Do Lenders Still Discriminate? A Robust Approach for Assessing Differences in Menus

David Hao Zhang Paul Willen*

October 13, 2020

Click for most recent version

Abstract

In the US, borrowers can choose to get lower interest rates on their mortgages by paying more discount points upfront. It is sometimes observed that minority borrowers pay the same average number of points as white borrowers conditional on the interest rate, but pay higher average interest rates than white borrowers conditional on points. How then can researchers tell if lenders charge minority borrowers more than observationally similar white borrowers? We show that the answer to this question involves a "menu problem," and intuitively appealing metrics of lender discrimination used in the literature can lead to false positives, false negatives, and contradictory results due to the potential hetereogeneity in borrower preferences across racial groups. This menu problem extends well beyond the mortgage market, but is often under-appreciated. To address this problem, we define (i) a new, robust, test statistic for equality in menus and (ii) a new difference in menus (DIM) metric for assessing whether one group of consumers would like to switch to another groups' menus, both based on pairwise dominance relationships in the data. We show how these metrics can be computed using methods from optimal transport, and devise a procedure for hypothesis testing in this class of problems based on directional differentiation. We implement our metrics on a new dataset matching 2018-2019 Home Mortgage Disclosure Act (HMDA) data to OptimalBlue rate locks. We find that Black and Hispanic borrowers were offered worse menus in terms of rates and points by lenders compared to Non-Hispanic White borrowers for conforming mortgages. Furthermore, the difference we detect is particularly concentrated among the more creditworthy borrowers.

*David Zhang is at Harvard Business School, dzhang@hbs.edu. Paul Willen is at the Federal Reserve Bank of Boston, paul.willen@bos.frb.org. We thank Isaiah Andrews, John Campbell, Edward Glaeser, Robin Lee, Ariel Pakes, Adi Sunderam, and Elie Tamer for their continuous advice on this paper. We are further indebted to Alex Bell, Neil Bhutta, Gary Chamberlain, Chris Foote, Andreas Fuster, Camilo Garcia-Jimeno, Aurel Hizmo, Jun Ishii, Larry Katz, Jonathan Roth, Andrei Shleifer, attendees of the 2020 System Applied Micro Conference and the 2020 System Econometrics Conference for their valuable comments. All errors are our own. The views expressed in this paper are solely those of the authors and not necessarily those of the Federal Reserve Bank of Boston nor the Federal Reserve System.

1 Introduction

Whether mortgage lenders discriminate against minority borrowers is an important question both in terms of academic research and in regards to its policy relevance.[1] However, the task of assessing whether lenders discriminate by offering minority borrowers worse prices relative to similar Non-Hispanic White borrowers in US mortgage markets is complicated by the fact there are two dimension to mortgage pricing: the interest rate and the amount of upfront fees charged by the lender. In particular, US mortgage borrowers can choose to pay more upfront fees, referred to as paying discount points in the industry, in return for lower interest rates, or conversely get the lender to pay some of their closing costs in exchange for a higher interest rate. We show that the availability of this choice between a higher upfront fee and a higher interest rate makes the detection of lender discrimination non-trivial, with the existing methods implemented in the large literature in this field susceptible to false and contradictory results. We then propose a novel identification argument and a new procedure for inference to deal with this menu problem, and apply it empirically to a new dataset to re-assess racial discrimination in mortgage markets.

In terms of background, many studies have found that minority consumers pay higher interest rates than observationally similar white consumers in the mortgage market.[2] While this can be interpreted as evidence that lenders systematically discriminated against minority borrowers by offering them worse pricing on their mortgages, another potential explanation is that minority consumers were simply more constrained in their choices of how much discount points to pay and how much lender credit, or negative discount points, to receive. The discount points explanation may still reflect structural disparities between racial groups, but has very different policy implications than one in which the lenders themselves are systematically offering minority consumers worse menus of rates and discount point options compared to white borrowers. Given data on rates and points (but not the menus borrowers faced, which are not typically available), our objective is to examine whether lenders discriminated

[1] Since the financial crisis, the Department of Justice obtained settlements of well over $500 million for overcharging Black and Hispanic borrowers in violation of the Fair Lending Act, as explained in Bhutta and Hizmo (Forthcoming), including $335 million for Bank of America (on behalf of Countrywide), $175 million for Wells Fargo, and $55 million for JP Morgan Chase. On Jun. 12, 2019, Sen. Elizabeth Warren wrote on twitter: "For generations, lenders have given African American & Latino families fewer loans at worse terms than similar white borrowers. Tech alone won't fix the problem. A new analysis found that discrimination is hardwired into lending algorithms. I want answers." `https://twitter.com/senwarren/status/1138909674781237253`.

[2] See e.g. Black and Schweitzer (1985), Boehm, Thistle, and Schlottmann (2006), Bocian, Ernst, and Li (2008), Ghent, Hernández-Murillo, and Owyang (2014), Cheng, Lin, and Liu (2015), Bartlett et al. (2019). Relatedly, Munnell et al. (1996) and Tootell (1996) finds that minority borrowers are more likely to be rejected for mortgages, and Black, Boehm, and DeGennaro (2003) finds that minority borrowers pay higher yield spreads when refinancing their mortgage.

against Black borrowers in the sense of offering them a worse distribution of menus compared to observationally similar white consumers, which we call *discrimination in menus.*

Our first contribution is to show that there exists a surprisingly non-trivial menu problem involved with assessing differences in the distribution of menus offered to minority and Non-Hispanic White borrowers when minority and Non-Hispanic White borrowers are likely to make different decisions (ie. have different preferences) over menu items. As an example, researchers may be tempted tempted to address a heterogeneity in preferences over points by racial group by "controlling" for points in their analyses. However, we show such regressions can give to contradictory results depending on which direction the regression is run. For example, it is possible that Black borrowers pay a higher interest rate than Non-Hispanic White borrowers controlling for points but pay the same number of points as White borrowers controlling for interest rate.[3] In addition, the menu problem goes far beyond this reverse regression problem. We show that even if the forward and reverse regression results were consistent, the approach of controlling for rates and points can lead to false positives in the sense of detecting discrimination when none exist, and false negatives in the sense of failing to detect discrimination when it does exist. Finally, we show that even the seemingly foolproof comparisons of means, i.e. checking if minority consumers on average pay both a higher interest rate and more discount points (or, generally, adjusting for a range of rate-point trade-off from external sources), can lead to false positives and false negatives if interpreted as evidence of discrimination in menus.

The menu problem we point out is important for the mortgage discrimination literature. There are two main existing methods by which a long literature has assessed discrimination in mortgage pricing given the rate-point trade-off. First, Courchane and Nickerson (1997) and Bhutta and Hizmo (Forthcoming) looks at whether black borrowers paid more in points conditional on rate in samples of FHA mortgages, for which Courchane and Nickerson (1997) found a differential in points paid by race while Bhutta and Hizmo (Forthcoming) did not. This difference in results based on a similar methodology may be explained changes in the patterns of the data over time and by the fact that Bhutta and Hizmo (Forthcoming) significantly improved upon the literature by using a much larger sample and by constructing a more uniform sample of loans.[4] Second, Woodward (2008), Woodward and Hall (2012), Bartlett et al. (2019) compares the interest rate of minority and White borrowers after adjusting for points using a known range of rate-point trade-offs, and found that minorities consistently paid more for mortgages, even in the FinTech era as shown in Bartlett et al.

[3]This problem is explored in the labor context in Goldberger (1984).

[4]Bhutta and Hizmo (Forthcoming) also has other findings, including that lenders receive more revenue from loans that were made to minorities and that the Black-White differences in rate and discount point can be rationalized by menus all having the same slope of +0.37% in rate per point.

(2019). As we discussed, both of the existing methods used in the literature, i.e. (i) controlling for points, and (ii) adjusting by a known range of slopes, can lead to false positives and false negatives and can contradict one another. Our robust solution to the menu problem would therefore allow researchers to assess discrimination in mortgage markets in a more consistent and theoretically sound way.

The menu problem also extends well beyond the mortgage discrimination setting. For example, when workers make decisions that trades off wages and hours worked, researchers may wish to assess whether the fact that men work longer hours than women can explain the gender differences in wage-hour outcomes as in the model of Goldin (2014). The menu problem implies that popular measures of gender inequality such as the the gender pay gap conditional on hours, or even the gender pay gap after adjusting for all relevant average compensating differentials, is not necessarily informative about whether the data on wages can be explained by heterogenous preferences over hours worked. Generally speaking, the menu problem is relevant whenever we wish to assess equality in opportunity given data on outcomes while allowing for heterogeneous preferences across groups.[5] Therefore, the problem that we point out and the solution we propose may be of broad interest.

As a solution to the menu problem, we propose (i) a new metric for detecting whether there exists a difference in the distribution of menus offered to two groups and (ii) a new lower bound measure for assessing differences in menus (DIM) for the extent to which one group of consumers would like to switch to another group's menus. Both metrics are based on pairwise dominance relationships in the data (i.e. getting a lower rate and paying fewer points dominates that with higher rates and higher points) which can be supplemented by by industry knowledge. Based on these pairwise relationships, we ask the question of whether the data *can* be rationalized by a model of equality in menus but heterogeneity in preferences, and if not we compute an average difference in menus perceived by one group of consumers when switching to another group's menus. Unlike the heuristic approaches used in the literature, our metrics are robust to unobserved differences in preferences across borrower groups which manifests in their choices. The sample counterparts to both of our metrics can be computed as solutions to optimal transport problems, which are computationally well-understood and can be efficiently computed through linear programming.

Furthermore, as a technical contribution we also derive a new approach to uniformly valid inference for the value of optimal transport problems, which we implement for our metrics to distinguish between statistical noise and actual differences in menus. Conventional

[5]The problem of unobserved choice sets also appears in demand estimation, as surveyed in Crawford, Griffith, and Iaria (2019). However, the problem is different in that here we are not concerned with demand estimation, and are instead focused on assessing whether two consumer groups could have faced the same distribution of menus based on data on their choices.

approaches to inference such as bootstrapping fails for optimal transport problems because the objective function can be non-differentiable (Fang and Santos, 2018). We first prove that optimal transport problems are directionally differentiable in the sense of Shapiro (1991) and Fang and Santos (2018). We then apply the asymptotic results of Fang and Santos (2018), which we combine with a Bonferroni correction following Romano, Shaikh, and Wolf (2014) and McCloskey (2017) to address sampling error in the directional derivatives. We show that this approach leads to asymptotically uniformly valid size control for hypothesis testing in the value of optimal transport problems, and test it in finite samples in a Monte Carlo simulation. Our new approach to inference in optimal transport may be useful for other researchers that may wish to conduct inference on the value of optimal transport problems, many of which are described in Galichon (2016).

Empirically, we use our metrics to assess racial discrimination in the 2018-2019 Home Mortgage Disclosure Act (HMDA) data matched to Optimal Blue rate locks. We show that we can detect inequality in menus offered by the same lender and county and within narrow covariate groups for conforming mortgages for both Black and Hispanic borrowers. Furthermore, we show that the on average Black borrower getting conforming mortgages would be willing to increase their interest rate by at least 2.0 basis points in order to switch to the menus of non-Hispanic White borrowers. Similarly, Hispanic borrowers are on average willing to pay 1.5 basis points in interest rate in order to switch menus with non-Hispanic White borrowers. Our finding that racial differences in lender pricing remains relevant for conforming mortgages is consistent with Bartlett et al. (2019), although the amount of interest rate discrimination we detect is smaller in magnitude. On the other hand, we do not detect interest rate discrimination in FHA mortgages, which is consistent with Bhutta and Hizmo (Forthcoming). Finally, the discrimination we detect is concentrated among borrowers with lower loan-to-value (LTV) ratios and higher credit scores.

As a point of clarification, throughout the paper we adopt the recent mortgage discrimination literature's definition of discrimination which may be different from how the term is used in other contexts. By discrimination, we mean lenders offering a different distribution of menus of rates and points for observably similar minority and Non-Hispanic White borrowers. Importantly, this captures not only differences that result from taste-based discrimination, but also disparate impact due to unobservables. For example, the act of giving more discretionary discounts to borrowers that are perceived have lower search frictions may be characterized as discrimination if minority borrowers are disproportionately unable to receive these discounts. We have two reasons for adopting this rather broad definition of discrimination. First, as explained in Bartlett et al. (2019), any systematic pricing differences between racial groups that do not reflect creditworthiness, which for conforming mortgages is largely determined by the pricing of mortgage insurance by Government Sponsored Enterprises (GSEs) based on a set of underwriting variables we observe, is likely illegal under the Fair Housing Act. Second, the process of mortgage pricing in the US is based on electronic ratesheets that take our observables as inputs and generates a range of rates and points as outputs, after which the loan originator may offer further discounts. Therefore, there is no clear channel by which lenders can use variables we do not observe to further price credit risk: it is not in the job descriptions of US mortgage loan originators who interacts with borrowers to go beyond the underwriting variables used in the ratesheets to subjectively price credit risk, and therefore any disparate impact from loan originator discounting based on the unobservables that they see might be considered discrimination from a regulatory point of view.[6]

The rest of this paper is structured as follows. Section 2 explains the motivation of our paper by exploring why heuristic approaches to analyzing discrimination in menus may be misleading. It also provides intuition for our approach. Section 3 formally defines our metrics for assessing discrimination in menus. Section 4 describes a methodology for conducting inference on our metrics. Section 5 shows our data and empirical results. Section 6 concludes.

2 Motivation and Intuition

In this section we discuss why intuitively appealing approaches for assessing discrimination in menus may be misleading, and provide intuition for our test of inequality in menus. By way of background, there are two dimensions of pricing for mortgages in the US, an upfront fee/discount points and the interest rate, where each point is customarily worth 1% of the loan amount. Consumers can have the option of picking a particular rate and discount point combination that best suits their preferences and financial constraints, the choices of which from an example ratesheet we plot in Figure 1. In particular, borrowers can pay discount points to reduce their interest rate, or receive money from the lender to help cover their closing costs by getting lender credit (paying negative points). The sense in which we think about lender discrimination in menus, then, is for minority borrowers to receive a worse rate-point schedule than white borrowers.

[6] Interestingly, insured mortgages from Black and Hispanic borrowers are likely more valuable to lenders from a cash flow perspective due to their significantly lower prepayment risk, as shown in Kau, Fang, and Munneke (2019) and Gerardi, Willen, and Zhang (2020). Therefore, even if lenders did have a channel to subtly use unobservables that are correlated with race to tweak ratesheets based on expected loan performance, it is not clear that it would necessarily disfavor minorities as we find in our empirical results. Furthermore, our paper can be interpreted as assessing "differences in menus" if the reader does not accept our definition of discrimination.

