Are University Admissions Academically Fair?*

Debopam Bhattacharya, Shin Kanaya, Margaret Stevens
University of Oxford

February 17, 2012.

Abstract: Selective universities are often accused of unfair admission practices which favour applicants from specific socioeconomic groups. We develop an empirical framework for testing whether such admissions are academically fair, i.e., they equalize the expected performance of the marginal admitted candidates – the admission-threshold – across socioeconomic groups. We show that such thresholds are nonparametrically identified from admissions data if unobserved, officer-specific heterogeneity affecting admission decisions is median-independent of applicant covariates and the density of past-admits' conditional expected performance is positive around the admission-threshold for each socioeconomic group. Applying our methods to admissions-data for a large undergraduate programme at Oxford and using first-year exam-performance as outcome, we find that the admission-threshold for male (respectively, private-school) applicants is about 3.9 (1.6) percentage points higher than for female (state-school) applicants. In contrast, average admission-rates are equal across gender and school-type, both before and after controlling for applicants' background-characteristics. The implied productivity loss is 1.8 (1) percentage points per applicant.

Keywords: University admissions, academic fairness, marginal admit, conditional median restriction, nonparametric identification, bounding productivity loss.

*Address for correspondence: Debopam Bhattacharya, Department of Economics, University of Oxford. Manor Road Building, Manor Road, OX1 3UQ, United Kingdom. email: debobhatta@gmail.com

1 Introduction

Background: Selective universities are frequently accused of biased admission practices which favour applicants from advantaged backgrounds and thus contribute to the perpetuation of socioeconomic inequality.[1] Universities usually respond to such allegations by claiming to admit students with the best academic potential, irrespective of their socioeconomic backgrounds.[2] Despite significant media and political interest in the issue, there does not seem to exist a rigorous empirical methodology for testing these claims on the basis of admissions data. Our purpose in this paper is to construct a formal econometric framework within which the "academic fairness" of admissions may be defined and empirically tested, based on pre-admission background data for all applicants and college-performance data for the admitted ones.

The notion of fairness we focus on – in accordance with the universities' claims – is an outcome-oriented one. Roughly speaking, it dictates that the marginal admitted individuals in different demographic groups (e.g. male and female) of applicants should have identical expected outcomes, where the expectation is computed based on characteristics observed by admission-officers at the application stage. This common value will be referred to as the admission threshold.

In economics, equalized marginal returns is a well-understood generic condition for optimal allocations. In the specific context of treatment assignment, it is equivalent to requiring the treatment-regime to maximize the expected value of the relevant population outcome subject to budget

[1] For example, in the UK, a highly publicized 2011 Sutton Trust report shows that 100 elite (mostly expensive private) schools - just 3% of schools for the relevant age-group - account for 31.9% of admissions to Oxford and Cambridge. Source:

http://www.suttontrust.com/news/news/four-schools-and-one-college-win-more-places-at-oxbridge

1.

[2] In British, European and Asian universities, undergraduate admissions are typically subject-specific and almost entirely academically focused. Extra-curricular achievements, leadership potential etc. typically play no role in admissions. For example, Oxford claims to be "...committed to recruiting the most able students, regardless of background", while Cambridge claims that its "aim is to offer admission to students of the greatest intellectual potential, irrespective of social, racial, religious and financial considerations". The closest US equivalent would be admission to post-graduate academic programs. Source:

A. http://www.ox.ac.uk/about_the_university/facts_and_figures/undergraduate_admissions_statistics/index.html

B. http://www.cam.ac.uk/admissions/undergraduate/apply/

constraints, c.f., Bhattacharya and Dupas (2012) and Bhattacharya (2011). However, empirically detecting who are the relevant marginal candidates and calculating their expected outcomes are difficult problems in general. The first challenge is that the definition of marginal is intertwined with the approval process and depends on all characteristics observed by the approval-officers, some of which may not be observable to an analyst (c.f., Heckman (1998), Persico (2009) and the references therein). Second, in some situations, it is difficult to observe the relevant outcomes or calculate their expected value. An example is the case of hiring workers, where it is difficult to measure an individual worker's productivity even after she is hired. Further, counter-factual outcomes such as potential productivity of rejected applicants are in fact never observed. Third, approval decisions for a large cohort of applicants, e.g., for university places, are usually made simultaneously by several officials who apply at least some personal discretion and/or display heterogeneity in taste or knowledge. This heterogeneity is likely to introduce idiosyncratic variation in individual decisions around a baseline university-wide policy and make the approval stochastic, even after conditioning on all applicant covariates observable to the officials. Defining and identifying the "overall" marginal candidates in such scenarios is a nontrivial problem.

In the university-application case, the first problem is essentially mitigated when the analyst can access the same application forms and standardized test-scores as those used by the admission officers. For example, an economist studying admissions in her own university can easily access these data, especially if she herself is involved in conducting admissions. Furthermore, in large universities, admission decisions for thousands of applicants are typically made within a short period of time. Consequently, it is difficult to fine-tune the admission process to judge each candidate based on a different set of characteristics and this leads to standardized assessment procedures based on a generic set of background variables. Therefore in this case, access to applicant records largely eliminates the unobserved applicant characteristics issue that plague studies of unfairness in some other situations, such as medical treatment, where patients are treated sequentially and different *criteria* are used to judge treatment appropriateness depending on the patient's age, ethnic and health background or gender. This reasoning further suggests that our methods can be directly used in all treatment situations where (i) approval criteria are standardized, (ii) relevant characteristics of the applicants are obtained through application forms and (iii) the forms are accessible to the analyst. Two pertinent examples are the approval of housing or consumer loans (c.f. Jiang et al (2011), discussed below) and the issuance of insurance coverage. In subsection 3.2 below we discuss a substantive assumption under which our methods can be applied to some other types of

treatment-assignment situations which do not have these three characteristics.

