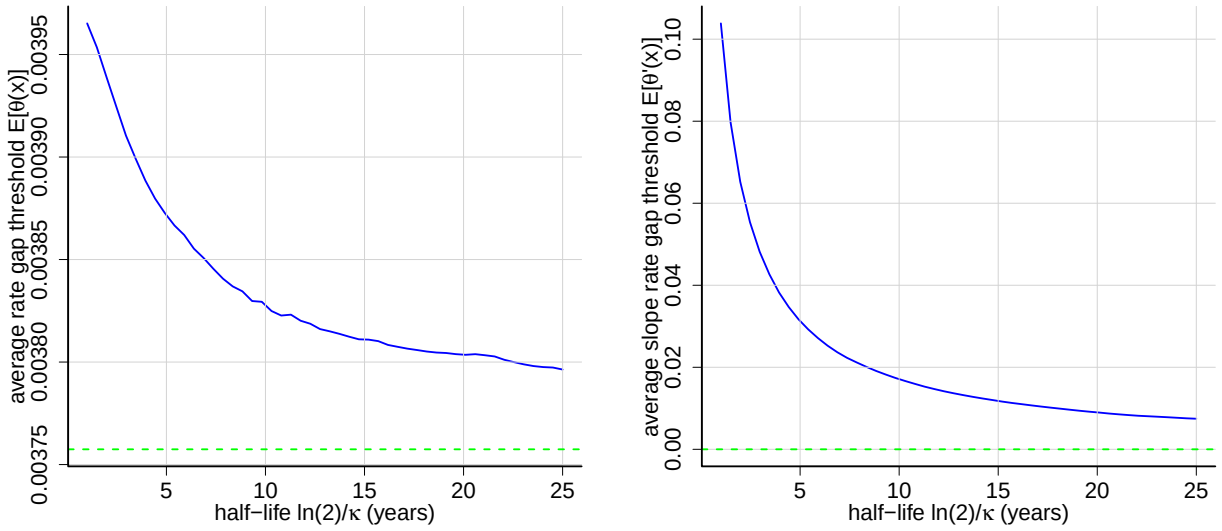

Figure A.1: **Rate gap threshold θ vs. mortgage process half-life.** Rate gap threshold θ in the case of m_t following an OU process for various speeds of mean-reversion κ. Left plot shows the ergodic average threshold $\mathbb{E}[\theta(x_t)]$, while right plot shows the ergodic average slope of the threshold $\mathbb{E}[\theta'(x_t)]$. Horizontal dash green line represents the limit $\chi \to +\infty$. Figure computed for $\rho = 5\%$, $\nu = 4.1\%$, $\psi_\nu = \psi_\chi = 2\%$, $\bar{m} = \mathbb{E}[m_t] = 5\%$ and $\sigma = 1\%$.

Figure A.1 suggests that the average rate gap threshold $\mathbb{E}[\theta(x)]$ 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 (which corresponds to $\kappa = 0$, i.e., the far right on the x-axis represented by the dashed green line) to a mean-reverting process introduces a non-negligible amount of state dependence in the barrier $\theta(x)$. The average slope $\mathbb{E}[\theta'(x)]$ is positive, meaning that the rate gap threshold is lowest at low mortgage rates, and

highest at high mortgage rates. Quantitatively, when the mortgage rate half-life is around 5 years (the type of persistence we observe in the data), the average slope of the decision rule is 0.04 – in other words, each p.p. increase in the current mortgage rate increases the rate gap threshold $\theta(x)$ by 4bps.

B Proofs: general equilibrium

B.1 MPE existence and uniqueness in homogeneous case

Proof of Proposition 3. Using Feynman-Kac, the pricing function P satisfies the equation

$$r(x)P(x,c;\chi) = c - f + \mathcal{L}P(x,c;\chi) + \left(\nu + \chi \mathbb{1}_{\{c-m(x)\geq\theta(x)\}}\right)(1 - P(x,c;\chi)).$$

The function P, solution of (10), 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. We will prove Proposition 3 in 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. $\psi_\chi = 0$), 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 $k^*(x,c) = \chi \mathbb{1}_{\{c\geq m(x)\}}$.

i. In this section, we restrict ourselves to the case where $\pi = 0$; in this case, there are no mortgage origination costs and thus no dead-weight losses. To make further progress, we study an alternative problem formulation. Consider the auxiliary problem

$$\tilde{P}(x,c;\chi) := \inf_{k\in\mathcal{K}} \mathbb{E}_x\left[\int_0^{+\infty} e^{-\int_0^t (r(x_s)+k_s+\nu)ds}(c - f + k_t + \nu)\,dt\right], \tag{B.1}$$

where the set $\mathcal{K}$ was defined in Appendix A.1. The advantage of this problem is that 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 Appendix A.1 allow us to argue that $\tilde{P}$ is twice continuously differentiable in x, continuous and increasing in c, satisfying the HJB equation

$$(r(x)+\nu)\tilde{P}(x,c;\chi) = c - f + \nu + \mathcal{L}\tilde{P}(x,c;\chi) + \min_{k\in[0,\chi]}\left\{k\left(1 - \tilde{P}(x,c;\chi)\right)\right\}. \tag{B.2}$$

Clearly the optimal Markov control is $\tilde{k}(x,c) = \chi \mathbb{1}_{\left\{\tilde{P}(x,c;\chi)\geq 1\right\}}$. Reinjecting this control into the HJB equation yields

$$(r(x)+\nu)\tilde{P}(x,c;\chi) = c - f + \nu + \mathcal{L}\tilde{P}(x,c;\chi) + \chi \mathbb{1}_{\left\{\tilde{P}(x,c;\chi)\geq 1\right\}}\left(1 - \tilde{P}(x,c;\chi)\right). \tag{B.3}$$

Notice that r_t is restricted to be on $\mathbb{R}_+$, which means that we must have $\tilde{P}(x,0;\chi) < 1$. Similarly, since r_t is bounded above by $\bar{r}$, it is also clear that 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

$$\tilde{P}(x, m(x);\chi) = 1 \tag{B.4}$$

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 (B.4) is unique and satisfies the equilibrium condition (11).