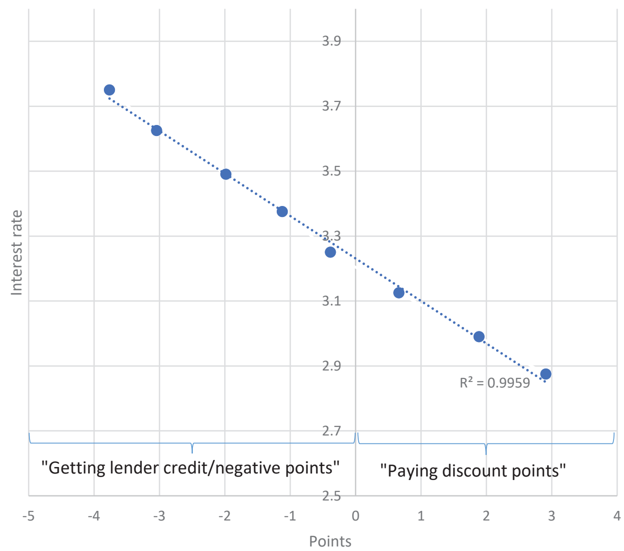
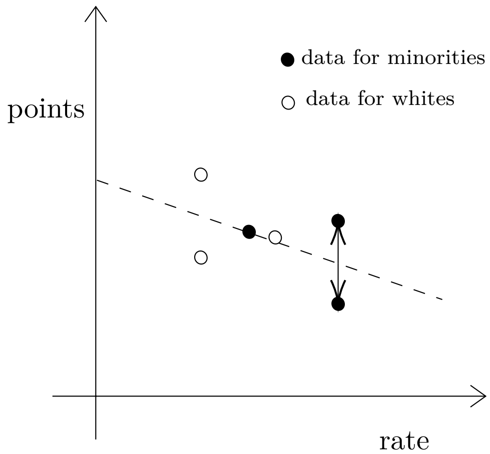


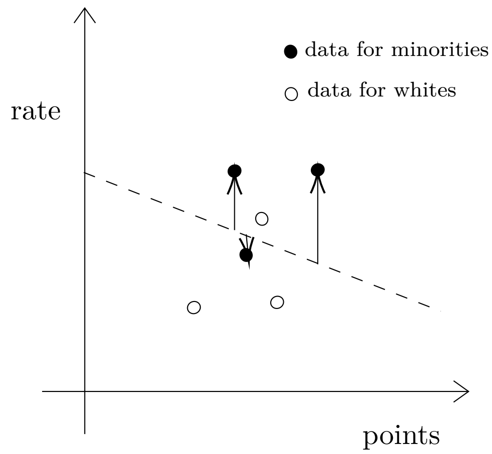
Figure 1: An example set of menu items from a lender ratesheet.

One natural heuristic approach to assessing whether lenders offered minority and Non-Hispanic White borrowers different menus is to control for one dimension of the menu. That is, for groups of borrowers from different races that are otherwise similar, estimate whether minority borrowers who received the same interest rate as white borrowers paid more points. This approach was used in Courchane and Nickerson (1997) and Bhutta and Hizmo (Forthcoming). Relatedly, one may also look at whether minority borrowers who paid the same number of points as white borrowers received higher interest rates.

A danger of the controlling for one dimension of the menu approach of assessing discrimination, however, is that it can lead to contradictory estimates depending on which item you choose to control for. For example, the situation in Figure 2 shows that it is possible for a regression of points on rate to show no discrimination against minorities while for an regression of rate on points to show discrimination with the same example data. In this figure, we represent example data from minority and white borrowers using black and white dots, respectively, regression line by the dash line, and the difference to the regression line by the solid arrows. Figure 2a shows that a regression of points on rate and borrower race would show a zero coefficient for minorities with the two arrows balancing each other out. On the other hand, Figure 2b shows that, using the same data, a regression of rate on points would show give a positive coefficient for minorities instead. This sort of contradiction is not particular to the linear regression case, and as we show in Appendix Figure A.1 can appear with general conditional expectations. This possibility for contradiction can be viewed as a version of the reverse regression problem of Goldberger (1984). While the reverse regression problem illustrates how the heuristic of conditioning on one dimension of the menu can be unreliable, our menu problem goes far beyond the reverse regression problem since there can be false positives and false negatives even when the forward and reverse regressions are consistent (and even when a simple comparison of means are consistent), which we will discuss in the rest of this section.



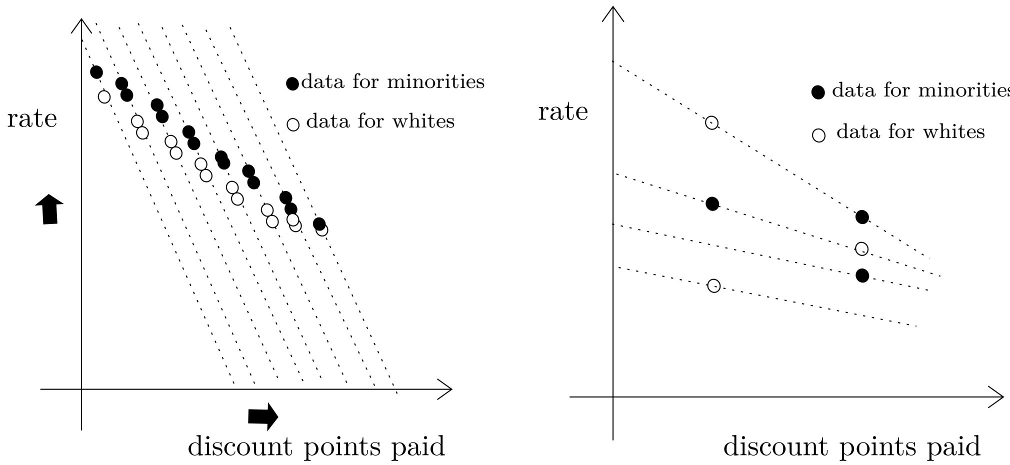
(a) Points on rates shows no discrimination



(b) With the same data, rates on points (flipped axes) shows discrimination

Figure 2: How the choice of which menu dimension to control for can lead to contradictory findings of discrimination

Figure 3 shows how even when forward and reverse regressions consistently detect discrimination or no discrimination, the heuristic of controlling for one dimension of the menu can lead to false positives and false negatives. In this figure, we represent example data from minority and white borrowers using black and white dots, respectively, and menus by dotted lines. The researcher wishes to evaluate whether the minority and white borrowers were offered the same distribution of menus. In the left panel, Figure 3a shows a false positive situation in which minority borrowers paid more in rate controlling for points, and more points controlling for rate, even though lenders offered both minority and white consumers the same distribution of menus. The only difference driving this result is that minority consumers chose to pay fewer points on every menu. In the right panel in Figure 3b, we illustrate a false negative situation in which minority consumers paid the same rate conditional on points but faced a worse distribution of menus since the bottom menu (the most advantageous menu) were only offered to white borrowers while the second to bottom menu were only offered to minority borrowers. Furthermore, Figure 3b can be constructed in a way such that the variance of the rate and discount points paid are equal, such that forward and reverse regressions both gives the same false negative but there does exist discrimination in menus.



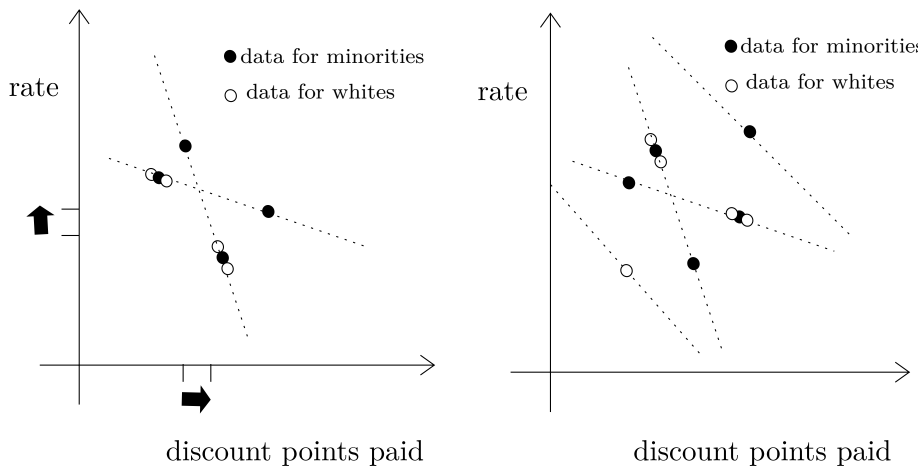
(a) **False positive**, minority borrowers paid more in rate (points) controlling for points (rate) but the menus were the same

(b) **False negative**, minority borrowers paid the same average rate controlling for points as white borrowers, but their menus were worse

Figure 3: False positives and false negatives from controlling for one direction of the menu

Intuitively, the assumptions underlying the approach of assessing discrimination in menus by looking at the average deviation to a regression line for the racial groups are that (i) all menus share the same shape, with unobserved heterogeneity in menus being due to an additive error term, and (ii) that this shape can be correctly estimated. The contradictory findings in Figure 2 resulted from the estimated slope of the menus being different depending on which regression is being run, thus breaking assumption (ii). In the situations in Figure 3a where consumers react the underlying heterogeneity in menus by changing their choices, a classic omitted variables bias emerges, again breaking assumption (ii).

A second intuitively appealing and seemingly surefire approach for assessing discrimination is to compare means, thus avoiding the problem of having to estimate menu slopes from the data. In other words, the researcher may wish to check if minority consumers paid more on average in both rates and discount points, such that they are disadvantaged in both dimensions, compared to observationally similar white borrowers. A variation of this approach is to take a pre-defined range of rate-point trade-offs as the slope estimated from external sources, which was done in Woodward (2008), Woodward and Hall (2012), and Bartlett et al. (2019). While this avoids the problem that regressions may incorrectly estimate menu slopes, it can still lead to false positives and false negatives when slopes are not constant across menus, thus breaking the assumption (i) that all menus share the same shape. This is a realistic problem because we know from ratesheet data that an unobserved heterogeneity in slopes does exist in the mortgage setting as the rate-point trade-off do vary substantially across lenders and over time (Figure 8). When slopes differ across menus, Figure 4a illustrates how a false positive in which minority consumers pay more on average in terms of both rates and points but faced the same distribution of menus as white borrowers can occur, and Figure 4b illustrates a false negative possibility in which minority borrowers paid the same average rates and points as white borrowers but did face worse menus.



(a) **False positive**, minority consumers paid more on average in both rates and points but their menus are the same

(b) **False negative**, minority consumers pay the same average rate and points as white borrowers, but their menus were worse

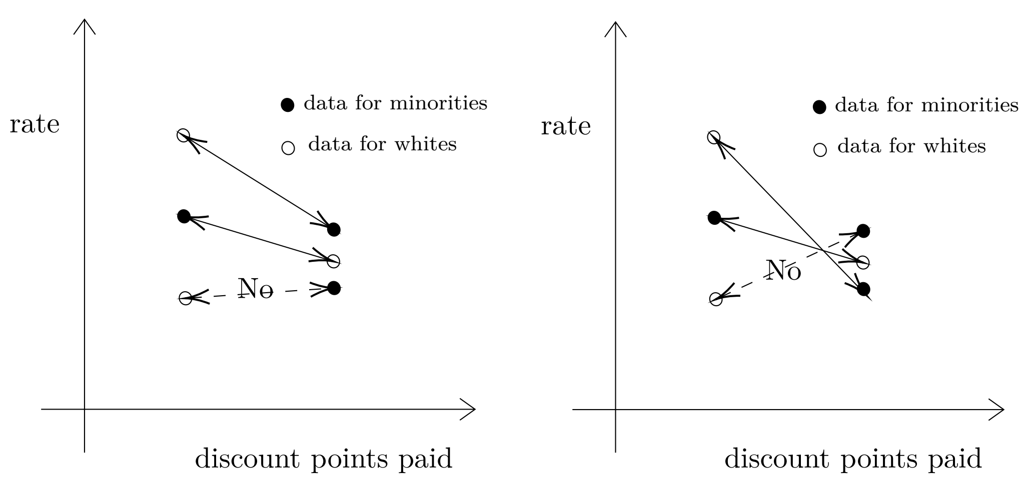
Figure 4: False positives and false negatives from checking if minority borrowers paid more on average in both rates and points

More specifically, the mechanism for when even a simple comparison of means would lead to false positives, as in the case of Figure 4a is that, in the example, minority borrowers respond less to differences in the slopes of the rate-point menus compared to white borrowers. This can be possible because there are two directions in which constraints can drive borrower choices of mortgage points. First, if borrowers are cash constrained when getting the loan, they may prefer to pay fewer points (or, get more lender credit/negative points) and get a low closing cost mortgage regardless of what rate-point tradeoffs lenders offer. Second, if borrowers are debt-to-income (DTI) constrained, they may need to pay more points to buy down the rate so as to be able to borrow more.[7] Therefore, the simple observation that minority borrowers on average pay more than white borrowers may reflect the fact that minority borrowers are more constrained in their choices, which is a form of "disadvantage" that is not necessarily due to lenders discriminating against them by offering them different menus.

[7]See e.g.: `https://www.thetruthaboutmortgage.com/dti-debt-to-income-ratio/`.

As we explained, both (i) controlling for a dimension of the menu and (ii) comparing means across borrower groups for across all menu dimensions, which are the existing methods used in the literature for assessing mortgage discrimination, can lead to false positives and false negatives as well as contradictory estimates. To address this problem, we define alternative metrics for assessing discrimination in menus based on whether the data *can* be rationalized by a model in which all groups of borrowers faced the same distribution of menus. Ignoring for now sampling error to build intuition, the common thread in the false positive situations of Figure 3a and 4a is that there exists a common distribution of menus underlying both minority and white choices, in the sense of there being a possible one-to-one match between minority and white borrowers where within each match both the borrowers' choices could have come from the same menu. This is the criteria we use for assessing equality in menus.

Our methodology can detect discrimination in menus where the existing heuristic approaches to the menu problem lead to false negatives. In particular, we illustrate in Figure 5 how the situation of Figures 3b in which regressions controlling for either rates or points would show a false negative but the data fails our one-to-one matching condition under the assumption that borrowers with pairwise strictly dominated choices (ie. paying more in terms of both rates and points) could not have shared menus with one another. In other words, there is no way to construct a common distribution of menus for minority and white borrowers that explains the data. Analogously, the false negative example from comparing means in Figure 4b also fails our criterion.



(a) Possibility #1 for matching fails (b) Possibility #2 for matching also fails

Figure 5: How data from Figure 3b fails a "perfect matching" condition

To summarize, our metrics are based whether a set of preferences can rationalize the data under equality in menus, they are *robust* to the false positives that is problematic for the heuristic approaches to assessment. Furthermore, in some situations, such as those in Figures 3b and 4b, our metrics can detect discrimination when heuristic approaches fail to do so. Nevertheless, a drawback of our approach is that it stills leave some possibility for false negatives, because the mere existence of a set of preferences that explains the data under equality in menus does not mean that it is the true set of preferences. This is a weakness compared to experimental that may allow the researcher to directly observe menus, but the advantage of our approach is that it requires only data on outcomes and few assumptions on the data generating process. We will discuss the power of our metrics in more detail in Figure 3.4.

3 Robust metrics for assessing discrimination in menus

In this section we define our robust metrics for assessing discrimination in menus, building on the intuition from Section 2 that equality in menus should imply the existence of a one-to-one "match" between minority and white consumers who could have faced the same menus in the data. We keep our model fairly simple, where menus are simply treated as a collection of items, which we define more formally in Section 3.1. We then define a direct test metric for equality in menus Section 3.2, a more welfare relevant differences in menus metric for whether one group of consumers would like to switch to another distribution of menus in Section 3.3. We leave inference on these metrics to Section 4.

3.1 Model

A menu item has values over k dimensions of attributes,[8] which we encode by $x \in \mathbb{X} \subset \mathbb{R}^k$. A menu $\mathbf{m} \subseteq \mathbb{X}$ is a collection of such menu items which are presented to the borrowers. When presented with a menu $\mathbf{m}$, we observe borrower i making a choice which maximizes their utility over menu items $u_i(x)$. That is, we observe choices x_i where:

$$x_i \in \arg\max_x \{u_i(x) : x \in \mathbf{m}\} \tag{1}$$

To keep the distribution of menus Lebesgue measurable and to implement the inference procedure of Section 4, we make the simplifying assumption that the set of items available to choose from is finite:

Assumption 1. *(Finiteness) The set of possible menu items, $\mathbb{X}$, is finite.*

[8] In our context, the two dimensions of a menu item in the mortgage context are rates and discount points.