A second advantage of the admission case is that one can easily match pre-application records with college outcomes of admitted candidates, thereby partially mitigating the unobserved outcome problem. The mitigation is partial because potential outcomes of rejected applicants will still remain unobserved.

Finally, the difficulty in defining and detecting marginal candidates under unobserved heterogeneity across admission officers, still needs to be resolved.

Our contribution: In the present paper, we construct an empirical model of admissions involving (i) observed applicant covariates, (ii) unobserved heterogeneity across admission officers and (iii) outcomes of past admitted students. We allow for the fact that not all admission offers translate into enrolment because applicants may accept alternative offers or fail to satisfy conditions specified in the current offer, such as securing a certain grade in the school-leaving public examination. Our primary contribution in this setting is to show that under reasonable behavioral assumptions and under "continuous density" type regularity conditions, the baseline admission threshold faced by applicants from a specific demographic group can be nonparametrically identified from admission data for current applicants and post-enrolment performance of past admitted students from that demographic group. It is not necessary to identify potential college outcomes of rejected candidates. A test of fairness can then be carried out by checking equality of the identified thresholds across the groups. Our key behavioral assumptions are that (a) admission officers form their subjective expectations on the basis of academic outcomes of past admits and (b) for each type of applicant, the expectational errors, i.e., the differences between the officers' subjective expectations and the true mathematical expectations, have zero median – i.e., the errors are equally likely to be positive or negative.[3] It is important to note that the latter assumption allows the distribution, and in particular the variance, of such errors to differ by demographic group, which is an important generalization. Indeed, one would expect that this variance is larger for historically under-represented groups, reflecting larger magnitudes of error in an officer's subjective beliefs regarding those types of individuals with whom the officer has had less experience. To our knowledge, the existing literature on detecting treatment fairness has largely ignored such unobserved, treater-specific heterogeneity.

In the context of university admissions, equal marginal returns – as emphasized above – is consistent with maximization of the average academic performance of the entering cohort. Therefore,

[3] If the expectational errors are systematically higher for one group, we can absorb that difference into our definition of thresholds. Thus the assumption of a zero value for the median is an innocuous normalization.

when admission thresholds are found to differ by demographics – e.g., due to affirmative action policies – it is useful to quantify the extent of discrepancy by the magnitude of the expected academic achievement foregone as a result. Under a stronger assumption of full independence between officer errors and applicant covariates, we outline a method of obtaining nonparametric bounds on this shortfall. Calculation of such efficiency losses appear to be novel in regards to the literature on detecting treatment fairness.

As a final step in our analysis, we apply the methods developed above to analyze admissions data from one large undergraduate programme of study at Oxford University, focusing on first year academic performance as the outcome of interest. The overall application success rates are seen to be almost identical across gender and type of school, both before and after controlling for key covariates. However, upon focusing on the marginal admitted candidates, we find that expected performance thresholds faced by applicants who are male or from independent schools exceed those faced by females or state school applicants. The magnitude of the gender difference is about 0.8 standard deviations and that for school-types about 0.2 standard deviations of the outcome. This finding is suggestive of some degree of affirmative action – either explicit or implicit – within the admissions process, which is not apparent from the equal success rates, thereby illustrating the usefulness of our approach.

Related Literature: On the econometric front, the present paper exactly complements a recent literature – pioneered by Manski (2004) – on the reverse problem of how treatment should be targeted for future populations, using information from past treatment outcomes. On the economic front, our proposed empirical methodology complements the existing *theoretical* literature on the economic analysis of affirmative actions in college admissions. Fryer and Loury (2005) have provided a critical review of this theoretical literature and a comprehensive bibliography. In regards to the educational literature, our work is complementary to a large volume of research on the usefulness of standardized test scores such as the SAT in predicting academic success in college and how this predictability varies across race and gender. See, for instance, staff research papers published online at the Central Institute of Education in the UK and the College Board in the US; Rothstein (2004) provides a comprehensive and critical review of this literature. Finally, the present work is substantively related to the empirical literature on the detection of so-called "taste-based" motives in law enforcement and in medical treatment, c.f., Persico (2009) (see section 3 below for further details and references).

Plan of the Paper: The rest of the paper is organized as follows. Section 2 sets up the formal

problem and defines the key parameters of interest. Section 3 discusses identification of admission thresholds using applicant-level admissions data, discusses the applicability of our methods in other types of treatment scenarios and contrasts our approach with alternative identification strategies in the empirical microeconomics literature. Section 4 deals with inference. Section 5 considers (partial) identification of the short-term productivity loss due to the use of different thresholds for different groups, as would happen under affirmative action policies. Section 6 contains the substantive application of our methods to admission at an Oxford undergraduate programme. Section 7 concludes. All technical material are collected in the appendix. We will use the notation $A := B$ to mean that A equals B by definition.

2 Set-Up

Let W denote an applicant's covariates which are observed by the university. Let $\mathcal{W}$ denote the support of W, $P_W(\cdot)$ denote the C.D.F. of W and let G denote one or more discrete components of W capturing the group identity of the applicant (such as sex, race or type of high school attended) which forms the basis of the alleged mistreatment. For an applicant i, let Y_i denote his/her future academic performance if admitted to the university. Let $\mu(w)$ denote a w-type student's true expected performance if he/she enrols.