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 $\nu + \chi \mathbb{1}_{\{x_t \leq x^*\}}$. Formally, if ω is a (unit mean) exponentially distributed random variable and if we introduce 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 $\Lambda_{\tau_{x^*,\chi}} = \omega$. As will be seen shortly, x^* will represent the latent state that was prevalent the last time a household refinanced. Consider the interest only "IO" and principal only "PO" net present values, defined via

$$IO(x; \chi) := \mathbb{E}_x \left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds} dt \right] \tag{B.5}$$

$$PO(x; \chi) := \mathbb{E}_x \left[e^{-\int_0^{\tau_{x,\chi}} r_s ds} \right] \tag{B.6}$$

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 $\nu + \chi \mathbb{1}_{\{x_t \leq x\}}$. Introduce the function m, defined via

$$m(x) := f + \frac{1 - PO(x; \chi)}{IO(x; \chi)} + \frac{\pi}{IO(x; \chi)}. \tag{B.7}$$

m is well defined, and is a continuous function of 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 (B.1) is decreasing in x, it must be the case that the function $\tilde{P}$ defined in (B.1) 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 (B.7) (with $\pi = 0$). To prove this statement, observe that

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

Rewriting this condition via the IO and PO terms, we have

$$1 = (m(x) - f) IO(x; \chi) + PO(x; \chi),$$

which directly implies (B.7) for $\pi = 0$. As $m(x)$ is increasing when $\pi = 0$, it must be the case that $\left(1 - PO(x; \chi)\right) / IO(x; \chi)$ 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 (B.7), is monotone increasing in x. Given this observation, we must have an equilibrium in which m is the mortgage rate, since m must satisfy

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

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 satisfy (B.7).

□

B.2 Prepayment rate invariance

Proof of Proposition 4. Refinancing occurs when $c = m(x^*) > m(x)$ and the households is paying attention. As show in Proposition 3, m is monotone increasing, which means that the refinancing set $\{c : c > m(x)\}$ can be translated to the set $\{x^* : m(x^*) > m(x)\}$, which in turn is equivalent to the set $\{x^* : x^* > x\}$. As the last set is not dependent on the mortgage pricing function $m(\cdot)$, the refinancing set is invariant to π. Consequently, the ergodic distribution over $(x_t, x^*_{t,\chi})$, and thus the ergodic average prepayment rate absent upfront closing costs $\psi_\chi = 0$, are invariant to π. □

B.3 Comparative statics

Proof of Proposition 5. Consider first the case $\pi = 0$. In that case, since $P = \tilde{P}$ can be defined via equation (B.1), it must be the case that P is decreasing in χ. Thus, the mortgage rate function $m(x)$ is an increasing function of χ, 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 χ. Consider $IO(x, x^*; \chi) := \mathbb{E}_x \left[\int_0^{\tau_{x^*,\chi}} e^{-\int_0^t r_s ds} dt \right]$, which solves the PDE

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

Differentiate this equation w.r.t. χ to obtain

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

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

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

Thus, $IO(x;\chi)$, defined in (B.5), is monotone decreasing in χ. This must mean that the mortgage rate function, defined via (B.7), is increasing in χ, whenever $\pi > 0$.

□

B.4 Infinite dimensional problem with heterogeneous households

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 χ in the population at time t (with associated joint density $f_t(c,\chi)$). Since types are a permanent household attribute, we must have

$$\int_c f_t(c,\chi)dc = h(\chi). \tag{B.8}$$

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 $c \neq m(S)$:

$$df_t(c,\chi) = -\left(\nu + \chi \mathbb{1}_{\{c \geq m(S_t)\}}\right) f_t(c,\chi)dt, \qquad c \neq m(S). \tag{B.9}$$

The density f_t, between t and $t+dt$, looses mass at rate ν 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

$$\lim_{c \uparrow m(S_t)} \partial_c f_t(c,\chi) - \lim_{c \downarrow m(S_t)} \partial_c f_t(c,\chi) = \nu h(\chi) + \chi \int_{m(S_t)}^{+\infty} f_t(c,\chi)dc. \tag{B.10}$$

The right-hand-side of this equation is the flux of households exogenously moving at rate ν and the flux of type-χ 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 χ. 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-χ households:

$$\begin{aligned} r(x)P(S,c;\chi) = c + \mathcal{L}P(S,c;\chi) + \left(\nu + \chi \mathbb{1}_{\{c \geq m(S)\}}\right)[1 - P(S,c;\chi)] \\ + \int \mathcal{T}[f](c,\chi)\frac{\delta P}{\delta f(c,\chi)}dcd\chi, \end{aligned} \tag{B.11}$$

with $\delta P/\delta f$ the functional derivative of P w.r.t. f at (c,χ) and the operator $\mathcal{T}$ defined via

$$\mathcal{T}[f](c,\chi) = -\left(\nu + \chi \mathbb{1}_{\{c > m(S)\}}\right) f(c,\chi) \tag{B.12}$$

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.

B.5 Approximate Pooling MPE existence and uniqueness with heterogeneity

Proof of Proposition 6. We establish the existence and uniqueness of the Approximate Pooling MPE using a method similar to Section B.1 for the case $\pi > 0$. To that effect, consider the

dynamic system $(x_t, x^*_{t,\chi})$, where

$$\begin{aligned} dx_t &= \mu(x_t)dt + \sigma(x_t)dB_t \\ dx^*_{t,\chi} &= \left(x_t - x^*_{t-,\chi}\right) dN^\chi_t, \end{aligned}$$

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

$$\mathcal{L}_{x,x^*}\phi(x,x^*) = \mathcal{L}\phi(x,x^*) + \left(\nu + \chi \mathbb{1}_{\{x \leq x^*_\chi\}}\right)(\phi(x,x) - \phi(x,x^*))$$

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

$$g(\chi) = \frac{h(\chi)\int_x \left[\left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*|x,\chi)dc\right)\right] f_\infty(x)dx}{\int_\chi h(\chi)\int_x \left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*|x,\chi)dc\right) f_\infty(x)d\chi dx}, \tag{B.13}$$

or the conditional one defined via

$$g(\chi|x) = \frac{h(\chi)\left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*, x|\chi)dc\right)}{\int_\chi h(\chi)\left(\nu + \chi \int_{x^* \geq x} f_\infty(x^*, x|\chi)dc\right)d\chi}. \tag{B.14}$$

Then, if the function $m(x; G)$ is increasing in x, where m is defined via

$$m(x;G) := f + \frac{1 + \pi - \mathbb{E}^G\left[PO(x;\chi)\right]}{\mathbb{E}^G\left[IO(x;\chi)\right]}, \tag{B.15}$$

we have an equilibrium, and this equilibrium is unique amongst all monotone equilibria. Thus, we 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 $\bar{P}_G\left(x, m(x^*)\right)$ of a mortgage with coupon $m(x^*)$,

$$\bar{P}_G(x, m(x^*)) := \mathbb{E}^G\left[\mathbb{E}_x\left[\int_0^{\tau_{x,\chi}} e^{-\int_0^t r_s ds}(m(x^*) - f)dt + e^{-\int_0^{\tau_{x,\chi}} r_s ds}\right]\right],$$

then clearly if m is increasing, $\bar{P}_G$ must be increasing in x^*, with $\bar{P}_G\left(x^*, m(x^*)\right) = 1 + \pi$ – in other words the equilibrium conditions are satisfied. □