Under Assumption 1, we can consider a probability distribution over possible menus $\mathbf{m} \sim \mathbf{M}$. This setup is fairly general and follows from consumers having standard (ie. complete and transitive) preferences over menu items.[9]

3.2 A robust test for inequality in menus

Suppose borrowers who are observationally similar in groups 1 and 2 faces distributions of menus, $\mathbf{M}_1, \mathbf{M}_2$, respectively, such that $\mathbf{m}_1 \sim \mathbf{M}_1$ for borrowers in group 1 and $\mathbf{m}_2 \sim \mathbf{M}_2$ for borrowers in group 2. The researcher wishes to compare the distribution menus across two groups of borrowers. More specifically, testing for equality of menus can be written as testing the null hypothesis that the distribution of menus being offered to both groups are equal. That is:

$$\mathcal{H}_0 : \mathbf{M}_1 = \mathbf{M}_2 \tag{2}$$

$$\mathcal{H}_1 : \neg\mathcal{H}_0 \tag{3}$$

To go from data on choices to statements about menus, we place restrictions on what choices could have been plausibly made from the same menus in terms of borrower preferences. For mortgages, it is plausible to assume that paying more in both interest rates and discount points is a dominated choice (and indeed would not be offered as a choice by the loan originator), which is the intuition we used in Figure 5 to reject equality in menus in that situation. We formalize this as Assumption 2:

Assumption 2. *(Dominance Restriction on Preferences) Paying more in rates and points is dominated. More formally, let* $x_1 = [r_1, y_1], x_2 = [r_2, y_2]$ *where* r_1, r_2 *represents rates and* y_1, y_2 *represents points. Then, if* $r_1 > r_2, y_1 \geq y_2$ *or* $r_1 \geq r_2, y_1 > y_2$ *then* $u_i(x_1) < u_i(x_2), \forall i$.

We illustrate in Figure 6 the restrictions on what observed choices may come from on the same menu under Assumption 2. The observed choice of the borrower, shown as the black dot, implies that they did not have the lower-left dashed quadrant available on their menu since otherwise they would have chosen it. Similarly, any choice on the upper-right quadrant could not have been from the same menu as the choice indicated, as that agent would have an incentive to switch to that choice. Note that while Assumption 2 is defined in the form of preferences, it could have also been defined in terms of menus which would have led to

[9]Our assumption that consumers maximize utility over menu items does rule out more behavioral representations of preferences over menus such as Gul and Pesendorfer (2001), in which the existence of some menu items may "tempt" consumers to change their rankings of other menu items. In that case, our statistical test of inequality in menus would still be valid, but the interpretation of our more welfare relevant differences in menus metric would be nuanced.

the same restrictions. Defining it in terms of preferences, though, is because it will allow us to define a welfare-relevant metric for assessing differences in menus later on.

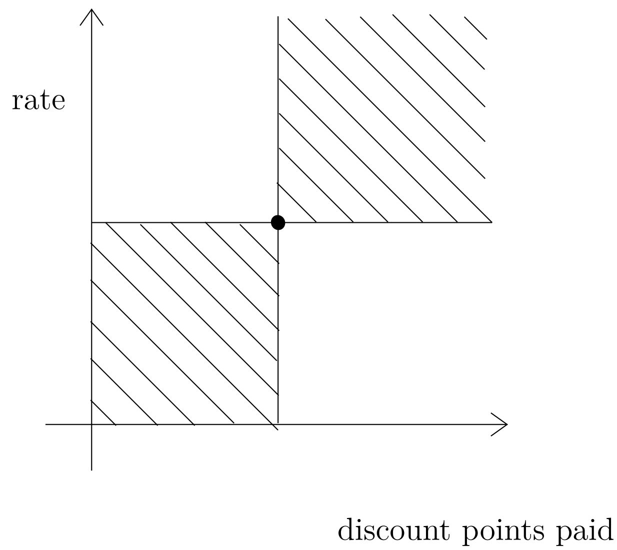


Figure 6: Restriction on what cannot lie on the same menus from dominance.

While Assumption 2 is sufficient for rejecting equality in menus in the example situation of Figure 5, in our empirical application it is too weak of a restriction to be informative. For our empirical analyses, we further adopt the industry rule of thumb of Bartlett et al. (2019) that each point paid reduces the interest rate on a mortgage by between 1/8 to 1/4 for conforming mortgages, with an expanded range for FHA mortgages. This is an assumption about menus rather than about preferences, which we formalize as Assumption 3:

Assumption 3. *(Restriction on Menus) In menus, each point paid reduces the rate by between* $[a, b]$*. More formally,* $x_1 = [r_1, y_1]$ *and* $x_2 = [r_2, y_2]$ *can only lie on the same menu* $\{x_1, x_2\} \subseteq \mathbf{m}$ *only if:*

$$a \leq \frac{r_2 - r_1}{y_2 - y_1} \leq b \tag{4}$$

with $x_1 = [r_1, y_1], x_2 = [r_2, y_2]$*,* r_1, r_2 *representing rates and* y_1, y_2 *representing points.*

We illustrate in Figure 7 the effect of defining a menu set based on Assumption 3. As Figure 7 indicates, range of possible choices that could have come from the same menu as the consumer with the choice illustrated by the black dot is more restricted under this assumption, compared with using only dominance relationships in terms of preferences as in Assumption 2. Thus, this improves our ability to detect discrimination in menus. Nevertheless, as we mentioned earlier, Assumption 2 is still needed to make welfare comparisons of menu distributions in Section 3.3.

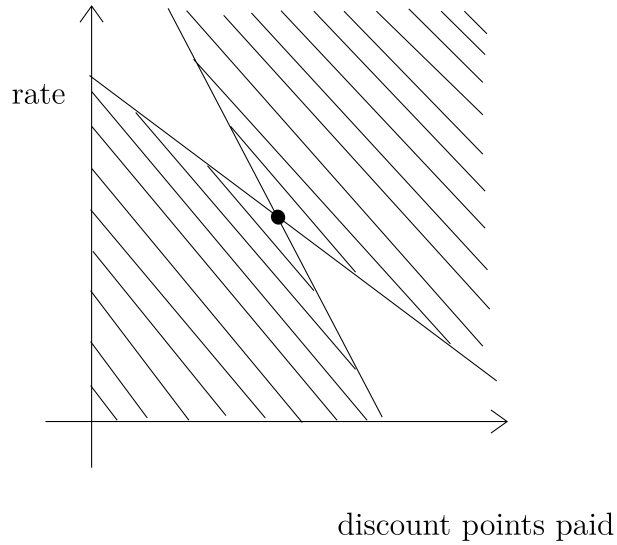
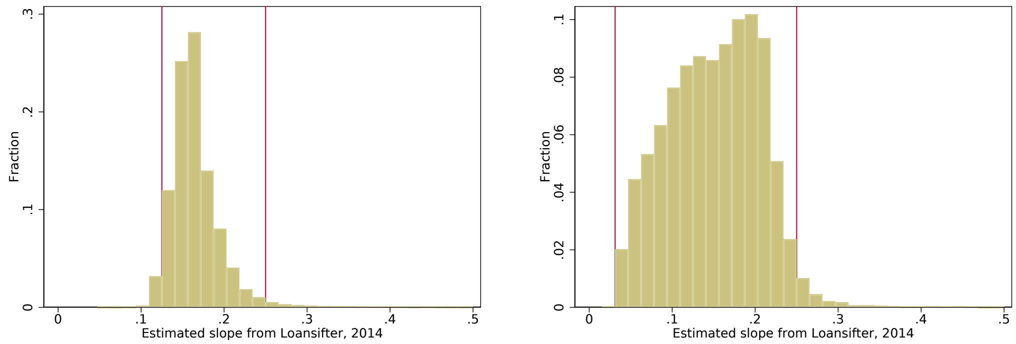


Figure 7: Restriction on what cannot lie on the same menus from Bartlett et al. (2019)'s "rule of thumb".

Empirically, we find that Bartlett et al. (2019)'s industry rule of thumb which motivated Assumption 3 covers a vast majority of menus based on data from a sample of lender ratesheets which enumerates the rate and discount point menus. Using the LoanSifter data from Fuster, Lo, and Willen (2019) in 2014, we estimate slopes of menus within their sample of 30 year conforming, purchase mortgages across seven different MSAs (Chicago, Houston, Los Angeles, Miami, New York City, Seattle, and San Francisco) and a range of loan amounts, FICO scores, and LTVs. The sample construction is discussed in more detail in Fuster, Lo, and Willen (2019). We estimate the slopes of the rate-point trade-off by taking the difference in the interpolated rate from 0 points to 2 points and dividing by 2. In this sample, the Bartlett et al. (2019)'s rule that each point paid is worth between 1/8 to 1/4 of a point covers 94.4% of all ratesheet observations for conforming mortgages. For FHA mortgages, we use an expanded rule that each point is worth between 1/32 to 1/4 in rate, which covers 97.6% of all observations. We illustrate this in Figure 8.



(a) Conforming mortgages, Bartlett et al. (2019) restriction in red

(b) FHA mortgages, our expanded restriction in red

Figure 8: Ratesheet evidence for our menu slopes Assumption 3.

Note: these figures are constructed using 2014 LoanSifter rate sheet data from Fuster, Lo, and Willen (2019) for conforming and FHA mortgages. The slope of the rate-point tradeoff is estimated by taking the interpolated rate from 0 to 2 points and dividing by 2. The red lines in Figure 8a represent 1/8 and 1/4, and the red lines in Figure 8b represent 1/32 and 1/4.

There is substantial heterogeneity between the menu slopes across lender-weeks which we see in Figure 8, perhaps reflecting market power or lender and time specific costs. As we explained earlier, the existence of this heterogeneity interacted with possible differences in preferences between the two groups makes the "comparing means" heuristic and its variations prone to false positives and negatives.

Under Assumption 3, we can define an indicator function for whether choices x_1, x_2 could have come from the same menu:

$$\phi(x_1 = [r_1, y_1], x_2 = [r_2, y_2]) = \begin{cases} 1 \text{ , if } a \leq \frac{r_2 - r_1}{y_2 - y_1} \leq b \\ 0 \text{ , otherwise.} \end{cases} \tag{5}$$

After defining this function, we have the following identification result for when vectors of choice probabilities $\mathbf{p}_1 = [p_1(x_1), p_1(x_2), \ldots], \mathbf{p}_2 = [p_2(x_1), p_2(x_2), \ldots]$ from observationally similar groups of borrowers can be rationalized under the null hypothesis of equality in the distribution of menus $\mathcal{H}_0 : M_1 = M_2$:

Theorem 1. *Under Assumptions 1 and 3, choice probabilities* $\mathbf{p}_1, \mathbf{p}_2$ *can be generated from the same underlying distribution of menus* $\mathbf{M}_1 = \mathbf{M}_2$ *if and only if there exists a coupling with probability mass function* $\boldsymbol{\pi}(x_1, x_2) : \mathbb{X} \times \mathbb{X} \to [0, 1]$ *with implied marginal densities* $\sum_{x_2} \boldsymbol{\pi} = \mathbf{p}_1, \sum_{x_1} \boldsymbol{\pi} = \mathbf{p}_2$ *such that:*

$$T \equiv 1 - E_{\boldsymbol{\pi}} \phi(x_1, x_2) = 0 \tag{6}$$

Proof. See Appendix A.1.1. □

Theorem 1 formalizes the intuition from Section 2 that equality in menus should imply a way to "match" observations to one another such that each pair can come from the same set of menus. In particular, $\boldsymbol{\pi}$ serves as a coupling or "matching function" where each of its entries $\pi(x_1, x_2)$ represents the extent to which choice probabilities $p_1(x_1)$ and $p_2(x_2)$ came from the same menu, and the requirement of Equation (6) is that this coupling should lie entirely in the area where $\phi(x_1, x_2) = 1$. Based on Theorem 1, we are ready to define a test statistic that looks at by how much is this matching deficient:

Definition 1. *Our test statistic for equality in menus, $\hat{T}$, given sample choice probabilities $\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2$, is:*

$$\hat{T} = \min_{\boldsymbol{\pi}(x_1,x_2)} 1 - E_{\boldsymbol{\pi}}\phi(x_1, x_2), s.t. \sum_{x_2} \boldsymbol{\pi} = \hat{\mathbf{p}}_1, \sum_{x_1} \boldsymbol{\pi} = \hat{\mathbf{p}}_2, \boldsymbol{\pi} \geq 0 \tag{7}$$

Where the extent to which $\hat{T} > 0$ indicates a failure of the perfect matching condition in Theorem 1 in sample, which is evidence against equality in menus. The statistic $\hat{T}$ in Equation (7) is a finite dimensional optimal transport objective that can be efficiently computed using a linear program.[10] To deal with sampling error inherent in $\hat{T}$, we discuss in Section 4 on how critical values for $\hat{T}$ from Equation (7), as an optimal transport objective, can be consistently simulated.

3.3 Metric for assessing differences in menus

The direct test of inequality in the distribution of menus in Section 3.2, while indicative of discrimination in menus, has the drawback that it may not be the object of interest for researchers. A statistical rejection of equality in menus may by itself carry little information about welfare, since the fact that the distribution of menus presented to one group is different in some aspect than the distribution of menus presented to another group may not be of welfare consequence, a problem which Abadie (Forthcoming) discusses in more detail. Rather, for the purposes of comparing menus across two distributions we want a metric for assessing whether menus from one group is meaningfully "better" than that of another group. For this purpose, we ask the question: *if black consumers were instead assigned white menus, how much better off would they be?*

Conceptually, we consider the object of interest to be the change in welfare when black consumers were instead assigned white menus, under an assignment rule $\pi(i, j)$ that map

[10] Galichon (2016) reviews optimal transport methods and its application to economics.

each consumer $i \in \mathcal{I}_1$ from group 1 to the menu of consumer $j \in \mathcal{I}_2$ from group 2. Giving all consumers the same welfare weight, this objective can be represented by Equation (8):

$$\Delta W_{\mathcal{I}_1 \to \mathcal{I}_2, \boldsymbol{\pi}} = \sum_{i \in \mathcal{I}_1, j \in \mathcal{I}_2} \pi(i,j)(u_i(\mathbf{m}_j) - u_i(\mathbf{m}_i)) \tag{8}$$

To get at $\Delta W_{\mathcal{I}_1 \to \mathcal{I}_2, \boldsymbol{\pi}}$, we proxy for the utility difference $u_i(\mathbf{m}_j) - u_i(\mathbf{m}_i)$ through a metric $d_{i \to j}(\mathbf{m}_i, \mathbf{m}_j)$ which measures by how much would consumer i be willing to increase the interest rates on their loan in order to switch from menu $\mathbf{m}_i$ to menu $\mathbf{m}_j$:

$$u_i(\mathbf{m}_j) - u_i(\mathbf{m}_i) = d_{i \to j}(\mathbf{m}_i, \mathbf{m}_j) \equiv \sup\{\delta \in \mathbb{R} : u_i(\{x + \delta e^r, x \in \mathbf{m}_j\}) \geq u_i(\mathbf{m}_i)\} \tag{9}$$

Where e^r represents a basis vector that is equal to 1 at the location indexing interest rates. If consumers have constant marginal utility over interest rates such that utility can be represented as $u_i(x = [r, y]) = r + f(y)$, then $d_{i \to j}(\mathbf{m}_i, \mathbf{m}_j)$ is directly proportional to the utility change for consumer i after switching to menu j. Even if consumers do not have constant marginal utility over interest rates it is still meaningful as a "willingness to pay" metric since $d_{i \to j}(\mathbf{m}_i, \mathbf{m}_j)$ measures by how much would consumer i be willing to increase the interest rates to switch from $\mathbf{m}_i$ to $\mathbf{m}_j$. In the rest of this section we will show how we can compute an informative lower bound for this metric given the data, $\underline{d}_{i \to j}(x_i, x_j) \leq d_{i \to j}(\mathbf{m}_i, \mathbf{m}_j), x_i \in \mathbf{m}_i, x_j \in \mathbf{m}_j$, which then leads our differences in menus measure.[11]

To define our lower bound, we make an additional assumption that menus are complete in points, such that all choices of points are available to borrowers, which we formalize as Assumption 4. This is an approximation since lenders may limit the choices of points to certain decimals (e.g. 0.134, 0.266, ...) rather than literally the full range, but the implications of such small gaps in menus is likely small. Another complication is that there may be information constraints on the part of borrowers such that they do not "see" their full choice set (ie. some borrowers may not know that they can pay/receive points), but as long as these information constraints are held constant in the counterfactual where they switch to another group's menus, our lower bound metric would remain valid.