To set up the empirical framework, we assume that we have data on pre-admission characteristics on a pool of applicants, drawn in an I.I.D. fashion from a distribution of potential applicants. For each applicant, we also observe whether they were made an offer of admission, for each offer we observe whether it was accepted and finally, for each accepted offer, we observe the outcome of interest (e.g., examination score after 1st year of university). We have such data for several years. When referring to variables from past years or expectations calculated on the basis of past variables, we will use the superscript P.

For a current applicant i in the data, let X_i, G_i denote his/her values of pre-admission characteristics and let D_i denote the dummy for whether (s)he received an offer and the dummy A_i equals 1 if and only if i was made an offer and accepted it. We observe (X_i, G_i, D_i, A_i) for every individual and the outcome Y_i if $A_i = 1$. Let

$$\mu^P(x, g) := E\left(Y^P | X^P = x, G^P = g, A^P = 1\right) \tag{1}$$

denote the conditional expectation of outcome Y for previous years' admits with characteristics x, g. We assume that when admission officers decide on whether to admit an (x, g) type student

in the current year, they base it on their subjective assessment of $\mu^P(x,g)$ which they surmise from data on (x,g) type students who were admitted in previous years. Note that $\mu^P(x,g)$ is in general different from $E(Y|X=x,G=g)$ which is typically unknown to admission officers in universities (or loan officers in banks in our loan application example above).[4] Indeed, a large literature in educational statistics on so-called "validation studies" use predicted performance of *admitted* candidates to infer the relative predictive ability of standardized test scores vis-a-vis high school grades and socioeconomic indicators and prescribe policies based on this analysis. See for example, Kobrin et al (2001), Kuncel et al (2008) and Sawyer (1996, 2010). Since our analysis evaluates what educational institutions actually do – rather than what they should have done – using $\mu^P(x,g)$ rather than $E(Y|X=x,G=g)$ is the correct approach here.

Toward that end, let $\mathcal{X}_g$ denote the support of X conditional on $G=g$ and $A^P=1$, i.e.,

$$\mathcal{X}_g = \left\{x : \Pr\left(A^P=1|X^P=x,G^P=g\right)>0\right\}.$$

These are the values of X which occur among the admits of type g in past years and so one can, in principle, calculate the values of $\mu^P(x,g)$ when $x\in\mathcal{X}_g$. We assume that a g-type applicant with a value of X outside $\mathcal{X}_g$ will be not offered admission with probability 1 (such as applicants with very low admission test scores) in the current year. On the other hand, an applicant i with $G_i=g$ and $X_i=x\in\mathcal{X}_g$ is offered admission if and only if $\mu^P(x,g)\geq\gamma_g+\varepsilon_i$, where ε_i denotes unobserved heterogeneity across admission officials and γ_g denotes the university-wide baseline threshold for applicants of demographic type g. Thus we may summarize the admission process as:

Assumption 1

$$D_i = \begin{cases} \mathbf{1}\left\{\mu^P(X_i,G_i)\geq\gamma_{G_i}+\varepsilon_i\right\},\ \textit{if } X_i\in\mathcal{X}_{G_i} \\ 0,\ \textit{if } X_i\notin\mathcal{X}_{G_i} \end{cases}. \tag{2}$$

Academically Fair Admissions: In this setting, we define an admission practice to be academically fair at the university level if and only if γ_g is identical across g. The underlying intuition is that the only way covariates G should influence the admission process is through their effect on the expected academic outcome. Having a larger γ for, say, females than males implies that a male

[4] If there existed trial data where admissions were randomized, then the latter can be calculated and used instead of $\mu^P(x,g)$. Alternatively, if Y^P were independent of A^P, given X,G (the so-called selection-on-observables case), then the two would be identical but this is somewhat irrelevant to the task at hand since admission officers are likely to act on the basis of $\mu^P(x,g)$, whether or not it equals $E(Y|X,G)$.

applicant with identical expected outcomes based on observable characteristics as a female applicant is more likely to be admitted. Conversely, under affirmative action, γ_g will be lower for those gs which represent historically disadvantaged groups. Therefore, we are interested in estimating the threshold γ_g for various values of g and testing if they are identical across g.

It is important to note that here we are not making any assumption about whether or not G affects the distribution of Y, conditional on X. In our set-up, a male applicant with identical X as a female candidate can have a higher probability of being admitted and yet the admission process may be academically fair if males have a higher expected outcome than females with identical X. This contrasts sharply with the notion of fairness employed, for example, in Bertrand et al (2005) which concluded racial bias if two job-applicants with identical CVs but of different race had different probabilities of being called for interview. In order for this finding to imply unfairness according to our criterion, one needs to assume that, conditional on the information in CVs, race has no impact on average worker productivity – a strong assumption, indeed.

For the identification/estimation of γ_g, we impose the following conditions:

Assumption 2 *(i)* $\{(X_i, G_i, D_i, A_i, Y_i A_i)\}_{i=1}^n$ *is an i.i.d. sequence, with* $E(|Y_i| \, |X_i) < \infty$*, almost surely; (ii)* $\text{median}(\varepsilon_i \mid X_i, G_i) = 0$ *almost surely and (iii) the distribution of* ε_i *has a strictly positive density (with respect to the Lebesgue measure) around* 0*, given* X_i *and* G_i*, almost surely.*