B.6 Integral representation of $\bar{P}_G$ for unconditional $G(\chi)$

Proof of Proposition 7. The HJB equation (10) holds for all χ, and thus, taking expectations w.r.t. the unconditional issuance type distribution $G(\chi)$, we have

$$r(x)\bar{P}_G(x,c) = c - f - \mathbb{1}_{\{m(x)\leq c\}}\mathbb{C}\mathrm{ov}^G\left(\chi, P(x,c;\chi)\right) + \mathcal{L}\bar{P}_G(x,c) + \left(\nu + \bar{\chi}_G \mathbb{1}_{\{m(x)\leq c\}}\right)\left(1 - \bar{P}_G(x,c)\right) \quad \text{(B.16)}$$

One can then use Feynman-Kac to conclude that $\bar{P}_G$ admits the integral representation (21). □

B.7 Invariance of lowest attainable mortgage rate

Proof of Proposition 8. Under the assumption that x is uni-dimensional and that $r(\cdot)$ is monotone increasing, call $\underline{x}$ the lowest bound for x. The monotone Pooling MPE implies m is increasing in x, so then $m(\underline{x})$ must be the lowest attainable mortgage rate. Then we have

$$P(x, m(\underline{x}); \chi) = P\left(x, m(\underline{x}); \chi'\right), \qquad \forall \chi, \chi',$$

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

$$P(\underline{x}, m(\underline{x}); \chi) = 1 + \pi, \qquad \forall \chi.$$

This implies that $m(\underline{x})$ is invariant to the distribution G. □

B.8 Origination distribution $G(\cdot)$

In this section, we describe the method used to determine the time-invariant origination distribution G. We first conjecture a pricing function $m(x)$. Given this pricing function, conditional on type χ, we derive the ergodic joint distribution over the latent state and coupons, denoted $f_\infty(x, c|\chi)$. This distribution can be thought of as the long-run fraction of time a type-χ household spends with coupon c in latent state x:

$$f_\infty(x, c|\chi)dxdc := \lim_{T\to+\infty} \frac{1}{T}\int_0^T \mathbb{1}_{\{x_t\in[x,x+dx], c_t\in[c,c+dc];\chi\}} dt$$

We then denote $f_\infty(x)$ the ergodic density of x (which only depends on our assumed stochastic process for x_t, and thus the infinitessimal operator $\mathcal{L}$) and $f_\infty(c|x,\chi) := f_\infty(x,c|\chi)/f_\infty(x)$ the ergodic coupon density conditional on the latent state x and the type χ. The ergodic origination distribution *conditional* on the latent state x is denoted $G(\chi|x)$, and its density is equal to

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

Instead, the ergodic *unconditional* origination distribution is denoted $G(\chi)$, and its density is given by

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

With the new origination distribution G in hand, we can now update our guess pricing to $m(x, G)$, and iterate on our procedure until successive iterations no longer affect G. At that point, the pricing function $m(x, G)$ is consistent with the ergodic coupon and gap distributions.

Importantly, when x is uni-dimensional, the derivations of $G(\chi)$ and $G(\chi|x)$ can simplify to a one-step procedure that does not require a loop. Indeed, so long as $m(x, G)$ is monotone in x, the latent states x at which households find it optimal to refinance are invariant to the pricing function m: they refinance whenever $x < x^*$, where $m(x^*) := c$. If x is multi-dimensional or the candidate m is not monotone, we must revert to the iterative procedure described above. In Appendix F we illustrate the implied pricing functions for our model calibration, and verify that they are indeed monotone in the latent state x.

B.9 Discussion of Krusell-Smith algorithm

An alternative method to solve the general equilibrium of our model with permanent attention heterogeneity would be to rely on methods originally developed by Krusell and Smith (1998) (thereafter, "KS"). The cross-sectional distribution f_t must then be summarized by a small number of moments that are good summary statistics for the aggregate behavior of our dynamic system. Assume for simplicity that there is a discrete number of types $\chi_1, ..., \chi_k$. Since the pricing problem of investors depends on the origination distribution G_t, which itself solely depends on the distribution f_t, one thoughtful choice of aggregate state variable is, for each type χ_i, the fraction of households that have a positive rate gap, $1 - F_t(m_t, \chi_i) := \Phi_{i,t}$. Indeed, note in that case that:

$$g_t(\chi_i) = \frac{(\nu + \chi_i \Phi_{i,t})h_i}{\sum_{j \leq k}(\nu + \chi_j \Phi_{j,t})h_j}$$

With KS, we then have to postulate state dynamics of the type $d\Phi_{i,t} = \mu_{\Phi,i}(x_t, \vec{\Phi}_t)dt$, for a set of k unknown drift functions $\mu_{\Phi,i}$ to be determined.

Solving the general equilibrium of our model then means that we need to solve for the mortgage function $m_t = m(x_t, \vec{\Phi}_t)$, as well as the drift functions $\mu_{\Phi,i}$, for $i \leq k$. For guess functions m and $\mu_{\Phi,i}$, for $i \leq k$, one can then solve the Feynman-Kac equation satisfied by the type-specific price function $P(x, \vec{\Phi}; \chi)$, and integrate over the distribution g_t (which only depends on $\vec{\Phi}_t$) in order to compute the pool price $\bar{P}$. Once the pool price is obtained for all states $(x, \vec{\Phi})$, one can then update our mortgage rate function m, by solving the implicit equation $\mathbb{E}^G\left[P(x, \vec{\Phi}, m(x, \vec{\Phi}); \chi)\right] = 1 + \pi$. Afterwards, KS requires us to simulate the dynamic system, and update our drift functions $\mu_{\Phi,i}, i \leq k$, using non-linear regression methods.

The KS logic relies on the idea that the dynamics for $\vec{\Phi}_t$ can be well described by a first-order

Markov process. We show that this is not the case. Indeed, we have, for any arbitrary type i:

$$\begin{aligned} d\Phi_{i,t} &= d\left(\int_{m_t}^{+\infty} f_t\left(c,\chi_i\right)dc\right) \\ &= -\left(\nu+\chi_i\right)\Phi_{i,t}dt - \left[\partial_S\left(F_t(m(S),\chi_i)\right)\cdot\mu_S + \frac{1}{2}\mathrm{tr}\left(\sigma_S'\partial_{SS'}\left(F_t(m(S),\chi_i)\right)\sigma_S\right)\right]dt \\ &\qquad - \sigma_S'\partial_S\left(F_t(m(S),\chi_i)\right)dZ_t, \end{aligned}$$

with μ_S, σ_S the drift and diffusion of the infinite-dimensional state vector $S_t = (x_t, f_t)$. While the first term in the above equation is linear in $\Phi_{i,t}$, the second term introduces distortions arising from aggregate shocks and their impact on the fraction of households with positive rate gaps, via the mortgage market interest rate m. This suggests that the dynamics of the potential state vector $\vec{\Phi}_t$ depart potentially significantly from those of a first-order Markov process.