Assumption 4. *(Completeness) The menus are complete in discount points. More specifically,* $\forall \mathbf{m}, \forall y', \exists x = [r, y'] \in \mathbf{m}$.

The effect of this Assumption 4 is illustrated in Figure 9. Under the assumption that the mortgage menus are complete in discount points, we can meaningfully say that the minority

[11] We note that while we define our $\underline{DIM}_{1 \to 2}$ over "willingness to pay" in terms of interest rates, we could have also defined it using points, but due to the possible existence of cash on hand constraints which are likely binding for many consumers constant marginal utility is very unlikely to hold for points which makes for a weaker welfare interpretation.

borrower whose choice is represented by the black dot would have preferred the menu of the white borrower whose choice is represented by the white dot, because there exists a level of discount points such that all possible choices that are in white borrower' menu dominates the minority borrower's choice. Otherwise, the minority borrower might not have preferred the white borrower's menu because the white borrower's menu could have been a singleton that the minority borrower dislikes. Therefore, adding in the assumption of menu completeness in points sharpens the comparison of menus.

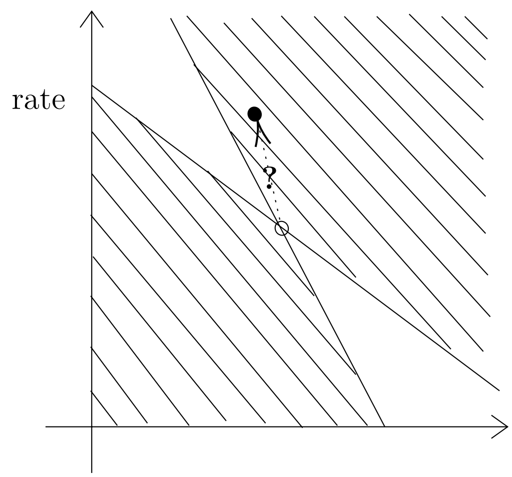


Figure 9: Impact of assuming that menus are complete in either rates or points

We illustrate in Figure 10 how we can construct a lower bound for the willingness to pay in terms of interest rates $\underline{d}_{i\to j}(x_i, x_j)$ under Assumption 4. There, a borrower 1 who made a choice x_1 from an unobserved menu $\mathbf{m}_1$ has made a choice that is dominated (in terms of paying a higher rate at the same level of points) by any possible the menu of borrower 2 who made a choice x_2 from a menu $\mathbf{m}_2$. This implies that, by revealed preference for borrower 1, the menu that borrower 2 faced is better than borrower 1's menu, or $\mathbf{m}_2 \succ_1 \mathbf{m}_1$. For borrower 1 to possibly become indifferent between $\mathbf{m}_1$ and $\mathbf{m}_2$, $\mathbf{m}_2$ needs to be shifted up by at least the amount indicated in the figure in the dimension of interest rates. In other words, a lower (sharp) lower bound for $d_{1\to 2}(\mathbf{m}_1, \mathbf{m}_2)$ is $\underline{d}_{1\to 2}(x_1, x_2)$ in the sense that borrower 1 is willing to pay at least $\underline{d}_{1\to 2}(x_1, x_2)$ more in interest rate in order to get borrower 2's menu. Similarly, the menu faced by borrower 2 would need to be shifted downwards by at most the negative $\underline{d}_{3\to 2}$ before it dominates x_3's choice. Therefore, borrower 3 with choice x_3 would need to receive at most $-\underline{d}_{3\to 2}$, or pay at least $\underline{d}_{3\to 2}$, before being willing to switch to borrower 2's menu.

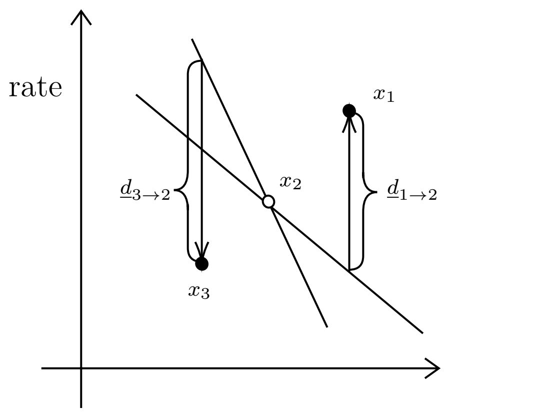


Figure 10: Lower bound for borrower 1's willingness to pay a higher interest rate in order to get borrower 2's menu.

Formalizing the intuition from Figure 10, we define our lower bound for the willingness of borrower 1 to switch to borrower 2's menu under Assumptions 1 to 4 as follows:

$$\underline{d}_{1\to 2}(x_1, x_2) \equiv r_1 - r_2 + a\max(y_1 - y_2, 0) + b\min(y_1 - y_2, 0) \leq d_{1\to 2}(x_1, x_2) \qquad (10)$$

Where in the first line j indexes points, and in the second line we take take it to the mortgage setting and let $x_1 = [r_1, y_1], x_2 = [r_2, y_2]$ where r_1, r_2 are rates and y_1, y_2 are points. The lower bound for how much borrower 1 would be willing to pay to switch to the menus of borrower 2, $\underline{d}_{1\to 2}(x_1, x_2)$, then allows us to define our differences in menus (DIM) metric:

Theorem 2. *Under Assumptions 1 to 4, choice probabilities* $\mathbf{p}_1, \mathbf{p}_2$ *implies a DIM metric:*

$$\underline{DIM}_{1\to 2} = \min_{\boldsymbol{\pi}(x_1,x_2)} E_{\boldsymbol{\pi}}\underline{d}_{1\to 2}(x_1, x_2), s.t. \sum_{x_2} \boldsymbol{\pi} = \mathbf{p}_1, \sum_{x_1} \boldsymbol{\pi} = \mathbf{p}_2, \boldsymbol{\pi} \geq 0 \qquad (11)$$

where $\underline{DIM}_{1\to 2}$ *serves as a lower bound for the average willingness to pay in terms of interest rates for borrowers in group 1 to switch menus with borrowers in group 2. Furthermore, if borrowers have constant marginal utility in interest rate, then:*

$$\underline{DIM}_{1\to 2} \leq \Delta W_{\mathcal{I}_1\to\mathcal{I}_2,\boldsymbol{\pi}} \qquad (12)$$

Proof. Follows from the definitions in Equation (9) and Equation (10). □

Theorem 2 shows that, when all consumers have the same constant marginal utility over interest rates (normalized to 1), our DIM metric is as a lower bound for the change in welfare for when consumers in group 1 are instead assigned menus from group 2 in an arbitrary way

$\Delta W_{\mathcal{I}_1,1\to 2,\pi_{1\to 2}}$. If instead consumers do not have constant marginal utility over interest rates, then the $\underline{DIM}_{1\to 2}$ metric could still be interpreted as the average increase in interest rates consumers in group 1 would be willing to pay in order to switch to menus from group 2. Furthermore, by Theorem 1, equality in menus would imply that $\underline{DIM}_{1\to 2} \leq 0$, so a finding that $\underline{DIM}_{1\to 2} > 0$ is also rejection of equality in menus in a welfare relevant way.

The sample analogue of the DIM metric follows immediately from Definition 2.

Definition 2. *Our empirical differences in menus metric,* $\underline{D\hat{I}M}_{1\to 2}$*, given choice probabilities* $\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2$*, is:*

$$\underline{D\hat{I}M}_{1\to 2} = \min_{\boldsymbol{\pi}(x_1,x_2)} E_{\boldsymbol{\pi}} \underline{d}_{1\to 2}(x_1,x_2), s.t. \sum_{x_2} \boldsymbol{\pi} = \hat{\mathbf{p}}_1, \sum_{x_1} \boldsymbol{\pi} = \hat{\mathbf{p}}_2, \boldsymbol{\pi} \geq 0 \tag{13}$$

In terms of inference, the sample DIM metric in Definition 2 is also the value of an finite dimensional optimal optimal transport problem. We discuss inference in the following Section 4.

3.4 When does our method have power?

Our metrics for differences in menus are robust in the sense that they are immune to the false positives problem which the existing methods suffer from, but may still generate false negatives in that there exists scenarios in which there is the data is rationalizable under equality in the distribution of menus but in fact the distribution is different. Generally speaking, our method only has power to the extent the borrowers' choices cannot be rationalized by an equal distribution of menus under the restrictions on preferences and menus that we have made. In this subsection we give some examples of scenarios of when this would and would not occur. In our later empirical analysis, show that our methodology does have enough power to be useful in detecting discrimination in mortgage markets.

Figure 11 illustrates a scenario in which white and minority borrowers both face the same default menu but some white borrowers are sometimes offered a discretionary discount in terms of points. In the figure, the menu represented by the dashed line is shifted leftward for a white borrower, but all minority borrowers face the original menu. This shift makes the bottom-left choice by the white borrower not matchable to any of the minority borrowers', so our one-to-one matching condition for equality in menus in Theorem 1 is broken. Our Theorem 1 would therefore have power to detect a difference in the distribution of menus offered to white and minority borrowers in this case. Futhermore, when the discretionary discount being given to white borrowers are large enough relative to the range of permissible menus, our DIM metric in defined in Theorem 2 would also show that minority borrowers would be willing to pay to switch to the white borrowers' menus.

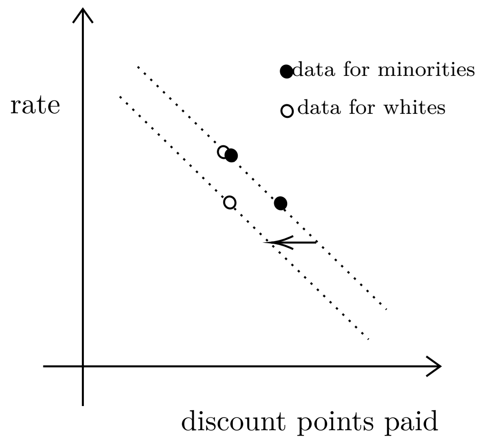


Figure 11: Power when discretionary discounts offered to only some white borrowers makes the data not rationalizable under equality in menus

Figure 12 gives an example situation in which minority and white borrowers were presented with different menus, with white borrowers on average paying less in both rates and points, but the data is rationalizable under equality in menus. More specifically, while the data was generated by menus represented by the dotted lines which differ in distribution between white and minority borrowers, the data can be rationalized with one minority and one white borrower both choosing from the hypothetical menu represented by the dashed line, and the other borrowers choosing from the correct dotted line menu. In this scenario, we cannot rule out that the true distribution of menus are represented by the dashed line plus a dotted line rather than the two dotted lines, and therefore our metrics from both Theorem 1 and Theorem 2 would fail to reject equality in menus.

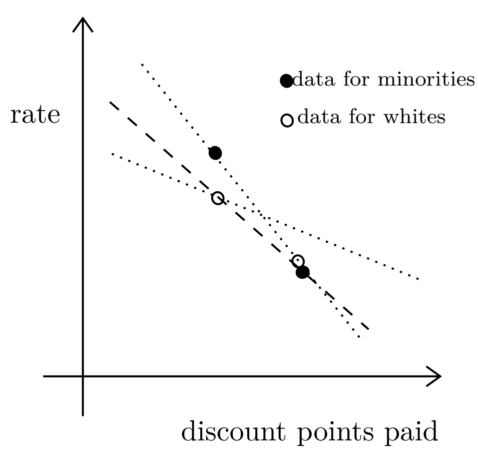


Figure 12: No power when the data can be rationalized by (incorrect but plausible) distribution of menus that are equal across the racial groups

While our approach cannot detect differences in menus in the situation of Figure 12, it does so with good reason: it is possible that the borrowers' choices were truly generated by the dashed line and the dotted line such that the racial groups did in fact face the same distribution of menus. It would be difficult to rule out that possibility without more stringent assumptions. Therefore, in our quest to be robust to the false positives problem that is suffered by existing methods, we allow for the possibility of false negatives when the data is rationalizable by an equal distribution of menus across the racial groups. Empirically in Section 5, we show that we are able to reject equality in menus for conforming mortgages, which is evidence that our method does have enough power to be useful in our mortgage setting.

4 Inference for Optimal Transport

In this section we devise a new procedure for conducting hypothesis testing on the values of optimal transport problems, which includes our metrics derived in Section 3. This is needed because the objective values of our model can be non-differentiable and therefore simple bootstrap methods are not consistent (Fang and Santos, 2018). While there are some existing methods in the literature which can be applied in more general contexts compared to optimal transport, they are either too conservative or converges too slowly based on our simulations, which we discuss at the end of this section. Our procedure may be useful for other researchers who wish to apply optimal transport methods to economics, some of which is listed in Galichon (2016).

We consider hypothesis testing on the value φ of a finite dimensional optimal transport problem with cost function $\phi(x_1, x_2), x_1, x_2 \in \mathbb{X}$ and marginal distributions $\mathbf{p}_1, \mathbf{p}_2$:

$$\hat{\varphi}(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2) = \min_{\boldsymbol{\pi}(x_1,x_2)} E_{\boldsymbol{\pi}}\phi \text{ s.t. } \sum_{x_2} \boldsymbol{\pi} = \hat{\mathbf{p}}_1, \sum_{x_1} \boldsymbol{\pi} = \hat{\mathbf{p}}_2, \boldsymbol{\pi} \geq 0 \tag{14}$$

Where the hypothesis is in the form of the value of the optimal transport $\varphi(\mathbf{p}_1, \mathbf{p}_2)$ as a function of the true marginal distributions $\mathbf{p}_1, \mathbf{p}_2$ are less than or equal to some value φ_0:

$$\mathcal{H}_0 : \varphi(\mathbf{p}_1, \mathbf{p}_2) \leq \varphi_0 \tag{15}$$

$$\mathcal{H}_a : \varphi(\mathbf{p}_1, \mathbf{p}_2) > \varphi_0 \tag{16}$$

The form of the null hypothesis in Equation (15) is especially relevant to us because both our test of equality in menus (that is, whether $T \leq 0$) and our lower bound DIM metric (that is, whether $DIM \geq DIM_0$) can be expressed in terms of it. We provide a methodology to conduct this hypothesis test by looking at the asymptotic distribution of $\hat{\varphi}$ and finding a critical value to compare the observed $\hat{\varphi}$ to under $\mathcal{H}_0$. This can then be inverted into a confidence interval for the true value of φ.

As an overview, we combine the directional derivatives approach of Fang and Santos (2018) and Shapiro (1990) with a size correction of Romano, Shaikh, and Wolf (2014) and McCloskey (2017) which allows us to conduct hypothesis testing for optimal transport with uniform size control. In doing so, we prove the directional differentiability for optimal transport problems on finite domains, and show how the general approach can be implemented as a Linear Program with Complementarity Constraints (LPCC).