Discussion: The presence of ε in (2) allows admission to be non-deterministic, given X and G; in particular, the admission cut-off for expected performance faced by applicants of type g varies randomly within g-type applicants (depending on which officer handled their file) with a median value of γ_g stipulated by a university-wide admission policy. For academically fair admissions, as explained above, one would expect γ_g to be identical across g. Part (ii) of Assumption 2 will hold when the systematic determinants of admission, such as past test scores, interview grades and demographic characteristics are observed by the econometrician but idiosyncratic preferences and/or the deviation of the admission officers' subjective expectation from the true $\mu^P(\cdot, \cdot)$ are not. Assumption (ii) basically says that for each type of applicant, an admission officer is equally likely to make positive and negative errors in calculating the expected academic performance. If applicant files are distributed randomly among admission officers, then ε is likely to be independent of (X, G), which is sufficient for Assumption 2 part (ii) to hold. One case where full independence might fail is when for some historically under-represented group g, the variance of ε is larger, reflecting larger magnitudes of error in an officer's subjective beliefs regarding those types of individuals with

whom the officer has had less experience. The conditional median restriction is robust to such scale dependence, as is well-known from Manski's (1975) maximum score analysis, and turns out to be sufficient here for identifying γ_gs for each value of g. Notice that the type of scale dependence mentioned above would be ruled out by the independence of ε and (X, G) or by the index-restriction assumption in Chandra and Staiger (2009, page 7). Observe also that this zero-median restriction is weaker than requiring the error distributions to be symmetric about zero and thus allows for arbitrary amounts of skewness.

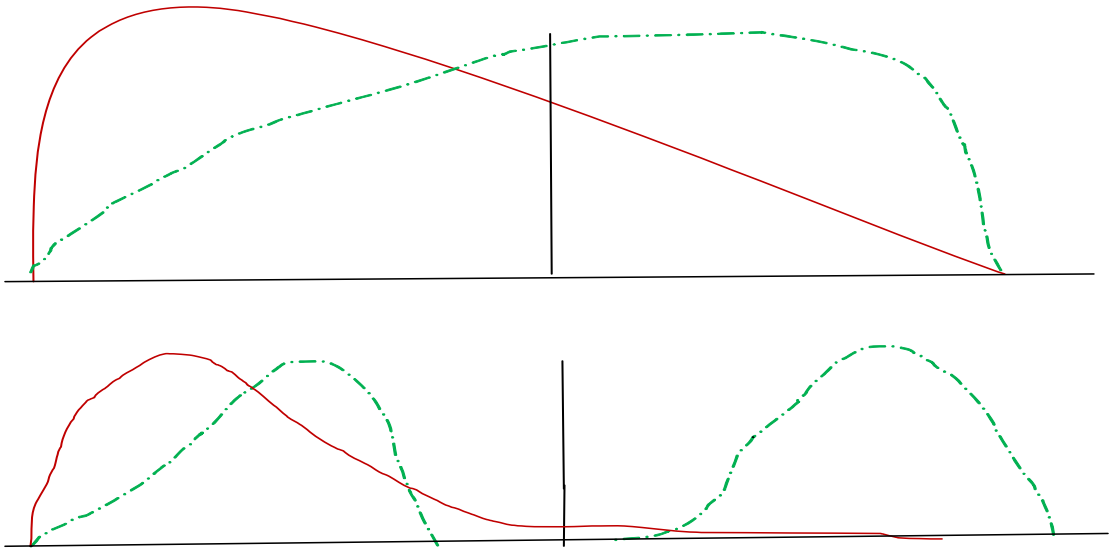
Finally, part (iii) of Assumption 2 is a regularity condition that aids the proof of identification. It will obviously hold for every standard continuous distribution.

Lastly, we will make a technical assumption which would imply that there exists a common feasible threshold. Toward that end, let Ω_g, Υ_g denote respectively the support for the distribution of X and of $\mu^P(X, G)$, given $G = g$.

Assumption 3 *$\Upsilon_g \equiv \Upsilon$ for all g; Υ contains an interval I such the density of $\mu^P(X, g)$ conditional on $G = g$ is strictly positive on I and γ_g lies in I for each g.*

To interpret this assumption, consider the case where G denotes gender and X contains one or more continuous variables like previous test scores. Then the above assumption basically says that the (conditional) expected outcome for men and that for women take values in the same set. So given any value $x \in \Omega_{male}$, there exists an $x' \in \Omega_{female}$ such that $\mu^P(x, male) = \mu^P(x', female)$. Note that this does not require Ω_g to be identical across g.

Under the above assumption, for every fixed g, there will exist $x^*(g) \in int(\Omega_g)$ such that $\mu^P(x^*(g), g) = \gamma$. So we can define individuals in subgroup g with $X = x^*(g)$ to be the "ideally marginal" admits in subgroup g, i.e., those (x, g)s who would be marginal in the absence of any ε, as would occur if the university conducted admissions as a single entity and had perfect knowledge of $\mu^P(\cdot, \cdot)$. If admissions are academically fair, then for every g, $\mu^P(x^*(g), g) = \gamma$; if not, and the marginal admits are denoted by $x(g)$ in group g, then $\mu^P(x(g), g) \equiv \gamma_g$ will differ across g. If the common support assumption did not hold, then it would be possible that admission is academically fair with a common γ which lies within the support of $\mu^P(X, g_1)$ conditional on $G = g_1$ but not of $\mu^P(X, g_2)$ given $G = g_2$. In that case, for group 1, we will have equality at the margin but for group 2, the marginal admits will have expected outcome exceeding the threshold if γ lies in a "hole" with respect to the support of $\mu^P(X, g_2)$ given $G = g_2$. Figure 1 illustrates the point. The common support assumption would hold if a situation like the top panel holds where both curves



The solid curve represents a fictitious conditional density of μ(X,male) and the dashed curve the density of μ(X,female). In the top panel, they have the same support and the common treatment threshold gamma is shown by the vertical line. In the bottom panel, the common threshold lies in the "hole" of the support of μ(X,female). So there is no x in the support of X for females where μ(X,female) can equal the common threshold.