Moreover, the KS approach also relies on the idea that absent aggregate shocks, the dynamic system admits a stationary cross-sectional distribution that does not depart significantly from the time-varying cross-sectional distribution in the presence of *small* aggregate shocks. In our paper, once we shut down aggregate shocks, interest rates and mortgage rates are constant, and the cross-sectional density f becomes a Dirac mass point. In the presence of aggregate shocks that are sufficiently large that the model-implied interest rate dynamics match those that we see in the data, the cross-sectional distribution departs significantly from the Dirac mass point, once again highlighting the shortcomings of KS in the context of our model.

B.10 Externality

One can reformulate the MPE as a game played by a continuum of households of measure 1. For simplicity, we will focus on symmetric equilibria. Consider $k_i(x,c)$ to be the refinancing intensity of a given borrower, and note $K(x,c) := \int_0^1 k_i(x,c)di$. Remember that we must have $k_i(x,c) \in [0,\chi]$ for all agents i. $K(x,c)$ is the average refinancing intensity in the population, whenever the latent state is x for a mortgage with coupon rate c. The price $P\left(x,c;K\right)$ of a mortgage depends on this average refinancing strategy, via

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

with τ_K the prepayment time for a mortgage with prepayment intensity $\nu + K(x_t,c)$. Clearly $P\left(x,c;K\right)$ is increasing in c and continuous in c under certain technical conditions. The mortgage rate $m(\cdot;K)$ is then defined implicitly via

$$P\left(x,m(x;K);K\right) := 1+\pi$$

Households' life-time *cost function*, when playing strategy k, and when all other households play strategy K, is then

$$\begin{aligned} J(x,c;k,K) &= \mathbb{E}_x\left[\int_0^{+\infty} e^{-\rho t}\left(c_t^{(k)} + \psi_\chi dN_t^{(k)} + \psi_\nu dN_t^{(\nu)}\right)\right], \\ dc_t^{(k)} &:= \left(m\left(x_t;K\right) - c_{t-}^{(k)}\right)\left(dN_t^{(k)} + dN_t^{(\nu)}\right), \end{aligned}$$

with $N_t^{(k)}$ (resp. $N_t^{(\nu)}$) the counting process for refinancings (resp. moves) with intensity $k(x_t, c)$ (resp. ν). The household *value function* is then

$$V(x, c; K) := \inf_{k \in \mathcal{K}} J(x, c; k, K)$$

Consistent with Maskin and Tirole (2001), an MPE of this dynamic game can then be defined via a strategy K^* that solves

$$K^* := \arg\min_{k \in \mathcal{K}} J(x, c; k, K^*)$$

The formulation of the household cost function $J(x, c; k, K)$ in this game makes it clear that changes in the strategy K played by all the other households has an impact on household i's cost $J(x, c; k, K)$, irrespective of the strategy k such agent i plays. For instance, the more "optimal" the rest of households behave (via strategy K), the greater mortgage rates will be (via the function $m(\cdot; K)$), and the higher the cost function $J(x, c; k, K)$ will be, irrespective of the strategy k played by agent i. This is the key externality of this game: a given agent does not internalize the impact that its prepayment decisions have on mortgage rates, and thus on other agents' payoffs.

Note that absent gains on sale (i.e. when $\pi = 0$), investors are always willing to offer a mortgage at a finite interest rate, even when $\chi \to \infty$, i.e., a unique MPE exists even when $\chi = +\infty$. In that case, the externality discussed above disappears.

However, when gains on sale need to be positive (i.e. when $\pi > 0$), the externality leads to deadweight losses that have additional consequences. In that case, we must impose $\chi < \infty$ in order for an MPE to exist. Indeed, lenders are only willing to roll origination costs into a higher rate if probabilistically there is sufficient inactive households to allow them to break even. When $\chi \to \infty$, the infinite variation property of Brownian shocks leads to mortgages with zero duration at issuance. It is then impossible for lenders to recoup their origination costs via the secondary market income $\pi > 0$ given a zero-duration mortgage with a finite mortgage coupon. Consequently, the combination of $\chi = \infty$ and $\pi > 0$ results in the absence of an MPE.

C Proofs: policy evaluations

C.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 π. Indeed, note $P^*(x, c)$ the price of such a prepayable instrument with coupon c when the latent aggregate state is x:

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

where τ_ν is a Poisson time with arrival rate ν. 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

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

The related HJB, for $c \leq m^*(x)$, is

$$[\nu + r(x)]\, P^*(x,c) = c - f + \nu + \mathcal{L}P^*(x,c)$$

The boundary condition on the surface $c = m^*(x)$ is $P^*(x, m^*(x)) = 1 + \pi$, where we will call $m^*(x)$ the "auto-RM rate". The optimality condition for the stopping time τ takes the form of the smooth pasting condition

$$\partial_x P^*(x, m^*(x)) = 0$$

This problem is well defined. 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

$$\begin{aligned}\hat{P}(x,c;\chi) :&= \inf_{k \in \mathcal{K}_\chi} \mathbb{E}_x \left[\int_0^{\tau_k \wedge \tau_\nu} e^{-\int_0^t r(x_s)ds} (c - f)\, ds \right. \\ &\qquad \left. + 1_{\{\tau_k \leq \tau_\nu\}} (1+\pi) e^{-\int_0^{\tau_k} r(x_s)ds} + 1_{\{\tau_k \geq \tau_\nu\}} e^{-\int_0^{\tau_\nu} r(x_s)ds} \right] \\ &= \inf_{k \in \mathcal{K}_\chi} \mathbb{E}_x \left[\int_0^{+\infty} e^{-\int_0^t (r(x_s) + \nu + k_s)ds} (c - f + \nu + k_t(1+\pi))\, ds \right],\end{aligned}$$

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 τ_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.[47] 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

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

C.2 Auto-RM vs. short rates

Proof of Proposition 9. We consider the case $\pi \geq 0$ – i.e. the case where mortgage origination costs are potentially incurred, and recouped by lenders via higher mortgage rates. As discussed in Appendix C.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

[47] Given the nature of Brownian motions, these prepayable instruments have, at the time of issuance, zero duration.