We make use the definition of Hadamard directional differentiation from Fang and Santos (2018), with some notational differences tailored to the optimal transport setting. Here, the value of an optimal transport represents a map $\varphi : \mathbb{D}_\varphi \to \mathbb{R}$, where $\mathbb{D}_\varphi = \mathcal{P}_\mathbb{X} \times \mathcal{P}_\mathbb{X}$, $\mathcal{P}_\mathbb{X}$ is the set of probability measures on $\mathbb{X}$. Let $\mathbb{D}_0 = \{P_1 - P_2 : P_1 \in \mathcal{P}_\mathbb{X}, P_2 \in \mathcal{P}_\mathbb{X}\}$ be the set of possible differences in probability measures, and $\theta = \{p_1, p_2\}$ be the marginal distributions, then:

Definition 3. *(Fang and Santos, 2018) A map $\varphi : \mathbb{D}_\varphi \to \mathbb{R}$ is said to to be Hadamard directionally differentiable at $\theta \in \mathbb{D}_\varphi$ tangentially to the set $\mathbb{D}_0$, if there is a continuous linear map $\varphi'_\theta : \mathbb{D}_0 \to \mathbb{R}$ such that:*

$$\lim_{n\to\infty} \left\| \frac{\varphi(\theta + t_n \mathbf{h}_n) - \varphi(\theta)}{t_n} - \varphi'_\theta(\mathbf{h}) \right\| = 0 \tag{17}$$

for all sequences $\{\mathbf{h}_n\} \subset \mathbb{D}_0$ and $\{t_n\} \subset \mathbb{R}^+$ such that $t_n \to^+ 0, \mathbf{h}_n \to \mathbf{h} \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n \mathbf{h}_n \in \mathbb{D}_\varphi$ for all n.

The main difference between the Hadamard directional differentiability and the typical notion of differentiability is that t_n is restricted to be positive in Definition 3 but not in the standard definition of differentiability. That is, loosely speaking the directional derivatives represent the change in the value of the function for a small change in its inputs "in the direction $\mathbf{h}$" for each $\mathbf{h}$.

We show in Theorem 3 that the value of all Monge-Kantorovich optimal transport problems with bounded cost functions on finite spaces is Hadamard directionally differentiable in the sense of Definition 3. In particular, Theorem 3 is a generalization of Sommerfeld and Munk (2018) which shows that the Wasserstein metric on finite spaces, which is the value of an optimal transport problem with the cost function restricted to distance metrics, is directionally differentiable.[12]

[12] It is also related to Tameling, Sommerfeld, and Munk (2019) who prove that the Wasserstein distance on countable metric spaces is directionally differentiable. Our Theorem 3 can be similarly extended to countable metric spaces under the assumption that the cost function ϕ is continuous.

Theorem 3. *The value φ of a optimal transport problem with cost function $\phi(x_1, x_2), x_1, x_2 \in \mathbb{X}$, where $M = \sup |\phi| < \infty$ and $\dim(\mathbb{X}) < \infty$, is Hadamard directionally differentiable, with derivative equal to:*

$$\varphi'_{\mathbf{p}_1,\mathbf{p}_2}(\mathbf{h}_1, \mathbf{h}_2) = \max_{\mathbf{u},\mathbf{v} \in \Psi^*(\mathbf{p}_1,\mathbf{p}_2)} \mathbf{h}_1^T \mathbf{u} + \mathbf{h}_2^T \mathbf{v} \tag{18}$$

where $\Psi^*(\mathbf{p}_1, \mathbf{p}_2) = \{\mathbf{u}, \mathbf{v} : \mathbf{p}_1^T \mathbf{u} + \mathbf{p}_2^T \mathbf{v} = \varphi(\mathbf{p}_1, \mathbf{p}_2), u(x_1) + v(x_2) \leq \phi(x_1, x_2) \forall x_1, x_2\}$ is the set of dual solutions to the linear programming problem, for all $\{\mathbf{p}_1, \mathbf{p}_2\} \in \mathbb{D}_\varphi$, tangentially to the set $\mathbb{D}_0$.

Proof. See Appendix A.1.2. □

Under i.i.d. sampling, we know that $\hat{\mathbf{p}}_1 - \mathbf{p}_1$ and $\hat{\mathbf{p}}_2 - \mathbf{p}_2$ approaches a multivariate Normal distribution:

$$\hat{\mathbf{p}}_1 - \mathbf{p}_1 \to \mathrm{N}\left(0, \begin{bmatrix} p_{1,1}(1 - p_{1,1}) & -p_{1,1}p_{1,2} & \dots \\ -p_{1,2}p_{1,1} & p_{1,2}(1 - p_{1,2}) & \dots \\ \dots & \dots & \dots \end{bmatrix}\right) \tag{19}$$

And likewise for $\hat{\mathbf{p}}_2 - \mathbf{p}_2$, such that by construction, Assumptions 2.1 and 2.2 of Fang and Santos (2018) is satisfied. Then, Theorem 2.1 of Fang and Santos (2018) immediately implies that:

$$r_n[\varphi(\{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2\}) - \varphi(\{\mathbf{p}_1, \mathbf{p}_2\})] = \varphi'_{\{\mathbf{p}_1,\mathbf{p}_2\}}(r_n[\{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2\} - \{\mathbf{p}_1, \mathbf{p}_2\}]) + o_p(1) \tag{20}$$

Such that the asymptotic distribution of $\varphi(\{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2\})$ can be obtained via the directional Delta method. The remaining challenge for inference is that the true $\mathbf{p}_1, \mathbf{p}_2$ used in $\varphi'_{\{\mathbf{p}_1,\mathbf{p}_2\}}$ in Equation 20 is not known, and thus must be estimated. While we could have used a plug-in version $\varphi'_{\{\hat{\mathbf{p}}_1,\hat{\mathbf{p}}_2\}}$, that would only converge pointwise and not uniformly, which is known in the moment inequalities literature to be a poor approximation to the finite sample properties of estimators for which there are discontinuities in the pointwise asymptotic distribution (Andrews and Soares, 2010). We deal with this problem using the logic of Romano, Shaikh, and Wolf (2014) and McCloskey (2017). More specifically, we construct a confidence band for $[\mathbf{p}_1, \mathbf{p}_2]$ at level β such that:

$$\limsup_{n\to\infty} \Pr([\mathbf{p}_1, \mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}) \geq 1 - \beta \tag{21}$$

Many uniform confidence bands satisfying Equation (21) are available, for example from Montiel Olea and Plagborg-Møller (2019). Then, we take our estimate of the directional derivative as the maximum directional derivative within this confidence band:

$$\hat{\varphi}'_\beta(\mathbf{h}_1, \mathbf{h}_2) = \max_{\mathbf{u},\mathbf{v} \in \Psi(\mathbf{p}_1,\mathbf{p}_2):[\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_\beta} \mathbf{h}_1^T \mathbf{u} + \mathbf{h}_2^T \mathbf{v} \tag{22}$$

Where $\Psi = \{\mathbf{u}, \mathbf{v} : \mathbf{p}_1^T \mathbf{u} + \mathbf{p}_2^T \mathbf{v} \leq \varphi_0, u(x_1) + v(x_2) \leq \phi(x_1, x_2) \forall x_1, x_2\}$ are the set of dual solutions under the null hypothesis $\mathcal{H}_0 : \varphi \leq \varphi_0$.

Next, we define the critical value for φ by using our estimated maximum directional derivative from Equation (22) at level $1 - \alpha + \beta$:

$$\hat{c}_{1-\alpha+\beta} = \inf\{c \in \mathbb{R} : \Pr(\hat{\varphi}'_\beta(\mathbf{h}_1, \mathbf{h}_2) \leq c) \geq 1 - \alpha + \beta\} \tag{23}$$

Where the distribution of $\mathbf{h}_1, \mathbf{h}_2$ is the asymptotic distribution of $\hat{\mathbf{p}}_1 - \mathbf{p}_1, \hat{\mathbf{p}}_2 - \mathbf{p}_2$. In the following Corollary 1, we will prove uniform coverage for when the observed value $\hat{\varphi}$ is under than the critical value $\hat{c}_{1-\alpha+\beta}$, and suggest a computationally tractable version of it as a linear program with complementarity constraints (LPCC).[13]

Corollary 1. *Suppose we have uniform confidence bands for* $[\mathbf{p}_1, \mathbf{p}_2] \in \hat{\mathcal{P}}_\beta$ *which provide uniform coverage as in Equation (21), then under* $\mathcal{H}_0 : \varphi \leq \varphi_0$*:*

$$\limsup_{n\to\infty} \Pr(\hat{\varphi} \geq \hat{c}_{n,1-\alpha+\beta}) \leq \alpha \tag{24}$$

where $\hat{c}_{n,1-\alpha+\beta}$ *is computed as in Equation (23).*

Proof. See Appendix A.1.3. □

Corollary 1 implies that uniformly valid hypothesis testing for the value of φ can be conducted by first computing a set of uniform confidence bands $[\mathbf{p}_1, \mathbf{p}_2] \in \hat{\mathcal{P}}_\beta$, and then maximizing over all directional derivatives within these bands as in Equation (22). Computationally, directly maximizing over the directional derivative defined in Equation (22) is difficult because it involves optimizing over a non-linear dual value constraint $\mathbf{p}_1^T \mathbf{u} + \mathbf{p}_2^T \mathbf{v} \leq \varphi_0$. To deal with this, we replace it with a complementary slackness condition $\boldsymbol{\pi}^T \mathbf{s} = 0, \boldsymbol{\pi} \geq 0, \mathbf{s} \geq 0$ where $u(x_1) + v(x_2) + s(x_1, x_2) = \phi(\mathbf{p}_1, \mathbf{p}_2)$ which implies that the elements of $\boldsymbol{\pi}$ and $\mathbf{s}$ cannot be positive simultaneously. Following the operations research shorthand, we represent this constraint by $\boldsymbol{\pi} \leq 0 \perp \boldsymbol{s} \geq 0$. The derivation of complementary slackness conditions like this can be found in standard texts on optimal transport/linear programming, and in particular Hsieh, Shi, and Shum (2017) uses a similar set of conditions for their projection method.

[13] As explained in Hsieh, Shi, and Shum (2017), LPCCs are well-understood computationally and are implemented in software such as Knitro: https://www.artelys.com/docs/knitro/2_userGuide/complementarity.html.

Based on this equivalency, the problem of finding critical values for the null hypothesis $\mathcal{H}_0 : \varphi \leq \varphi_0$ versus the alternative $\mathcal{H}_a : \varphi > \varphi_0$ be written as the following LPCC:

$$\hat{\varphi}'_\beta(\mathbf{h}_1, \mathbf{h}_2) = \max_{\mathbf{u},\mathbf{v},\mathbf{p}_1,\mathbf{p}_2,\mathbf{s},\boldsymbol{\pi}} \mathbf{h}_1^T\mathbf{u} + \mathbf{h}_2^T\mathbf{v} \tag{25}$$

$$\sum_{x_2} \boldsymbol{\pi} = \mathbf{p}_1 \tag{26}$$

$$\sum_{x_1} \boldsymbol{\pi} = \mathbf{p}_2 \tag{27}$$

$$E_{\boldsymbol{\pi}}\phi \leq \varphi_0 \tag{28}$$

$$u(x_1) + v(x_2) + s(x_1, x_2) = \phi(\mathbf{p}_1, \mathbf{p}_2) \tag{29}$$

$$[\mathbf{p}_1, \mathbf{p}_2] \in \hat{\mathcal{P}}_\beta \tag{30}$$

$$\boldsymbol{\pi}, \boldsymbol{s} \geq 0 \tag{31}$$

$$\boldsymbol{\pi} \leq 0 \perp \boldsymbol{s} \geq 0 \tag{32}$$

To test our econometric approach, we conduct a Monte Carlo simulation with two possibilities for points $\{0, 1\}$ and five possibilities for rate $\{3, 3.25, 3.5, 3.75, 4\}$, where each point is worth between 1/8 and 1/4 in rate. Furthermore, black and white borrowers choose each rate-point options with probability $\frac{1}{10}$ such that the null discrimination of no discrimination in menus is satisfied. We compute $\hat{\mathcal{P}}_\beta$ using the plug-in sup-t band of Montiel Olea and Plagborg-Møller (2019), let $\beta = \frac{1}{10}\alpha$, and show the simulated probability that we reject equality in menus $\mathcal{H}_0 : \varphi_0 = 0$ at the 1%, 2.5%, 5%, and 10% levels in Table 1. As Table 1 shows, our approach has approximately correct size across a wide range of sample sizes and significance levels.

Table 1: Control of our size-corrected directional derivative approaches to inference

	Significance level			
Sample size	1%	2.5%	5%	10%
$n_1 = n_2 = 500$	1.2	2.6	4.4	10.3
$n_1 = n_2 = 1000$	0.5	2.2	4.7	9.6
$n_1 = n_2 = 5000$	0.8	2.3	4.8	9.8
$n_1 = n_2 = 10000$	1.0	3.4	5.8	10.8
$n_1 = n_2 = 50000$	0.9	2.8	5.0	9.8

Note: Computed via 2000 sample draws and 500 draws of $\mathbf{h}_1, \mathbf{h}_2$ from the estimated asymptotic multivariate Normal distribution for $\mathbf{p}_1, \mathbf{p}_2$ within each sample draw.

Compared to existing methodology that could be applied to the optimal transport context, the advantage of our procedure is that it achieves uniform coverage without being overly conservative. In particular, Hsieh, Shi, and Shum (2017) has a projection method for parameter inference in mathematical programming problems, which is a broader set of problems than optimal transport, but their approach is conservative.[14] Another approach that is theoretically valid in this setting is the m-out-of-n subsampling of Politis and Romano (1994), but in addition to requiring the researcher to choose a subsample size m it can require large samples for convergence.

We look at the control of existing approaches at the 5% significance level under our simulation setting in Appendix Table A.1. In that table, HSS (2017) refers to to the projection method of Hsieh, Shi, and Shum (2017), and m out of n subsampling approach refers to the method of Politis and Romano (1994), and in the final column the size-corrected directional derivatives approach is replicated from Table 1 for comparison. Table A.1 shows that the HSS (2017) approach rejects the null with probability close to 0%, implying that it is conservative. On the other hand, the m out of n subsampling approach tends to reject at rates greater than 5% for all values of m we tried, and appears to require very large sample sizes in order to converge to the correct rejection rate.

In the following Section 5 we also report one-sided confidence intervals from the inversion of our hypothesis test. Nothing in our econometric theory precludes us from also testing the other direction and reporting two-sided confidence intervals instead.[15] However, since our economic theory is focused on getting a *lower bound* for the existence and welfare effects of discrimination, one-sided confidence intervals are particularly suitable for our purposes.

In summary, in this section we devised a new procedure for hypothesis testing in optimal transport that is uniformly valid and not overly conservative for our empirical context. Our empirical analysis in the following Section 5 shows that we are able to strongly reject equality in menus for conforming mortgages using our methodology.

[14] In our empirical context, this conservativeness tend to make the confidence intervals uninformative. Our approach is more tailored to directionally differentiable functions, and although it can be less conservative, it cannot be readily extended to all mathematical programming problems covered in Hsieh, Shi, and Shum (2017) because not all mathematical programming programs are directionally differentiable.

[15] The other direction is however computationally more challenging since it would involve taking the minimum of a maximum.

5 Empirical Estimates of Mortgage Discrimination

5.1 Data

We apply our methodology to a new dataset constructed via matching the 2018-2019 public Home Mortgage Disclosure Act (HMDA) data to OptimalBlue rate locks. The public HDMA data contains information on borrower race and ethnicity, and we take the borrowers with a HMDA derived race of “Black or African American” as our sample of Black borrowers, borrowers with a derived ethnicity of “Hispanic or Latino” as our sample sample of Hispanic borrowers, and borrowers with a HMDA derived race of “White” along with a HMDA derived ethnicity of “Not Hispanic or Latino” as our sample of Non-Hispanic White borrowers.