Figure 1:

have positive height at the cutoff-point γ, marked by the vertical line. To be clear, this common support assumption has nothing to do with the identification of group-specific thresholds, analyzed in the following section. Instead, the purpose of this assumption is that it enables us to interpret the inequality between group-specific thresholds as symptomatic of academically unfair admissions.

3 Identification of γ_g

3.1 Identification method

The basic identification idea is to use for each fixed g, the median restriction and the observed $\Pr(D=1|X=x,G=g)$ to identify the values of X defining the marginal admits, viz., those x for which $\Pr(D=1|X=x,G=g)=1/2$ and then average $\mu^P(x,g)$ – separately identified from admitted students in previous years – across these marginal admits to yield γ_g.

Our identification is facilitated by the following regularity condition:

Assumption 4 *For each value g in the support of G, the distribution of the random variable $\mu^P(X,G)$ conditional on $G=g$ has a strictly positive density (with respect to the Lebesgue measure) on an open interval around γ_g.*

The above assumption guarantees that there exists some $x \in \mathcal{X}_g$, such that $\Pr(D=1|X=x,G=g)=1/2$. It will hold as long as X has at least one continuously distributed component and $\mu^P(X,G)$ varies sufficiently with that component. We emphasize that a "large" support for X is not necessary here, because for generic budget constraints, γ_g will be located in the interior of the support of $\mu^P(X,g)$.

We formally state the identification statement through the following proposition. Its proof also illustrates the intuition and hence is included in the main text.

Proposition 1 *Suppose that assumptions 1, 2 and 4 hold. Then for each g, the threshold γ_g is point-identified.*

Proof: Note that if there exists an x such that $\Pr(D=1|X=x,G=g)=1/2$, then we must have that

$$\Pr\left[\mu^P(x,g)-\gamma_g \geq \varepsilon | X=x, G=g\right]=1/2,$$

implying that

$$\mu^P(x,g)-\gamma_g=0,$$

by (ii) and (iii) of Assumption 2. Therefore, by averaging over all such x, one obtains that

$$E_X\left[\mu^P(X,g) - \gamma_g \mid \Pr\left(D=1|X,G=g\right) = \frac{1}{2}, G=g\right] = 0. \tag{3}$$

The latter implies that γ_g can be identified via the equality

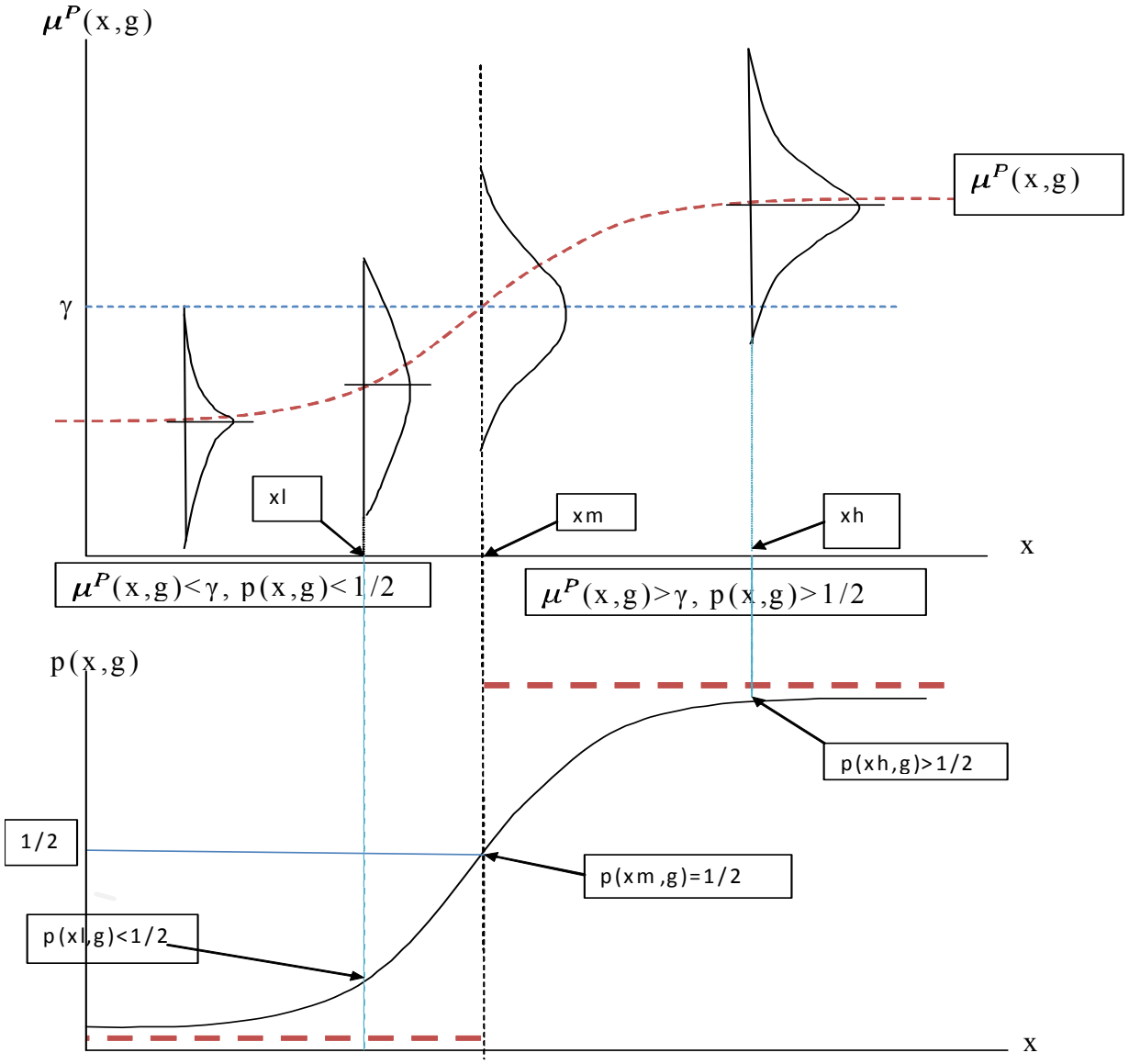
$$\begin{aligned}\gamma_g &= E_X\left[\mu^P(X,g) \mid \Pr\left(D=1|X,G=g\right) = \frac{1}{2}, G=g\right] \\ &\vdots \\ &= \frac{\int \mu^P(x,g) \times 1\left\{p(x,g) = \frac{1}{2}\right\} dF_{X|G}(x|g)}{\int 1\left\{p(x,g) = \frac{1}{2}\right\} dF_{X|G}(x|g)},\end{aligned}$$