the following set of inequalities

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

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. □

C.3 Auto-RM impact on initial debt-to-income ratio

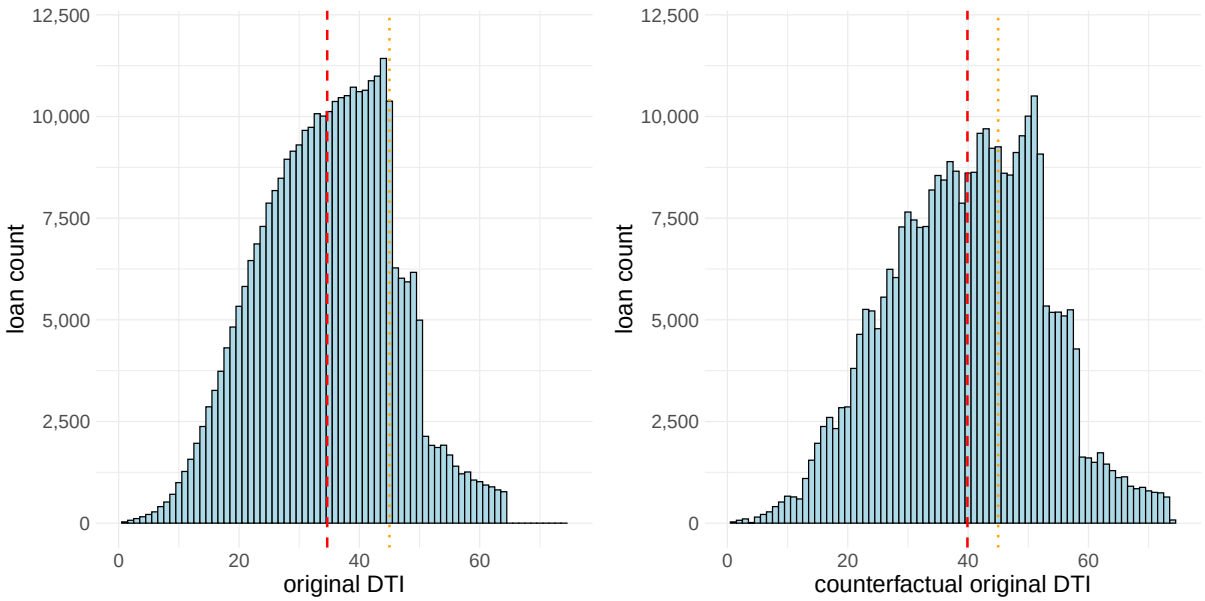

Figure C.1: **DTI distribution and counterfactual DTI distribution.** Left figure shows the DTI distribution in the SFLP data. Right figure shows the counterfactual DTI distribution if mortgage rates were higher than those actually realized, with a difference corresponding to the ergodic average difference between (a) mortgage rates in the Approximate Pooling MPE and (b) mortgage rates in the auto-RM equilibrium.

D Numerical method

D.1 Equilibrium mortgage rates

We use a finite-difference ("FD") approximation of the HJB (B.2) in order to compute the mortgage rate function $m(x)$ when the short rate follows a one-dimensional process. In that case, we have showed that optimal refinancing obtains whenever the household is attentive and $r_t \leq r^*$, where we have defined r^* implicitly via $c = m(r^*)$. In words, r^* is the rate that was prevalent at the time of the last refinancing event. Rather than defining the household state as (x, c), we will instead

focus on the monotone transformation (r, r^*). We discretize the short rate $r \in \{r_1, ..., r_{n_r}\}$ to n_r points, and attention types $\chi \in \{\chi_1, ...\chi_{n_\chi}\}$ to n_χ points, so that the state vector is $\left(r_i, r_j^*, \chi_k\right)$. Further, let $\overrightarrow{P}$ be the (stacked) vector of shadow mortgage prices on a mortgage to attention type χ household with coupon $m(r^*)$ at current interest rate r, where $m(r^*)$ is to be derived. Let $\mathbf{A}_r$ be the FD upwind scheme matrix representation of the linear generator for the short rate r_t (a tri-diagonal matrix), and $\mathbf{A}_{r^*}$ the corresponding matrix for r_t^* (utilizing the monotonicity of the pricing function, i.e., $m(r^*) > m(r) \iff r^* > r$).[48] Then, in this discretized space, we have

$$\underset{(n_r^2 n_\chi)\times(n_r^2 n_\chi)}{\mathbf{R}} \cdot \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{P}} = \underset{(n_r^2 n_\chi)\times n_r}{\mathbf{S}^m} \cdot \underset{n_r \times 1}{\overrightarrow{m}} + \underset{(n_r^2 n_\chi)\times(n_r^2 n_\chi)}{\mathbf{A}_r} \cdot \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{P}} + \underset{(n_r^2 n_\chi)\times(n_r^2 n_\chi)}{\mathbf{A}_{r^*}} \cdot \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{P}} \tag{D.1}$$

where $\mathbf{S}^m$ simply appropriately expands the applicable n_r coupons rates $m(r^*)$ summarized by the vector $\overrightarrow{m}$ to all the states of the system $\left(r_i, r_j^*, \chi_k\right)$. The break-even condition of our Approximate Pooling MPE with issuance distribution g is then

$$\underset{n_r \times (n_r n_\chi)}{\mathbf{G}} \cdot \underset{(n_r n_\chi)\times(n_r^2 n_\chi)}{\mathbf{S}^\pi} \cdot \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{P}} = (1+\pi) \underset{n_r \times 1}{\overrightarrow{1}} \tag{D.2}$$

where $\mathbf{S}^\pi$ selects the n_r states $\left(r_i, r_j^*\right)$ with $r = r^* \iff i = j$ out of all n_r^2 states $\left(r_i, r_j^*\right)$ for any χ,[49] while G is simply the discretized density over n_χ states of χ for all n_r states $\left(r_i, r_j^*\right)$ with $r_i = r_j^* \iff i = j$. In other words, this condition imposes that at issuance, given that one is facing a density g of types, the mortgage market value is equal to $1 + \pi$.[50]

Define $\mathbf{A} \equiv \mathbf{A}_r + \mathbf{A}_{r^*}$. Stacking the equations and rearranging, we have

$$\begin{aligned} (\mathbf{R} - \mathbf{A})\overrightarrow{P} - \mathbf{S}^m \cdot \overrightarrow{m} &= \overrightarrow{0} \\ \mathbf{G} \cdot \mathbf{S}^\pi \cdot \overrightarrow{P} &= (1+\pi)\overrightarrow{1} \end{aligned} \tag{D.3}$$

which can be written to

$$\begin{bmatrix} \underset{(n_r^2 n_\chi)\times(n_r^2 n_\chi)}{\mathbf{R} - \mathbf{A}} & \underset{(n_r^2 n_\chi)\times n_r}{-\mathbf{S}^m} \\ \underset{n_r \times (n_r^2 n_\chi)}{\mathbf{G} \cdot \mathbf{S}^\pi} & \underset{n_r \times n_r}{\mathbf{0}} \end{bmatrix} \begin{bmatrix} \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{P}} \\ \underset{n_r \times 1}{\overrightarrow{m}} \end{bmatrix} = (1+\pi) \begin{bmatrix} \underset{(n_r^2 n_\chi)\times 1}{\overrightarrow{0}} \\ \underset{n_r \times 1}{\overrightarrow{1}} \end{bmatrix} \tag{D.4}$$