Starting from the uniquely matched HMDA-OptimalBlue matched dataset, we further restrict our analysis to standard 30 year, new purchase fixed rate first lien mortgages on owner-occupied site built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortizing features. Table 2 compares the HMDA data, which includes the complete set of mortgages originated in the US with such characteristics, to our matched sample. We find that our matched sample has very similar average loan sizes, LTV, rates, points, and the percent composition of Black and Hispanic borrowers compared to the HMDA data, as can be seen from Table 2. One known caveat to this is that lenders using the OptimalBlue platform tend to be smaller lenders, with the larger lenders being more likely to have their own platform for rate locking. Since these smaller lenders are not likely to keep any loans in portfolio, any use of signals not used by the GSEs to price mortgages is likely not allowed by law (Bartlett et al., 2019).

Table 2: Comparison of means between the 2018-2019 HMDA data and our HMDA-OptimalBlue matched data

	Loan Size	LTV	Rate	Points	% Black	% Hispanic	*N*
Panel A: Conforming mortgages							
HMDA data	$261,566	84.4	4.50	0.08	4.5	9.0	*3,730,152*
Matched sample	$258,205	83.5	4.59	0.12	3.9	8.6	*817,588*
Panel B: FHA mortgages							
HMDA data	$215,144	96.1	4.57	0.07	14.1	19.7	*1,437,088*
Matched sample	$220,031	95.7	4.63	0.13	13.7	18.8	*360,202*

Note: This table compares the 2018-2019 HMDA data and our HMDA-OptimalBlue matched sample for 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4, and outliers for rate below 2 and above 10.25, were excluded.

To control for the impact of lender and borrower characteristics within each loan program, we exactly match observations from Black and Non-Hispanic White borrowers without replacement on groups of covariates, taking a random observation when multiple white borrowers can be matched to a Black borrower. This creates a sample containing equal numbers of Black and Non-Hispanic White borrowers that are exactly matched on their covariates. If lenders offered Black and Non-Hispanic White borrowers the same distribution of menus conditional on covariates, our covariate-matched sample of Black and Non-Hispanic White borrowers should then have faced the same distribution of menus. The covariates that we used to match are lender, county, month of lock, and 8 categories of FICO scores and 9 categories of LTVs as defined in the GSE Loan-Level Price Adjustment (LLPA) matrix. Thus, we control for the effects of the interactions of all these covariates in our assessment of equality in menus. This set of controls is similar to what is used in Bartlett et al. (2019). Furthermore, we compute the rate spread to the Freddie Mac Weekly Survey rate during the week of the rate lock, following Bhutta and Hizmo (Forthcoming).

Summary statistics for our matched sample in Table 3. For our empirical analysis, we de-mean within each lender-county-month covariate group and round rates to the nearest eighths and points to the nearest 1/2, and we show in Table 3 that this step does not substantively change the mean differences in rates or points.[16]

As shown in Table 3, Black borrowers paid 4.8 basis points more in interest rate and 2.5 basis points more in points for conforming mortgages. On the other hand, they only paid 2.9 basis points more in interest rate and -3.2 basis points less in points for FHA mortgages. While this comparison of means cannot be interpreted as evidence for or against discrimination in menus (as we noted in Section 2), it does show that the distribution of data underlying conforming and FHA mortgages are different. Nevertheless, we will show in the next Section 5.2 that this difference in the distribution of data between conforming and FHA mortgages is however not sufficient by itself to reconcile the results of Bartlett et al. (2019) and Bhutta and Hizmo (Forthcoming), the former of which found evidence of discrimination in conforming mortgages while the latter did not in FHA mortgages. We find that the choice of heuristic used in analysis also contributes to the discrepancy in results.

[16]Both the HMDA data and the OptimalBlue data contains information about discount points paid which sometimes disagree with one another. While the literature has used both data sources, we focus on the HMDA information because it appears to have less measurement error. More specifically, we find in Appendix Table A.2 that regression of the HMDA information on points on origination charges and total loan costs has a much stronger R^2, with a coefficient closer to 1, compared to the OptimalBlue information on points. Furthermore, in a regression controlling for the HMDA points the effect of OptimalBlue points has minimal additional explanatory power for origination charges and total loan costs. We interpret this as suggestive evidence for there being more measurement error in the OptimalBlue definition of points. Our results using the OptimalBlue data on points are qualitatively similar.

Table 3: Summary statistics on the covariate matched sample

Panel A: Black and Non-Hispanic White Covariate Matched Sample						
	Conforming			FHA		
	Black	White	*Difference*	Black	White	*Difference*
Raw						
Rate Spread (bps)	36.2	31.4	*4.8*	48.3	45.4	*2.9*
Points (bps)	12.6	10.1	*2.5*	10.6	13.6	*-3.0*
De-meaned & rounded						
Rate Spread (bps)	3.1	-1.7	*4.8*	2.2	-0.7	*2.9*
Points	1.5	-0.9	*2.4*	-1.6	1.6	*-3.2*
Sample size	6,398	6,398		4,711	4,711	
Panel B: Hispanic and Non-Hispanic White Covariate Matched Sample						
	Conforming			FHA		
	Hispanic	White	*Difference*	Hispanic	White	*Difference*
Raw						
Rate Spread (bps)	38.3	34.5	*3.8*	49.3	45.4	*3.9*
Points (bps)	15.6	11.8	*3.8*	13.5	13.3	*-0.2*
De-meaned & rounded						
Rate Spread (bps)	2.5	-1.2	*3.7*	2.6	-1.2	*3.8*
Points	2.3	-1.3	*3.6*	-0.1	0.2	*-0.3*
Sample size	14,758	14,758		6,156	6,156	

Note: This table lists the summary statistics for our Black and White as well as Hispanic and Non-Hispanic White lender-county-month and covariate matched sample of 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were dropped, and all rate spreads were between -1.25 and 3. De-meaned & rounded rates and points were demeaned by lender-county-month and covariate group and then rounded into the nearest eighths for rate and 1/2 for points.

5.2 Results from heuristic analyses

We show the results from the heuristic approaches we discussed our sample in Table 4. In Table 4, Heuristic 1 in columns (1)-(2) refers to the approach of comparing points paid controlling for rate which was used in Courchane and Nickerson (1997), Black, Boehm, and DeGennaro (2003), and Bhutta and Hizmo (Forthcoming), and Heuristic 2 in columns (3)-(8) refers to the comparing means after adjusting for points using external values for the rate-point trade-off which was the general approach used in Woodward (2008), Woodward and Hall (2012), and Bartlett et al. (2019). The trade-offs used for Heuristic 2, in columns (3), (4), (5), and (6), respectively, were 1/8 and 1/4 for conforming and 1/32 and 1/4 for FHA mortgages. Finally, Columns (7)-(8) shows results from an alternative form of Heuristic 1, where we control for points and compare rate instead.

Table 4: Assessments of lender discrimination using two heuristic approaches in the Black and Non-Hispanic White matched sample

	Heuristic 1		Heuristic 2				Alternate Heuristic 1	
	Conforming	FHA	Conforming		FHA		Conforming	FHA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	points	points	rate_1/8	rate_1/4	rate_1/32	rate_1/4	rate	rate
black	4.059***	-0.732	2.977***	3.324***	2.650***	2.223***	2.728***	2.612***
	(0.989)	(1.321)	(0.325)	(0.373)	(0.434)	(0.508)	(0.320)	(0.436)
Rate Decile FE	Yes	Yes	No	No	No	No	No	No
Points Decile FE	No	No	No	No	No	No	Yes	Yes
N	12271	9200	12271	12271	9200	9200	12271	9200

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table conducts heuristic analyses of data for our Black and Non-Hispanic lender-county-month and covariate-matched sample of 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded, and rate spreads below -55 basis points and above 90 basis points were excluded.

Columns (1) and(2) of Table 4 shows that approach taken in Heuristic 1 would detect lender overcharge in conforming mortgages but not FHA mortgages: with black borrowers paying 4.1 basis points more in discount points for conforming mortgages but a statistically insignificant -0.7 basis points *fewer* discount points for FHA mortgages. Thus, Column (2) replicates Bhutta and Hizmo (Forthcoming)'s result that lenders appear to have not overcharged Black borrowers in terms points after controlling for rate in FHA mortgages. However, Columns(3)-(6) shows that Heuristic 2 would consistently detect discrimination for both conforming and FHA mortgages, with magnitudes of 2.9-3.3 basis points for conforming mortgages and 2.4-2.9 basis points for FHA mortgages. Thus, our result using Heuristic 2 is consistent with Bartlett et al. (2019) which focused on conforming mortgages. Therefore, the result of no overcharge in FHA mortgages is sensitive to the choice of heuristic used.

Furthermore, comparing Column (2) to Column (8) in Table 4 shows that the choice of which menu dimension to control for in implementing Heuristic 1 (ie. whether we control for rate and compare points, or the other way around) can lead to contradictory results for the FHA sample. In particular, using an alternative specification for Heuristic 1 where we control for the decile of points instead of rate, we do find that Black borrowers pay a 2.6 basis points higher rate after controlling for points, even though they do not significantly differ in points paid after controlling for rate. The results for the Hispanic and Non-Hispanic White matched sample where shown in Appendix Tables A.3, with similar qualitative results.

Note that we do not interpret results from Table 4 as evidence for or against the hypothesis that lenders overcharged Black borrowers by offering them worse menus of rates and points: both heuristics approaches to analysis can lead to false positives and false negatives, as we showed in Section 2, so the results from such approaches are difficult to interpret. We present results from our analysis in the next Section 5.3.

5.3 Analysis of differences in menus using our metrics

In this section we present our assessment of whether lenders offered minority borrowers worse menus of rates and points. Table 5 presents results from testing for equality in menus between black and white borrowers using our Definition 1. We find in columns (1) and (3) that our test statistic for inequality is positive and highly significant for both Black and Hispanic borrowers for conforming mortgages. More specifically, for conforming mortgages in the Black vs Non-Hispanic White matched sample, our test statistic in column (1) is $\hat{T} = 2.77$, which indicates that 2.77% of Black borrowers' choices in the data could not have been matched to that of Non-Hispanic White borrowers' that could have been on the same menu. The p-value for this test statistic is less than < 0.001, indicating that the probability that this came out of random chance is less than 0.1%. For Hispanic borrowers, our test statistic is $\hat{T} = 1.53$ in column (3), with a p-value less than < 0.001. For FHA mortgages, on the other hand, we are unable to reject equality in menus between Black and Non-Hispanic White borrowers, with a test statistic of $\hat{T} = 0$ in column (2), but are able to for Hispanic and Non-Hispanic White borrowers with a test statistic of $\hat{T} = 2.62$ in column (4).

Table 5: Results from our test of equality in menus ($\hat{T}$).

	Black vs Non-Hispanic White		Hispanic vs Non-Hispanic White	
	(1)	(2)	(3)	(4)
	Conforming	FHA	Conforming	FHA
Test statistic ($\hat{T}$)	2.77^{***}	0.00	1.53^{***}	2.62^{***}
95% CI	[2.10, ∞)	[0.00, ∞)	[0.89, ∞)	[1.16, ∞)
p-value	<0.001	1.000	<0.001	<0.001
N	12,796	9,422	29,516	12,312

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: this table shows results from our test for equality in menus based on our Definition 1, with $\hat{T}$ in units of percentage points. We use the Black and Non-Hispanic White and Hispanic and Non-Hispanic White lender-county-month and covariate-matched samples of 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded. P-values were computed using 2,000 draws from the asymptotic Normal distribution implied by $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$ using our procedure in Section 4, and confidence intervals are through inversion of the hypothesis test.

Table 6 presents our results for our differences in menus metric as in Definition 2. Column (1) shows that for conforming mortgages Black borrowers on average are willing to pay at least $\underline{D\hat{I}M}_{1\to 2} = 2.03$ basis points more in interest rates in order to get the Non-Hispanic White borrowers' menus for conforming mortgages. This again rejects equality in menus, and indicates that the distribution of menus faced by black borrowers is worse than those faced by white borrowers. Similarly, Column (3) shows that Hispanic borrowers are willing to pay at least $\underline{D\hat{I}M}_{1\to 2} = 1.52$ basis points more in order to get the Non-Hispanic White borrowers' menus. Our lower bound for how much more in interest rates Non-Hispanic White borrowers would be willing to pay to switch to minority menus, on the other hand, is consistently negative. While these magnitudes are small, as explained in Bartlett et al. (2019) even a small difference in interest rate at origination leads to a large differences in payments over the lifetime of the mortgage.

Table 6: Results for our lower bound for the average interest rate increase (bps) needed for consumers to remain indifferent after switching to another group's menus ($\underline{D\hat{I}M}$)

	Black vs Non-Hispanic White		Hispanic vs Non-Hispanic White	
	Conforming	FHA	Conforming	FHA
	(1)	(2)	(3)	(4)
Minority to white ($\underline{D\hat{I}M}_{1\to2}$)	2.03***	-2.44	1.52***	-0.91
95% CI	[1.45, ∞)	[-3.23, ∞)	[1.90, ∞)	[-1.59, ∞)
White to minority ($\underline{D\hat{I}M}_{2\to1}$)	-6.89	-7.22	-6.32	-8.59
95% CI	[-7.50, ∞)	[-7.98, ∞)	[-6.72, ∞)	[-9.27, ∞)
N	12,796	9,422	29,516	12,312

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: this table shows results for our metric for differences in menus (DIM) based on our Definition 2, with $\underline{D\hat{I}M}$ in units of basis points. We use the Black and Non-Hispanic White and Hispanic and Non-Hispanic White lender-county-month and covariate group matched samples of 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded. P-values were computed using 2,000 draws from the asymptotic Normal distribution implied by $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$ using our procedure in Section 4, and confidence intervals are through inversion of the hypothesis test.

In summary, we find that lenders offered Black and Hispanic borrowers a different distribution of menus than Non-Hispanic White borrowers for conforming mortgages, with our test statistic in columns (1) and(3) of Table 5 strongly rejecting equality in menus and our differences in menus metric in columns (1) and (3) of Table 6 suggesting that Black and Hispanic borrowers would be on average willing to pay to switch to Non-Hispanic White borrowers' menus. On the other hand, the evidence from FHA mortgages is more mixed: column (2) of Table 5 shows that we cannot reject equality in menus between Black and Non-Hispanic White borrowers for FHA mortgages, and while column (4) of Table 5 shows we are able to do so for Hispanic and Non-Hispanic White borrowers we are unable to reject a zero DIM metric in column (4) of Table 5 in terms of the average increase in rate Hispanic borrowers would be willing to pay in order to receive the menus of Non-Hispanic White borrowers.

5.4 Further analyses of differences in menus

In Section 5.3 we showed that lenders offered Black and Hispanic borrowers a less advantageous distribution of menus compared to observationally similar Non-Hispanic White borrowers for conforming mortgages, but were silent as to why this differential pricing appears. One possibility is that there are performance differences in the mortgages by race which lenders price on. However, Kau, Fang, and Munneke (2019) and Gerardi, Willen, and Zhang

(2020) shows that on originated mortgages compared to observationally similar white borrowers, black borrowers has far lower prepayment rates, making pools of black borrowers' loans more valuable than pools of white borrowers' loans. Therefore, the evidence on loan performance does not necessarily support the hypothesis of performance-based adverse pricing for minority borrowers.