where $p(x,g)$ denotes $\Pr\left(D=1|X=x,G=g\right)$. Now, assumption 4 guarantees that for every fixed g, the set $\Pi_g = \left\{x \in \mathcal{X}_g : \mu^P(x,g) = \gamma_g\right\}$ – identical to the observable set of $x \in \mathcal{X}_g$ satisfying $\Pr\left(D=1|X=x,G=g\right) = 1/2$ – is nonempty. Finally, $x \in \mathcal{X}_g$ guarantees that we can compute $\mu^P(x,g)$ for each $x \in \Pi_g$ from past cohorts, which completes the proof of identification. **Q.E.D.**.

Thus, operationally, the identification strategy for γ_g is to first detect current year's applicants of type (x,g) for whom the predicted probability (conditional on $G=g$) of getting an offer is exactly a half. These are the marginal candidates of type g whose X take values in the set Π_g. Then calculate predicted outcome, using data on past years' admits. Finally, average these predicted outcomes across last years' g-type admits with values of x in Π_g. This average yields γ_g.

Graphical Intuition: The above identification argument can be visualized through the graph depicted in figure 2.To interpret this graph, fix a value of G, say, $G=g$. Suppose X is one-dimensional and that $\mu^P(X,g)$ is continuously distributed. In the top panel of the graph, we plot $\mu^P(x,g)$ against x by the dashed curve and mark γ by the horizontal dashed curve. In the bottom panel, we plot the corresponding admission probability $p(X,g)$ against X in the absence of errors (dashed curve) and in the presence of errors (solid curve).

In the absence of errors, the admission probability would be zero for those values of x where $\mu^P(x,g) < \gamma$ and equal to one where $\mu^P(x,g) \geq \gamma$. Now consider what happens when there are stochastic perception errors. Such errors will make the perceived expectation at any value x of X to have a distribution around the dashed $\mu^P(x,g)$ curve. This is shown by the density humps in the graph's top panel which, given the zero median restriction, will be centered at the true $\mu^P(x,g)$. Now, whether a particular applicant with value x of X will be admitted is probabilistic, depending on whether the noisy subjective expectation exceeds γ. For a point like xh on the right, we have $\mu^P(xh,g) > \gamma$; consequently $p(xh,g)$ which equals the area under the density curve at xh above γ in the upper panel and marked by the vertical height of the solid curve in the lower panel will exceed one-half; similarly, the probability at a point like xl, $p(xl,g) < 1/2$ and only at a point like



Graphical Illustration of Identification for fixed G=g

Figure 2:

xm where $\mu^P(xm, g) = \gamma$, will the density hump be centred at γ, making the probability of success at exactly one-half. Notice that this argument does *not* require the density curves to be symmetric or have the same spread. All that is required is that the area under the curve for $X = x$ below $\mu^P(x, g)$ should be equal to that above $\mu^P(x, g)$, i.e., that the perception errors are equally likely to be positive and negative.

Once we have identified the group-specific thresholds γ_g, we can test if admission is outcome-oriented by testing the equality of γ_g across g. This implication is facilitated by our common support assumption (3) in the previous section for $\mu^P(\cdot, \cdot)$.

Remark 1 *It is useful to note that our method remains applicable in situations where universities get applications from students with different educational backgrounds. For example, among UK university applicants, quite a few take the International Baccalaureate instead of the A-level exams. Since our methodology is based entirely on the predicted outcomes and predicted probability of offer and not on the background covariates themselves, it is easy to include such students into the analysis. One simply uses IB scores instead of A-level scores as the corresponding X for these students and predicts outcomes and probability of offer from the corresponding regressions. Thereafter, all applicants are pooled together and the analysis proceeds exactly as before. In some real situations, one or more applicant characteristics may be more "qualitative" such as performance in admission interviews. However, for large applicant pools, such information is usually given a numerical score or grade by university officials to enable easy comparison in the final stage of selection. This score can be used as a component of X in our proposed methodology.*

Remark 2 *Our analysis does not require background information for past years' applicants who were rejected. Universities typically do not store this information and hence it is useful to have a method which does not require them.*

3.2 Median Independence in other treatment scenarios

Our assumption of median independence is intuitive in the admissions context of the present paper where an analyst can directly access the application forms and test scores. But it is worthwhile to check under what conditions it may hold in other contexts such as medical treatment where physicians are likely to observe many determinants of treatment outcome which are unobservable to an analyst. Toward that end, suppose the doctors and the analyst observe outcomes from randomized controlled trials where potential patient outcomes are denoted by the standard notation