Let

$$\underset{(n_r^2 n_\chi + n_r)\times(n_r^2 n_\chi + n_r)}{\mathbf{A}^{\pi m}} \equiv \begin{bmatrix} \mathbf{R} - \mathbf{A} & -\mathbf{S}^m \\ \mathbf{S}^\pi \cdot \mathbf{G} & \mathbf{0} \end{bmatrix} \tag{D.5}$$

Then, if $\mathbf{A}^{\pi m}$ is invertible, we can calculate shadow mortgage prices and equilibrium mortgage rates via a simple matrix inversion:

$$\begin{bmatrix} \overrightarrow{P} \\ \overrightarrow{m} \end{bmatrix} = (1+\pi)\left(\mathbf{A}^{\pi m}\right)^{-1} \begin{bmatrix} \overrightarrow{0} \\ \overrightarrow{1} \end{bmatrix} \tag{D.6}$$

[48] If the short-rate process is multi-dimensional, then $\mathbf{A}_{r^*}(m)$ is a function of m, and an iterative procedure is required to solve for the equilibrium $m(r)$.

[49] Essentially, in matrix form, *conditional* on χ, this is simply selecting the diagonal.

[50] Note that $\mathbf{G}$ can implement a state-dependent issuance distribution $G(r)$ but only allows for a one-step solution unless the resulting mortgage vector $\overrightarrow{m}$ is monotone. If not, a loop evaluating the American option pricing problem is required.

Let $\overrightarrow{g}$ be the constant $n_\chi \times 1$ density (with $\overrightarrow{g}^\top \cdot \overrightarrow{1} = 1$) and let $\overrightarrow{r}$ be the constant n_r discount rates. Order the system by (r, r^*, χ), we have

$$\underset{(n_r^2 n_\chi)\times(n_r^2 n_\chi)}{\mathbf{R}} = \underset{n_r n_\chi \times n_r n_\chi}{\mathbf{I}} \otimes \underset{n_r \times n_r}{\text{diag}}\left(\overrightarrow{r}\right) \quad \text{and} \quad \underset{n_r \times (n_r n_\chi)}{\mathbf{G}} = \underset{1\times n_\chi}{\overrightarrow{g}^\top} \otimes \underset{n_r \times n_r}{\mathbf{I}} \tag{D.7}$$

where $\otimes$ denotes the Kroenecker product. Meanwhile, $\mathbf{S}^m$ is a selector matrix made up of n_r columns of ones each of length $n_r n_\chi$, while $\mathbf{S}^\pi$ simply has 1's on the "diagonal" $r = r^*$.

D.2 Ergodic density

We next derive the ergodic density of our dynamic system. We leverage the fact that the adjoint of the linear generator for the stochastic process (r_t, r_t^*) is, when approximated via a FD scheme, simply the transpose of the matrix $\mathbf{A} = \mathbf{A}_r + \mathbf{A}_{r^*}$ as all entries of $\mathbf{A}$ are real. The ergodic density $\overrightarrow{f}_\infty$ is then solution to the Kolmogorov forward equation ("KFE")

$$\overrightarrow{0} = \mathbf{A}^\top \cdot \overrightarrow{f}_\infty \quad \text{s.t.} \quad \overrightarrow{1}^\top \cdot \overrightarrow{f}_\infty = 1 \tag{D.8}$$

Using an approach presented in Achdou, Han, Lasry, Lions, and Moll (2021), we simply pick an arbitrary reference state $\left(r_i, r_j^*, \chi_k\right)$ and allocate arbitrary probability mass to it, say $f_\infty\left(r_i, r_j^*, \chi_k\right) = 1$.[51] We can then simply change the first condition in (D.8) to reflect this change, by replacing the corresponding row of $\mathbf{A}^\top$ with a vector of zeros and a single one that represents the diagonal for $\left(r_i, r_j^*, \chi_k\right)$, and also replace the appropriate row of $\overrightarrow{0}$ with our arbitrary probability mass. We can invert the system and solve for an auxiliary density vector. This auxiliary vector, even though it gives the right relative weights to the different states versus the reference state $\left(r_i, r_j^*, \chi_k\right)$, is not guaranteed to fulfill the second condition in (D.8), i.e., to sum to 1. Thus, we need to normalize this auxiliary vector, which then finally yields $\overrightarrow{f}_\infty$.

E Maximum likelihood estimation

Probability per period. Take χ and dt as given, then the probability of refinancing in an effective period is given by

$$p = 1 - e^{-\chi dt} \iff \chi = -\frac{\ln(1-p)}{dt}.$$

We estimate the probability per period by using an observation for each individual

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

Under our assumption that attention times are independent and identically distributed, we have

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

[51] Here, the reference state can be arbitrarily chosen, as any point (r, r^*, χ) has strictly positive probability. This is an artefact of the inattention friction, with a small number of households facing long periods of inactivity. Without time-dependent inaction, we would have to pick a state for which the ergodic probability of being visited is strictly positive.

Thus, using N_sim individuals observed over the panel from 0 to T, we have N_sim x_i observations.

E.1 Clustering

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 χ.

Mazimizing the likelihood function for finite vector p. 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

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

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 $\mathbf{p}$ with possible values in $\mathbf{P}$ of length $N \leq N_{HH}$. We want to use MLE to estimate the optimal vector $\mathbf{P}$ (which implies an optimal vector $\mathbf{p}$) 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

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

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 $\mathbf{s}$ and $\mathbf{t}$, we want to estimate the maximum likelihood that those observations were generated by a vector $\mathbf{p} = [p_1, ..., p_N]$ where $p_i \in \mathbf{P}$ with $\mathbf{P} \in [0,1]^N$, i.e., having $N < N_{HH}$ entries. For a given $\mathbf{P}$, the log-likelihood (omitting the binomial coefficient term) is given by

$$L(\mathbf{P}) = \sum_{i=1}^{N_{HH}} \max_{p\in\mathbf{P}} \{s_i \log p + (t_i - s_i)\log(1-p)\} \tag{E.1}$$