To further explore the possibility that expected differences in loan performance may be driving averse pricing for Black and Hispanic borrowers, we look at the levels of differences in menus across LTV and FICO categories, and detect more discrimination among the lower LTV and higher FICO (ie. more creditworthy) borrowers. In particular, columns (1) and (3) of Table 7 shows that shows that we detect large differences in menus for conforming mortgages in the LTV under 75 and LTV between 75 and 80 categories, such that Black and Hispanic borrowers would be willing to pay 6 and 5 basis points to switch to Non-Hispanic White menus, respectively. While we can detect some differences in menus in the 80 to 90 LTV range, it is smaller,and over a 90 LTV we can no longer say that Black and Hispanic borrowers would be willing to switch to Non-Hispanic White menus. Similarly, Columns (1) of Table 8 shows that we only significantly detect differences in menus for Black borrowers in the FICO$\geq$ 740, FICO between 720 and 740, and FICO between 700 and 720 categories, and less for the FICO categories below 700. This apparent concentration of discrimination in menus within the more creditworthy group of borrowers suggests that concerns about loan default are likely not the driver behind the differential pricing received by Black and Hispanic borrowers. Rather, other factors such as differences in the negotiation and the shopping process by race may be the driver of the discrimination in menus we detect, the analysis of which we leave for future research.

Table 7: Analysis of differences in menus comparing borrowers across quartiles of minority composition by Census tract

	Black vs Non-Hispanic White		Hispanic vs Non-Hispanic White	
	(1)	(2)	(3)	(4)
	Conforming	FHA	Conforming	FHA
Panel A: Test of Equality in Menus ($\hat{T}$)				
LTV $\leq$ 75 ($\hat{T}$)	8.04***	-	7.88***	-
95% CI	[3.09, ∞)	-	[4.68, ∞)	-
75 $<$ LTV $\leq$ 80 ($\hat{T}$)	7.23***	-	7.06***	-
95% CI	[5.00, ∞)	-	[5.39, ∞)	-
80 $<$ LTV $\leq$ 90 ($\hat{T}$)	2.56***	-	0.94*	-
95% CI	[-0.02, ∞)	-	[0, ∞)	-
90 $<$ LTV $\leq$ 95 ($\hat{T}$)	0.90**	-	0.60	-
95% CI	[0.11, ∞)	-	[0, ∞)	-
LTV $>$ 95 ($\hat{T}$)	1.27**	0.17	0.63*	1.26**
95% CI	[0.19, ∞)	[0, ∞)	[0, ∞)	[0.22, ∞)
Panel B: Difference in Menus ($D\hat{I}M$) Metric				
LTV $\leq$ 75 ($\hat{T}$)	5.98***	-	5.41***	-
95% CI	[2.92, ∞)	-	[3.72, ∞)	-
75 $<$ LTV $\leq$ 80 ($\underline{D\hat{I}M}_{1\to2}$)	6.19***	-	5.18***	-
95% CI	[4.60, ∞)	-	[4.23, ∞)	-
80 $<$ LTV $\leq$ 90 ($\underline{D\hat{I}M}_{1\to2}$)	1.86**	-	0.02	-
95% CI	[0.41, ∞)	-	[-1.00, ∞)	-
90 $<$ LTV $\leq$ 95 ($\underline{D\hat{I}M}_{1\to2}$)	-0.35	-	0.12	-
95% CI	[-1.11, ∞)	-	[-0.45, ∞)	-
LTV $>$ 95 ($\underline{D\hat{I}M}_{1\to2}$)	-1.08	-2.39	-2.91	-1.19
95% CI	[-1.94, ∞)	[-3.21, ∞)	[-3.93, ∞)	[-1.85, ∞)
$N_{\text{LTV}\leq 75}$	840	-	3,142	-
$N_{75<\text{LTV}\leq 80}$	3,002	-	8,492	-
$N_{80<\text{LTV}\leq 90}$	1,714	-	3,988	-
$N_{90<\text{LTV}\leq 95}$	4,844	-	9,742	-
$N_{\text{LTV}>95}$	2,396	9,302	4,152	12,060

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Panel A shows results for our test of equality in menu ($\hat{T}$) in units of percentage points based on our Definition 1 and Panel B shows our metric for differences in menus ($\underline{D\hat{I}M}$) in units of basis points based on our Definition 2, with $\underline{D\hat{I}M}$ in units of basis points. We compare borrowers in the quartiles of minority (Black + Hispanic) percent by Census tract in the 2018-2019 HMDA data. The 1st quartile has a maximum of 3.1% minority borrowers, the 2nd quartile has between 3.1% to 9.1%, the 3rd quartile between 9.1% to 27.2%, and the 4th quartile above 27.2%. We match on lender-county-month and covariate groups for 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded. P-values were computed using 2,000 draws from the asymptotic Normal distribution implied by $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$ using our procedure in Section 4, and confidence intervals are through inversion of the hypothesis test.

Table 8: Analysis of differences in menus comparing borrowers across quartiles of minority composition by Census tract

	Black vs Non-Hispanic White		Hispanic vs Non-Hispanic White	
	(1)	(2)	(3)	(4)
	Conforming	FHA	Conforming	FHA
Panel A: Test of Equality in Menus ($\hat{T}$)				
FICO $\geq$ 740 ($\hat{T}$)	2.92***	4.81	2.70***	0.61
95% CI	[2.13, ∞)	[0, ∞)	[1.67, ∞)	[0, ∞)
720 $\leq$ FICO < 740 ($\hat{T}$)	2.94**	4.03	0.77	4.44*
95% CI	[0.04, ∞)	[0, ∞)	[0, ∞)	[0, ∞)
700 $\leq$ FICO < 720 ($\hat{T}$)	6.21**	4.24	1.32	10.77***
95% CI	[0.81, ∞)	[0, ∞)	[0, ∞)	[5.37, ∞)
660 $\leq$ FICO < 700 ($\hat{T}$)	1.49	1.57	1.96	6.15***
95% CI	[0, ∞)	[0, ∞)	[0, ∞)	[3.65, ∞)
FICO < 660 ($\hat{T}$)	7.04	0.12	8.21*	0.03
95% CI	[0, ∞)	[0, ∞)	[0, ∞)	[0, ∞)
Panel B: Difference in Menus ($D\hat{I}M$) Metric				
FICO $\geq$ 740 ($\underline{D\hat{I}M}_{1\to2}$)	1.82***	-0.19	2.23***	-4.51
95% CI	[1.16, ∞)	[-3.49, ∞)	[1.88, ∞)	[-6.97, ∞)
720 $\leq$ FICO < 740 ($\underline{D\hat{I}M}_{1\to2}$)	1.78*	-10.05	-0.45	-2.47
95% CI	[-0.06, ∞)	[-15.65, ∞)	[-1.74, ∞)	[-5.54, ∞)
700 $\leq$ FICO < 720 ($\underline{D\hat{I}M}_{1\to2}$)	2.26*	-0.84	-1.96	3.81***
95% CI	[-0.12, ∞)	[-3.07, ∞)	[-3.59, ∞)	[1.24, ∞)
660 $\leq$ FICO < 700 ($\underline{D\hat{I}M}_{1\to2}$)	-0.27	-1.69	0.49	0.51
95% CI	[-2.65, ∞)	[-3.01, ∞)	[-1.31, ∞)	[-0.57, ∞)
FICO < 660 ($\underline{D\hat{I}M}_{1\to2}$)	-1.76	-2.86	-4.96	-2.59
95% CI	[-6.25, ∞)	[-3.89, ∞)	[-9.96, ∞)	[-3.55, ∞)
$N_{\text{FICO}\geq740}$	9,914	354	22,964	662
$N_{720\leq\text{FICO}<740}$	1,082	184	2,588	360
$N_{700\leq\text{FICO}<720}$	920	428	2,002	748
$N_{660\leq\text{FICO}<700}$	738	2,884	1,680	4,008
$N_{\text{FICO}<660}$	142	5,572	282	6,534

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Panel A shows results for our test of equality in menu ($\hat{T}$) in units of percentage points based on our Definition 1 and Panel B shows our metric for differences in menus ($\underline{DIM}$) in units of basis points based on our Definition 2, with $\underline{D\hat{I}M}$ in units of basis points. We compare borrowers in the quartiles of minority (Black + Hispanic) percent by Census tract in the 2018-2019 HMDA data. The 1st quartile has a maximum of 3.1% minority borrowers, the 2nd quartile has between 3.1% to 9.1%, the 3rd quartile between 9.1% to 27.2%, and the 4th quartile above 27.2%. We match on lender-county-month and covariate groups for 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded. P-values were computed using 2,000 draws from the asymptotic Normal distribution implied by $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$ using our procedure in Section 4, and confidence intervals are through inversion of the hypothesis test.

6 Discussion

For the menu problem we identified, our methodology is unlikely to be the final word. While our approach has the advantage of requiring few assumptions to be valid, there may be other assumptions that can be used for identifying differences in menus. A promising path for future research would be to use other identifying assumptions for measuring differences in menus, or by running experiments to address the menu problem and assess its relevance in specific contexts.

At a higher level, our conceptual separation between menus and preferences is not without caveats. In many circumstances, external factors such as neighborhoods may influence both menus and preferences simultaneously (see, e.g. Katz, Kling, and Liebman (2001), Chetty, Hendren, and Katz (2016), Chetty and Hendren (2018)), such that a distinction between menus and preferences may not always be sensible. Nevertheless, it is sometimes useful, particularly from a policy perspective, to be able to designate differences in outcomes to either inequality in menus or heterogeneity in preferences. We do this for the mortgage market, and find that lenders continue to offer minorities borrowers worse menus in terms of rates and points compared to observationally similar white borrowers for conforming mortgages, particularly for more creditworthy borrowers.

References

Abadie, Alberto. Forthcoming. "Statistical Non-significance in Empirical Economics." *American Economic Review: Insights*.

Andrews, Donald W. K., and Gustavo Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica* 78(1): 119–157. doi:10.3982/ECTA7502.
Available at `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7502`.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. 2019. "Consumer-Lending Discrimination in the FinTech Era." *Working Paper*.

Bhutta, Neil, and Aurel Hizmo. Forthcoming. "Do Minorities Pay More for Mortgages?" *Review of Financial Studies*.

Black, Harold A., Thomas P. Boehm, and Ramon P. DeGennaro. 2003. "Is there discrimination in mortgage pricing? The case of overages." *Journal of Banking & Finance* 27(6): 1139 – 1165. ISSN 0378-4266. doi:https://doi.org/10.1016/S0378-4266(02)00250-9.
Available at `http://www.sciencedirect.com/science/article/pii/S0378426602002509`.

Black, Harold A, and Robert L Schweitzer. 1985. "A Canonical Analysis of Mortgage Lending Terms: Testing for Lending Discrimination at a Commercial Bank." *Urban Studies* 22(1): 13–19. ISSN 0042-0980.

Bocian, Debbie Gruenstein, Keith S. Ernst, and Wei Li. 2008. "Race, ethnicity and subprime home loan pricing." *Journal of Economics and Business* 60(1): 110 – 124. ISSN 0148-6195. doi:https://doi.org/10.1016/j.jeconbus.2007.10.001. Financing Community Reinvestment and Development.
Available at `http://www.sciencedirect.com/science/article/pii/S0148619507000847`.

Boehm, Thomas P., Paul D. Thistle, and Alan Schlottmann. 2006. "Rates and race: An analysis of racial disparities in mortgage rates." *Housing Policy Debate* 17(1): 109–149. doi:10.1080/10511482.2006.9521563.
Available at `https://doi.org/10.1080/10511482.2006.9521563`.

Cheng, Ping, Zhenguo Lin, and Yingchun Liu. 2015. "Racial discrepancy in mortgage interest rates." *The Journal of Real Estate Finance and Economics* 51(1): 101–120.

Chetty, Raj, and Nathaniel Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*." *The Quarterly Journal of Economics* 133(3): 1163–1228. ISSN 0033-5533. doi:10.1093/qje/qjy006.
Available at `https://doi.org/10.1093/qje/qjy006`.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106(4): 855–902. doi:10.1257/aer.20150572.
Available at `https://www.aeaweb.org/articles?id=10.1257/aer.20150572`.

Courchane, Marsha, and David Nickerson. 1997. "Discrimination Resulting from Overage Practices." *Journal of Financial Services Research* 11(1): 133–151. ISSN 0920-8550.

Crawford, Gregory S, Rachel Griffith, and Alessandro Iaria. 2019. "Preference estimation with unobserved choice set heterogeneity using sufficient sets." *Discussion Paper 11675, Center for Economic Policy Research*.

Fang, Zheng, and Andres Santos. 2018. "Inference on Directionally Differentiable Functions." *The Review of Economic Studies* 86(1): 377–412. ISSN 0034-6527. doi:10.1093/restud/rdy049.
Available at `https://doi.org/10.1093/restud/rdy049`.

Fuster, Andreas, Stephanie H. Lo, and Paul S. Willen. 2019. "The Time-Varying Price of Financial Intermediation in the Mortgage Market." *Working Paper*.

Gal, Tomas, and Harvey J Greenberg. 2012. *Advances in sensitivity analysis and parametric programming*, vol. 6. Springer Science & Business Media.

Galichon, Alfred. 2016. *Optimal transport methods in economics*. Princeton University Press.

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

Ghent, Andra C., Rubén Hernández-Murillo, and Michael T. Owyang. 2014. "Differences in subprime loan pricing across races and neighborhoods." *Regional Science and Urban Economics* 48: 199 – 215. ISSN 0166-0462. doi: https://doi.org/10.1016/j.regsciurbeco.2014.07.006.
Available at `http://www.sciencedirect.com/science/article/pii/S0166046214000751`.

Goldberger, Arthur S. 1984. “Reverse Regression and Salary Discrimination.” *The Journal of Human Resources* 19(3): 293–318. ISSN 0022166X.
Available at http://www.jstor.org/stable/145875.

Goldin, Claudia. 2014. “A Grand Gender Convergence: Its Last Chapter.” *American Economic Review* 104(4): 1091–1119. doi:10.1257/aer.104.4.1091.
Available at http://www.aeaweb.org/articles?id=10.1257/aer.104.4.1091.

Gul, Faruk, and Wolfgang Pesendorfer. 2001. “Temptation and Self-Control.” *Econometrica* 69(6): 1403–1435. doi:10.1111/1468-0262.00252.
Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00252.

Hsieh, Yu-Wei, Xiaoxia Shi, and Matthew Shum. 2017. “Inference on Estimators defined by Mathematical Programming.”

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. “Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment*.” *The Quarterly Journal of Economics* 116(2): 607–654. ISSN 0033-5533. doi:10.1162/00335530151144113.
Available at https://doi.org/10.1162/00335530151144113.

Kau, James B, Lu Fang, and Henry J Munneke. 2019. “An unintended consequence of mortgage financing regulation–a racial disparity.” *The Journal of Real Estate Finance and Economics* 59(4): 549–588.

McCloskey, Adam. 2017. “Bonferroni-based size-correction for nonstandard testing problems.” *Journal of Econometrics* 200(1): 17 – 35. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2017.05.001.
Available at http://www.sciencedirect.com/science/article/pii/S0304407617300556.

Montiel Olea, José Luis, and Mikkel Plagborg-Møller. 2019. “Simultaneous confidence bands: Theory, implementation, and an application to SVARs.” *Journal of Applied Econometrics* 34(1): 1–17. doi:10.1002/jae.2656.
Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2656.

Munnell, Alicia H., Geoffrey M. B. Tootell, Lynn E. Browne, and James McEneaney. 1996. “Mortgage Lending in Boston: Interpreting HMDA Data.” *The American Economic Review* 86(1): 25–53. ISSN 00028282.
Available at http://www.jstor.org/stable/2118254.