Y_0 and Y_1. Let $Y = Y_1 - Y_0$ and suppose doctors make treatment decision for current patients, based on the scalar random variable $I := I(W, Z) := E(Y|X, Z)$ where W is observable to the analyst and Z is not. Suppose that a current patient with characteristics (w, z) is given the treatment if and only if $I(x, z) > \gamma$, i.e., the treatment rule is

$$D = 1\{I(X, Z) \geq \gamma\}.$$

Let $\mu(w)$ denote $E(I|W = w)$ which, by the law of iterated expectations, equals $E(Y|W = w)$, which is observable to the analyst. Rewriting the treatment rule in terms of observable and unobservable components, we get

$$\begin{aligned} D &= 1[\mu(W) + \{I - \mu(W)\} \geq \gamma] \\ &:= 1[\mu(W) + \varepsilon \geq \gamma]. \end{aligned}$$

This takes the form of (1) provided the restriction $median(\varepsilon|W) = 0$ holds almost surely. The latter will hold if and only if

$$med(I - \mu(W)|W) = 0, \text{ i.e., } E\{I|W\} = median\{I|W\}, \text{ a.s.} \tag{4}$$

Note that condition (4) makes no further assumption about the distribution of Z given W and allows for arbitrary heteroskedasticity of Z in W. For example, if $I = W'\beta + Z$, then $\varepsilon = Z - E(Z|W)$ and the zero conditional median is equivalent to requiring that $E(Z|W) = med(Z|W)$. This is far weaker than requiring that Z be independent of W and, in particular, allows the variance of the unobservable Z to vary by W. In particular, this condition will hold whenever Z has a symmetric distribution, in particular a Gaussian law, conditional on W. Note also that (4) is a statement about conditional means and medians which may or may not have "causal interpretations" of any kind.

3.3 Comparison with other identification strategies

In the empirical microeconomics literature, several alternative approaches have been proposed for distinguishing "taste-based discrimination" from "statistical discrimination" (c.f., Persico (2009) for a review). The issue is to test whether (disparities in) observed treatment rates across demographic groups can be justified as the consequence of treaters maximizing a specific "legitimate" objective, based on applicant characteristics which they observe. This is essentially identical to our

definition of academic fairness in the college admissions context[5] However, the existing approaches, which differ depending on the application context and data availability, are not applicable to our setting. For instance, in the context of law-enforcement, Knowles, Persico and Todd (2005)[6] attempt to identify racially motivated, inefficient treatment using the fact that criminals can alter their potential outcomes in response to the crackdown regime. Such responses are not feasible in the admissions context where applicant outcome depends on long-term human capital accumulation.

In the context of medical treatment, Chandra and Staiger (2009) attempt to identify difference in outcome thresholds for surgery by assuming an index-restriction on unobservables distributions. As explained above, this approach fails when the unobservable's distribution has covariate-dependent scale, as is quite likely when decision-makers have comparatively less experience with applicants from specific groups and thus make errors with larger variances for such groups. In the medical context, Bhattacharya (2010) suggests an alternative approach to testing outcome-oriented treatment assignment via a partial identification analysis using a combination of observational data and experimental findings. Such experimental results are typically difficult to come by in the college admissions context.

In the medical setting, Anwar and Fang (2011) consider a test of taste-based prejudice in emergency room discharge using the re-admission rate as the outcome of interest. The key assumption is that physicians have at their disposal a continuous choice variable related to diagnostic tests which they can choose optimally in order to determine suitability for discharge. The identification strategy is then based on comparing the re-admission rates of patients of different race who had undergone the diagnostic test at the physician-optimized level of intensity. In the admissions set-up, there is no such continuous choice variable available to admission officers.

In recent work, Bertrand et al (2010) have examined the consequences of affirmative action in admission to Indian engineering colleges on the graduates' earnings. In their context, admission is based on a single exam score and admission thresholds vary by applicants' social caste. These thresholds are publicly fixed and commonly known, thereby removing the key empirical challenge – that of defining and identifying the marginal admits and rejects – arising in general admissions

[5] We deliberately refrain from using the term discrimination because it is suggestive of prejudice. Neither we, nor the existing research in our reading, can pinpoint the bahavioral source of any observed discrepency from the economic ideal of identical marginals. As such, any such discrepency is stated to have arisen from "taste-based" motives, by definition.

[6] Related recent papers include Anwar and Fang (2006), Grogger and Ridgeway (2006), Antonovic and Knight (2009) and Brock et al (2011).

contexts where entrance is based on several background variables.[7]

In ongoing work, Jiang, Nelson and Vytlacil (2011) analyze the identification of a deterministic model of loan approval using information on approved loans alone. Their setting and their goal are different from those of the present paper. In particular, JNV wish to identify the analog of the $\mu^P(\cdot)$ function in the deterministic model $D = 1\left(\mu^P(W) > 0\right)$ but when they only observe the distribution of $W|D = 1$. In contrast, we observe W for all applicants, the relevant μ function is identified directly from past outcomes data, the determination of D involves additional heterogeneity and the goal is to identify the threshold γ's which potentially vary by W. Like us, JNV also assume, realistically, that all characteristics of loan-applicants that the banks observe and systematically use are available to the analyst via the application forms but, unlike us, they cannot allow for any unobserved heterogeneity in the approval equation, given their data limitations.

4 Estimation and Inference

We now consider the calculation of γ_g from admissions data collected for one or more cohorts of applicants. Note that any particular cohort may be viewed as a random sample from the overall population of all applicants. Therefore, the values of γ_g calculated based on a specific cohort will suffer from sampling uncertainty; so a test of equal γ_g's requires a distribution theory, which we derive in this section.