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

Numerical implementation We now want to maximize over the choice of $\mathbf{P}$, that is,

$$\max_{\mathbf{P}\in[0,1]^N} L(\mathbf{P})$$

The maximization procedure requires $p_i \in [0,1]$, and thus $\mathbf{P} \in [0,1]^N$. Write $p = 1 - e^{-\tilde{\lambda}}$, and $\tilde{\lambda}_n = \sum_{i=1}^n \lambda_i$, so that $\lambda_i = \tilde{\lambda}_i - \tilde{\lambda}_{i-1}$ with $\tilde{\lambda}_0 = 0$ is the difference between subsequent $\tilde{\lambda}_i$'s. Wlog, we order the $\tilde{\lambda}_i$'s to be increasing, so that $\lambda_i \geq 0$. Thus, for a vector of $\tilde{\boldsymbol{\lambda}}$, we recover $\mathbf{P}$ by observing

that

$$\tilde{\boldsymbol{\lambda}} = \mathbf{L} \cdot \boldsymbol{\lambda} \quad \text{where} \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & 0 \\ 1 & & \ddots & 0 \\ 1 & \cdots & 1 & 1 \end{bmatrix}$$

The advantage of this expression is that $\lambda_i \in [0, \infty)$, so traditional optimization solvers can be used as each λ_i satisfies box constraints, yielding a unique solution due to the fact that $\tilde{\lambda}_i$ is increasing in i. Thus, defining $\ell(\boldsymbol{\lambda}) \equiv L\left(1 - e^{-\mathbf{L}\cdot\boldsymbol{\lambda}}\right)$, we write the maximization as

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} \ell(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^N} L\left(1 - e^{-\mathbf{L}\cdot\boldsymbol{\lambda}}\right)$$

For high-dimensional $\mathbf{P}$, evaluating $\max_{p \in \mathbf{P}} \{s_i \log p + (t_i - s_i) \log(1-p)\}$ for all N values of $\mathbf{P}$ might be expensive. However, note that

$$f_i(p) := s_i \log p + (t_i - s_i) \log(1-p)$$

is well behaved, i.e., single-peaked and globally concave. Thus, picking the maximum of $f_i(p)$ at the closest values of $\mathbf{P}$ to the right and left of $\hat{p}_i = \frac{s_i}{t_i}$ suffices for each i.

E.2 Alternative empirical MLE specifications

Here, we present alternative MLE specifications, in particular different cuts of the data w.r.t. the required refinancing gap. The main group-based specification with $gap > 0.5\%$ can be found in Table 1. For the alternative specifications, Table E.1 shows the MLE specification for $gap > 0\%$, and Table E.2 shows the MLE specification for $gap > 1\%$. Finally, the above median splits w.r.t. FICO ("FICO+") and loan balance ("Loan+") are show in Table E.3.

Spec: weighted by avg loan, $gap > 0\%$

χ	$p(\chi)$	StDev p	$H(\chi)$
0.0	0.0	0.0006	0.788
0.1799	0.0147	0.0001	0.064
0.3953	0.0323	0.0002	0.08
0.8517	0.685	0.0006	0.046
2.4855	0.1871	0.0024	0.021

Table E.1: 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 $\overline{\chi} = 13.53\%$.

Spec: weighted by avg loan, $gap > 1\%$			
χ	$p(\chi)$	StDev p	$H(\chi)$
0.0	0.0	0.0009	0.857
0.2845	0.0234	0.0003	0.043
0.753	0.0608	0.0006	0.054
2.0478	0.1569	0.0018	0.031
7.6738	0.4724	0.0063	0.015

Table E.2: 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 $\overline{\chi} = 22.80\%$.

Spec: weighted by avg loan, $gap > .5\%$					
	FICO+			Loan+	
χ	$H(\chi)$	$G(\chi)$	χ	$H(\chi)$	$G(\chi)$
0.0	0.763	0.551	0.0	0.784	0.569
0.2343	0.078	0.104	0.2343	0.057	0.076
0.5627	0.089	0.157	0.5627	0.083	0.148
1.3882	0.054	0.129	1.3882	0.057	0.138
5.2775	0.017	0.059	5.2775	0.019	0.069

Table E.3: 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 $\boldsymbol{\chi}$ vector constant at the baseline MLE. The average attention rate for FICO+ is $\overline{\chi} = 23.01\%$ under $H(\chi)$ and $\overline{\chi} = 60.28\%$ under $G(\chi)$, for Loan+ is $\overline{\chi} = 24.11\%$ under $H(\chi)$ and $\overline{\chi} = 65.59\%$ under $G(\chi)$.

E.3 Parametric distribution estimation

Instead of a finite number of groups, here we assume some parametric density $f(\cdot;\theta)$ for χ or $p(\chi) = 1 - e^{-\chi dt}$, with θ possibly multi-dimensional.

E.3.1 Maximizing parametric likelihood function over probabilities p.

For analytical tractability, we will directly parameterize the distribution of p via $f(p;\theta)$, where $p \in [0,1]$. Then, the likelihood for a given observation i, i.e., $(s_i, t_i) = (s,t)$ is given by

$$L(t,s;\theta) = \int_0^1 \underbrace{f(p;\theta)}_{\text{Prob of type } p \text{ given } \theta} \times \underbrace{\binom{t}{s} p^s (1-p)^{t-s}}_{\text{Prob of } s \text{ out of } t \text{ when type is } p} dp$$

Special case: beta distribution. Suppose we parameterize p via a beta distribution. Let $\theta = (\alpha, \beta)$ with

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where $\Gamma(n)$ is the Gamma function. Note the beta distribution directly applied to p nests the exponential distribution w.r.t. χ with parameter $\hat{\theta}$ above when $(\alpha, \beta) = \left(1, \hat{\theta}/dt\right)$.[52] An observation (t, s) being generated by a distribution with parameters (α, β) has the likelihood

$$\begin{aligned}\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \binom{t}{s} p^s (1-p)^{t-s}\, dp &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{t}{s} \int_0^1 p^{s+\alpha-1}(1-p)^{t-s+\beta-1}\, dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{t}{s} \frac{\Gamma(s+\alpha)\Gamma(t-s+\beta)}{\Gamma(\alpha+t+\beta)} \\ &= \binom{t}{s} \frac{B(s+\alpha, t-s+\beta)}{B(\alpha,\beta)},\end{aligned}$$

resulting in a Beta-Binomial distribution where $B(\alpha, \beta)$ is the Beta function. Thus, the log-likelihood for an observation i, i.e., $(s_i, t_i) = (s, t)$, is given by

$$\mathcal{L}(t, s; \theta) = \log \binom{t}{s} + \log B(s+\alpha, t-s+\beta) - \log B(\alpha, \beta)$$

Ignoring the parts unaffected by (α, β), i.e., $\log \binom{t}{s}$, the log-likelihood of the data is

$$L(\theta) \equiv \sum_{i=1}^{N_{HH}} \left[\log B(s_i+\alpha, t_i-s_i+\beta) - \log B(\alpha,\beta)\right] \tag{E.2}$$

F Conditional origination distribution $G(\chi, r)$ and gain on sale π

This appendix presents the main results of the paper assuming the (i) *conditional* origination distribution $G(\chi, r)$ as depicted in the right panel of Figure 4 and (ii) a possibly different gain on sale π. As the following results show, the state-dependence of the distribution does not impact the main results of the paper in any substantive way, while variations in gain on sale pass through to the equilibrium mortgage rate in predictable ways, with higher impact in high interest-rate environments. The only substantive change comes from $\pi = 0$'s impact on the ergodic average rates, in which case the separating MPE per type is mildly downward rather than strongly upward sloping as it was for $\pi = 4.6pts$. The reason is that for $\pi > 0$ more attentive types by nature impose higher deadweight cost of refinancing, which is passed on through the rate.