Politis, Dimitris N., and Joseph P. Romano. 1994. "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions." *Ann. Statist.* 22(4): 2031–2050. doi:10.1214/aos/1176325770.
Available at `https://doi.org/10.1214/aos/1176325770`.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2014. "A Practical Two-Step Method for Testing Moment Inequalities." *Econometrica* 82(5): 1979–2002. doi:10.3982/ECTA11011.
Available at `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11011`.

Santambrogio, Filippo. 2015. "Optimal transport for applied mathematicians." *Birkäuser, NY* 55(58-63): 94.

Shapiro, Alexander. 1990. "On concepts of directional differentiability." *Journal of Optimization Theory and Applications* 66: 477–487. doi:10.1007/BF00940933.

Shapiro, Alexander. 1991. "Asymptotic analysis of stochastic programs." *Annals of Operations Research* 30: 169–186.

Sommerfeld, Max, and Axel Munk. 2018. "Inference for empirical Wasserstein distances on finite spaces." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1): 219–238. doi:10.1111/rssb.12236.
Available at `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12236`.

Tameling, Carla, Max Sommerfeld, and Axel Munk. 2019. "Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications." *The Annals of Applied Probability* 29(5): 2744–2781. doi:10.1214/19-AAP1463.

Tootell, Geoffrey M. B. 1996. "Redlining in Boston: Do Mortgage Lenders Discriminate Against Neighborhoods?*." *The Quarterly Journal of Economics* 111(4): 1049–1079. ISSN 0033-5533. doi:10.2307/2946707.
Available at `https://doi.org/10.2307/2946707`.

Villani, Cédric. 2008. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media.

Woodward, Susan E. 2008. "A Study of Closing Costs for FHA Mortgages." *U.S. Department of Housing and Urban Development Report*.

Woodward, Susan E., and Robert E. Hall. 2012. "Diagnosing Consumer Confusion and Sub-optimal Shopping Effort: Theory and Mortgage-Market Evidence." *American Economic*

Review 102(7): 3249–76. doi:10.1257/aer.102.7.3249.
Available at http://www.aeaweb.org/articles?id=10.1257/aer.102.7.3249.

A.1 Proofs of results

A.1.1 Proof of Theorem 1

Proof. In the forward direction, if such a $\pi(x_1, x_2)$ exists, then it is possible for there to be a series of menus $\mathbf{m} = \{x_1, x_2\}$ where $\phi(x_1, x_2) = 1$, each appearing with probability $g_1(\mathbf{m}) = g_2(\mathbf{m}) = \pi(x_1, x_2)$, in which group 1 consumers chose x_1 and group 2 consumers chose x_2. Under this construction, the distribution of menus across the two groups are equal such that $\mathbf{M}_1 = \mathbf{M}_2$, and the choice probabilities are rationalized. The reverse direction follows from the fact that, denoting $c(x_1, x_2|\mathbf{m})$ by the probability that group 1 consumers choose $x_1 \in \mathbf{m}$ and group 2 consumers choose $x_2 \in \mathbf{m}$ given a menu $\mathbf{m}$, we can compute such a $\pi(x_1, x_2) = \sum_{\mathbf{m}} c(x_1, x_2|\mathbf{m}) g(\mathbf{m})$ for any $c(x_1, x_2|\mathbf{m}), g(\mathbf{m})$ where $g(\mathbf{m}) = g_1(\mathbf{m}) = g_2(\mathbf{m})$. □

A.1.2 Proof of Theorem 3

Proof. First, we show that φ is Gâteaux directionally differentiable in the sense that the limit:

$$\varphi'_{\mathbf{p}_1,\mathbf{p}_2}(\mathbf{h}_1, \mathbf{h}_2) = \lim_{t \to 0^+} \frac{\varphi(\mathbf{p}_1 + th_1, \mathbf{p}_2 + th_2) - \varphi(\mathbf{p}_1, \mathbf{p}_2)}{t} \tag{33}$$

exists for all and is equal to that given by Equations (18) for $\{\mathbf{p}_1, \mathbf{p}_2\} \in \mathbb{D}_\varphi$ and $\{\mathbf{h}_1, \mathbf{h}_2\} \in \mathbb{D}_0$. To do this, without loss of generality letting $\phi^* = \phi + M, M = \sup|\phi|$ such that $\phi^* \geq 0$ and $\varphi = \varphi^* - M$, we transform the problem to standard linear programming form:

$$\varphi^* = \min_{\boldsymbol{\pi}} E_{\boldsymbol{\pi}} \phi^* \text{ s.t. } E_{x_2}\boldsymbol{\pi} \geq \mathbf{p}_1, E_{x_1}\boldsymbol{\pi} \geq \mathbf{p}_2, \boldsymbol{\pi} \geq 0 \tag{34}$$

and then use Theorem 3.1 of Gal and Greenberg (2012) which sets out conditions for the Gâteaux directional differentiability of standard linear programs with inequality constraints. In particular, we need to check primal and dual stability in the sense that, let $\Pi^*(p_1, p_2)$ be the set of primal solutions to the linear programming problem in Equation (34), then the set of primal solutions reachable from perturbations in the direction $\{h_1, h_2\}$ is non-empty, such that:

$$\Pi^\infty(\{\mathbf{p}_1, \mathbf{p}_2\}, \{\mathbf{h}_1, \mathbf{h}_2\}) = \{\boldsymbol{\pi} : \{\boldsymbol{\pi}^k\} \to \boldsymbol{\pi} \text{ for some } \{\epsilon_k \to 0^+\}, \tag{35}$$

$$\text{with } \boldsymbol{\pi}^k \in \boldsymbol{\Pi}^*(\mathbf{p}_1 + \epsilon_k \mathbf{h}_1, \mathbf{p}_2 + \epsilon_k \mathbf{h}_2)\} \neq \emptyset \tag{36}$$

and analogously for dual solutions. This can be done by referencing existing results:

1. Since $\{\mathbf{p}_1 + \epsilon_k \mathbf{h}_1, \mathbf{p}_2 + \epsilon_k \mathbf{h}_2\}$ are a series of probability measures for $\epsilon_k \leq 1$, primal stability in the sense of Equation (35) is guaranteed by Theorem 5.19 in Villani (2008).
2. Similarly, dual stability is guaranteed by Theorem 1.52 in Santambrogio (2015), in particular by taking the sequence of c-concave Kantorovich potentials corresponding to $(\mathbf{p}_1 + \epsilon_k \mathbf{h}_1, \mathbf{p}_2 + \epsilon_k \mathbf{h}_2)\}$.

Second, we show that φ is Lipschitz, such that the Gâteaux directionally differentiability of φ is equivalent to Hadamard directionally differentiability following Shapiro (1990). For the l_1 norm $l_1(\{\mathbf{p}_{1,1}, \mathbf{p}_{1,2}\}, \{\mathbf{p}_{2,1}, \mathbf{p}_{2,2}\}) = \sum_x |\mathbf{p}_{1,1}(x) - \mathbf{p}_{1,2}(x)| + |\mathbf{p}_{2,1}(x) - \mathbf{p}_{2,2}(x)|$, we will show that:

$$|\varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_{2,1}, \mathbf{p}_{2,2})| \leq M l_1(\{\mathbf{p}_{1,1}, \mathbf{p}_{1,2}\}, \{\mathbf{p}_{2,1}, \mathbf{p}_{2,2}\}) \tag{37}$$

More specifically, let $\mathbf{p}_1^- = \min\{\mathbf{p}_{1,1}, \mathbf{p}_{1,2}\}, \mathbf{p}_1^+ = \max\{\mathbf{p}_{1,1}, \mathbf{p}_{1,2}\}$, and analogously for $\mathbf{p}_2^-, \mathbf{p}_2^+$. By construction we know that $\varphi(\mathbf{p}_1^-, \mathbf{p}_2^-) \leq \varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}), \varphi(\mathbf{p}_{2,1}, \mathbf{p}_{2,2})$ and $\varphi(\mathbf{p}_1^+, \mathbf{p}_2^+) \geq \varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}), \varphi(\mathbf{p}_{2,1}, \mathbf{p}_{2,2})$, and therefore:

$$\varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_1^+, \mathbf{p}_2^+) \leq \varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_{2,1}, \mathbf{p}_{2,2}) \leq \varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_1^-, \mathbf{p}_2^-) \tag{38}$$

Furthermore, we know that $\varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_1^-, \mathbf{p}_2^-) \leq M l_1$ since taking the optimal plan from $\pi^- = \varphi(\{\mathbf{p}_1^-, \mathbf{p}_2^-\})$ and then constructing a plan $\pi^{-,*} = \pi^- + (\mathbf{p}_{1,1} - \mathbf{p}_1^-)(\mathbf{p}_{1,2} - \mathbf{p}_2^-)$ yields an upper bound for the value value $\varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) \leq E_{\pi^{-,*}} \phi = \varphi(\{\mathbf{p}_1^-, \mathbf{p}_2^-\}) + \sum(\mathbf{p}_{1,1} - \mathbf{p}_1^-)(\mathbf{p}_{1,2} - \mathbf{p}_2^-)\phi \leq \varphi(\{\mathbf{p}_1^-, \mathbf{p}_2^-\}) + M \sum(\mathbf{p}_{1,1} - \mathbf{p}_1^-) \sum(\mathbf{p}_{1,2} - \mathbf{p}_2^-) \leq \varphi(\{\mathbf{p}_1^-, \mathbf{p}_2^-\}) + M l_1$. Similarly, we have that $\varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_1^+, \mathbf{p}_2^+) \geq -M l_1$. Substituting into Equation (38) yields:

$$-M l_1 \leq \varphi(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}) - \varphi(\mathbf{p}_{2,1}, \mathbf{p}_{2,2}) \leq M l_1 \tag{39}$$

Which implies Equation (37) and that the mapping φ is Lipschitz.

□

A.1.3 Proof of Corollary 1

Proof. By law of total probability and property of limsup, we know that:

$$\limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left(\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}\right) \tag{40}$$

$$\leq \limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left([\mathbf{p}_1,\mathbf{p}_2] \notin \hat{\mathcal{P}}_{n,\beta}\right) + \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left((\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}) \cap [\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}\right) \tag{41}$$

$$\leq \limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left([\mathbf{p}_1,\mathbf{p}_2] \notin \hat{\mathcal{P}}_{n,\beta}\right) + \limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left((\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}) \cap [\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}\right) \tag{42}$$

By Equation (21) for the uniform confidence band, we know that:

$$\limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left([\mathbf{p}_1,\mathbf{p}_2] \notin \hat{\mathcal{P}}_{n,\beta}\right) \leq \beta \tag{43}$$

By law of conditional probability, then Equation (20) and the Portmanteau Theorem, we have:

$$\limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left((\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}) \cap [\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}\right) \tag{44}$$

$$= \limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left((\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}) | [\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}\right) \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left([\mathbf{p}_1,\mathbf{p}_2] \in \hat{\mathcal{P}}_{n,\beta}\right) \tag{45}$$

$$\leq \alpha - \beta \tag{46}$$

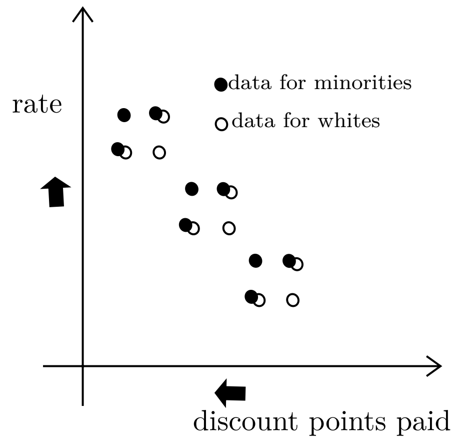
Combining the two parts, we have:

$$\limsup_{n\to\infty} \Pr_{\mathbf{p}_1,\mathbf{p}_2}\left(\hat{\varphi}_n \geq \hat{c}_{n,1-\alpha+\beta}\right) \leq \beta + (\alpha - \beta) = \alpha \tag{47}$$

As required. □

A.2 Additional Tables and Figures

Figure A.1: How the choice of which menu dimension to condition on can yield contradictory results



Note: this Figure shows that a "contradictory" assessment of discrimination can appear nonlinearly, and complements the linear regression case of Figure 2.

Table A.1: Other approaches to inference, compared to our size-corrected directional derivatives approach

	HSS (2017)	m out of n subsampling			Size-corrected directional derivatives
		$m = n^{2/3}$	$m = n^{1/2}$	$m = n^{1/3}$	
$n_1 = n_2 = 500$	0.0	31.8	21.9	16.2	4.4
$n_1 = n_2 = 1000$	0.0	28.3	18.4	12.5	4.7
$n_1 = n_2 = 5000$	0.0	20.0	11.3	8.6	4.8
$n_1 = n_2 = 10000$	0.1	18.3	10.3	8.4	5.8
$n_1 = n_2 = 50000$	0.0	13.7	8.6	6.5	5.0

Note: Entries represent the probability of rejecting the null at 5% level. The control of the subsampling approach is taken with subsample size m as indicated, with $n = \frac{n_1 n_2}{n_1+n_2} = \frac{1}{2}n_1$. The control of our size-corrected directional derivatives approach were computed via 2000 sample draws and 500 draws of h_1, h_2 from the estimated asymptotic multivariate Normal distribution for p_1, p_2 within each sample draw.

Table A.2: Regressions of origination costs and total loan costs as a percent of the loan amount on HMDA's information on points (hmda_points) versus OptimalBlue's information on points (ob_points)

	Origination costs			Total loan costs		
	(1)	(2)	(3)	(4)	(5)	(6)
hmda_points	0.915***		0.902***	0.940***		0.941***
	(0.00277)		(0.00308)	(0.00372)		(0.00463)
ob_points		0.469***	0.0200***		0.474***	0.00531
		(0.00377)	(0.00220)		(0.00434)	(0.00361)
_cons	0.580***	0.560***	0.575***	2.194***	2.174***	2.190***
	(0.000432)	(0.00129)	(0.000591)	(0.000577)	(0.00149)	(0.000927)
N	1224911	1221338	1221208	1224417	1220921	1220715
R^2	0.553	0.231	0.553	0.233	0.093	0.232

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: The sample consists of the conforming and FHA purchase mortgages originated within the Retail channel within our 2018-2019 HMDA-OptimalBlue matched sample. In each regression we excluded observations with extreme outliers for points (below -4 or above 4) and for origination costs and total loan costs as a percent of the loan amount below -3% or above 10%. All regressions include lender by county by year by product type fixed effects. Standard errors were also clustered at the lender by county by year by product type level.

Table A.3: Assessments of lender discrimination using two heuristic approaches in the Hispanic and Non-Hispanic White matched sample

	Heuristic 1		Heuristic 2				Alternate Heuristic 1	
	Conforming	FHA	Conforming		FHA		Conforming	FHA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	points	points	rate_1/8	rate_1/4	rate_1/32	rate_1/4	rate	rate
hispanic	5.153***	2.021*	2.945***	3.437***	3.449***	3.331***	2.623***	3.334***
	(0.633)	(1.113)	(0.225)	(0.432)	(0.254)	(0.369)	(0.221)	(0.370)
Rate Decile FE	Yes	Yes	No	No	No	No		
Points Decile FE	No	No	No	No	Yes	Yes		
N	28273	12055	28273	12055	28273	12055	28273	12055

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table conducts heuristic analyses of data for our Hispanic and Non-Hispanic lender-county-month and covariate matched sample of 30 year, new purchase, fixed rate, first lien mortgages on owner-occupied site-built properties without prepayment penalties, balloon, interest only, negative amortization, or non-amortization features. Outliers for points above 4 and below -4 were excluded, and rate spreads below -55 basis points and above 90 basis points were excluded.