[52] Translated to a density over χ, we have, for $\theta = (\alpha, \beta)$,

$$\hat{f}(\chi; \theta) = f(1-e^{-\chi dt}; \theta) \cdot dt \cdot e^{-\chi dt} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(1-e^{-\chi dt}\right)^{\alpha-1} \left(e^{-\chi dt}\right)^{\beta} dt.$$

Note that for $\alpha = 1$, we have $\frac{\Gamma(\beta+1)}{\Gamma(1)\Gamma(\beta)} = \beta$, so that $\hat{f}(\chi; \theta)\Big|_{\alpha=1} = \beta dt \cdot e^{-\chi\beta dt}$ is the density of the exponential distribution over χ with parameter $\hat{\theta} = \beta dt$. Finally, note that $\log \beta = -\log B(1, \beta)$.

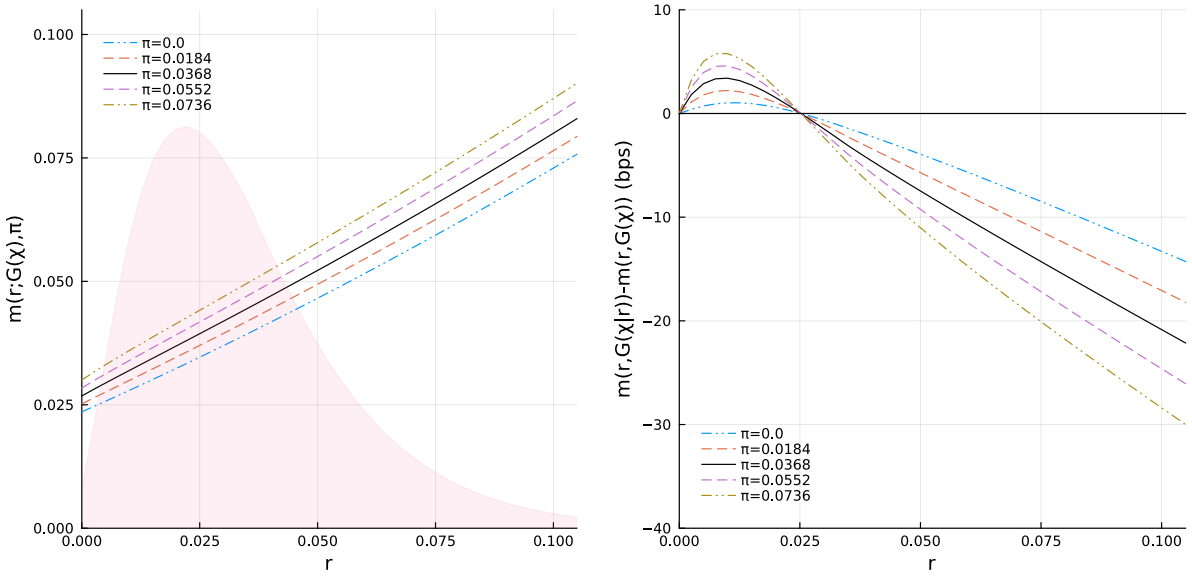

Figure F.1: **Equilibrium mortgage rates under different** π: Left Panel: Unconditional origination distribution $G(\chi)$ mortgage rates $m(r, G; \pi)$ for different gain on sale π, with the solid black line being the baseline. Right Panel: Difference between the conditional origination distribution $G(\chi|r)$ and unconditional origination distribution $G(\chi)$ (as used in main part of paper) equilibrium mortgage rates, $m(r, G(\chi|r)) - m(r, G(\chi))$.

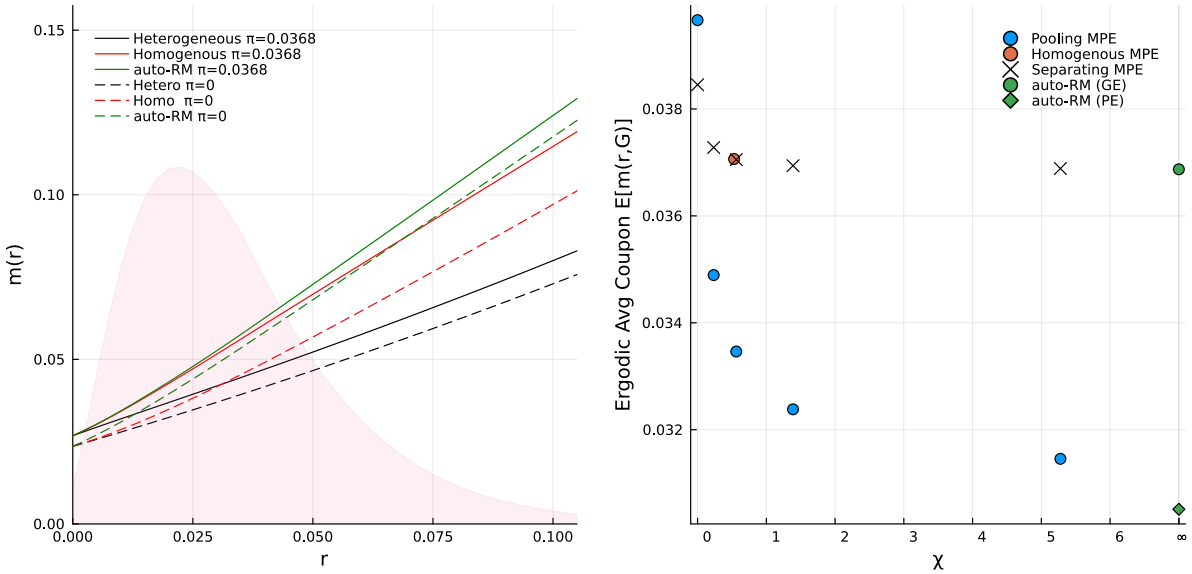

Figure F.2: **Equilibrium mortgage rates and ergodic average coupons under** $\pi = 0$: Left Panel: Solid (dashed) lines depict the approximate pooling equilibrium (black line), the homogenous equilibrium (red line), and auto-RM (green line) for the baseline $\pi = 80\% \times 4.6pts$ (zero gain $\pi = 0$) setting. Right Panel: Ergodic average coupons under no gain on sale $\pi = 0$.