Motivated by the restriction of (3), we first present an estimator of γ_g. Observe that our identification strategy is fully nonparametric and does not require any functional form assumption. With a large enough sample size, one can consider fully nonparametric estimation of $\mu^P(x,g)$, $p(x,g)$ and, eventually, γ_g. But for our sample size, this is difficult to implement due to curse of dimensionality. We therefore resort to estimating $\mu^P(x,g)$ and $p(x,g)$ via parametric models here. For estimating γ_g we consider both parametric as well as non-parametric kernel based approaches; in our empirical application, we report the results from both approaches. In the appendix we

[7] Our use of the term "marginal" is different from the notion of marginal individuals in Carneiro, Heckman and Vytlacil (2009). Firstly, the set-up in their paper involves an instrumental variable, satisfying an exclusion restriction and with large support, which affects allocation to treatment. No such IV seems to be available in our context. Without such an IV, the analog of CHV's "marginal individuals" of type (x,g) in our set-up are those for whom the corresponding admission officer's unobservable error ε satisfies $\varepsilon = \mu^P(x,g) - \gamma_g$. But since we are primarily interested in identifying the university-wide baseline γ_g from knowledge of $\mu^P(x,g)$, such individuals are not of primary interest to us. Instead, the relevant g-type marginal individuals for us are those whose x satisfy $\mu^P(x,g) = \gamma_g$.

state and prove formal theorems describing the distribution theory for this semiparametric case, c.f., theorem 1 in subsection 8.3. For the sake of pedagogical completeness, in the last part of the appendix we state and prove the asymptotic distribution of $\hat{\gamma}_g$ resulting from a fully nonparametric estimation of $\mu^P(x,g)$, $p(x,g)$, c.f., theorem 2, subsection 8.3.

In the semiparametric approach, we estimate $\mu^P(x,g)$ and $p(x,g)$ parametrically in the first step and then in the second step, we estimate γ_g by a weighted average of $\hat{\mu}^P(X_i,G_i)$, where the weights are a decreasing function of the distance between $\hat{p}(X_i,G_i)$ and $1/2$,

$$\hat{\gamma}_g = \frac{\sum_{i=1}^{n} K_h\left(\hat{p}(X_i,G_i) - 1/2\right)\hat{\mu}^P(X_i,G_i)\,\mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(\hat{p}(X_l,G_l) - 1/2\right)\mathbf{1}\{G_l = g\}}.$$

Here $K_h(z) = K(z/h)/h$; $K(\cdot)$ is a kernel function; h is a smoothing parameter (bandwidth); $\hat{p}(x,g)$ and $\hat{\mu}^P(x,g)$ are first-step estimators of $p(x,g)$ $(:= \Pr[D_i = 1|(X_i,G_i) = (x,g)])$ and $\mu^P(x,g)$, respectively. This is exactly the average predicted outcome for those (x,g) types for whom the predicted probability of getting an offer, i.e., $p(x,g)$ is close to a half, where closeness is determined by the kernel and the bandwidth h.

We may contrast this with a benchmark, fully parametric approach, which is easier to implement and does not require subjective bandwidth choice. In this case, we estimate $\mu^P(x,g)$ and $p(x,g)$ parametrically in the first step and then in the second step, project the estimated $\mu^P(x,g)$ on (functions of) the estimated $p(x,g)$, using linear regression. Then γ_g is estimated by the predicted value of the final regression, evaluated at $\hat{p}(x,g) = 1/2$.

In order to increase our sample size, we will make the following stationarity assumption, which will let us combine the current year performance data (whenever available) with past data to calculate the estimates.

Assumption 5 *The past and current cohorts are drawn from the same population distribution of applicants and the admission thresholds have not changed over the entire period of analysis.*

In view of this stationarity assumption 5, we drop the P superscript from now on and combine all available years of data for estimating conditional means and probabilities in order to maximize precision.[8]

[8] Also, note that if the outcome is available for every individual i admitted in the curent year as well, then we may use the outcome Y_i itself instead of $\hat{\mu}^P(X_i,G_i)$. However, if such data are not available, we cannot average them and no one in the past data may have exactly the same value of x_i for applicant i in the current data (especially if X contains continuous components like test scores) whose Y we could use directly. This would force us to average

For the fully parametric case, due to the smoothness of the estimator of γ_g in the regression parameters, the estimator will be root-n normal and one can use the bootstrap to get its standard errors. The semiparametric case is somewhat different from standard 2-step estimators where the first step is nonparametric and the second step involves some form of averaging of the first step estimators, leading eventually to$\sqrt{n}$-normal estimates. Here, due to kernel smoothing at the second step, even if the first step is parametric, one cannot estimate γ_g at the parametric rate. We now outline the distribution theory in this case.

Distribution of the semiparametric estimator: For the first stage, one may use any parametric model satisfying some mild conditions (c.f., assumptions 11 and 12, below) e.g., a probit or logit model for $p(w)$; and a linear (regression) model for $\mu(w)$. Define $\tilde{\gamma}_g$ as the infeasible estimator that would result if the true values $\mu(X_i, g)$ and $p(X_i, g)$ were used instead of their estimates:

$$\tilde{\gamma}_g := \frac{\sum_{i=1}^{n} K_h\left(p(X_i, g) - 1/2\right) \mu(X_i, g) \mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(p(X_l, g) - 1/2\right) \mathbf{1}\{G_i = g\}}, \tag{5}$$

for each $g \in \{f, m\}$. We show in the appendix (see theorem 1) that our semiparametric estimator $\hat{\gamma}_{sp}(g)$ has the same asymptotic distribution as $\tilde{\gamma}_g$. Since $\tilde{\gamma}_g$ is a nonparametric regression of the dependent variable $\tilde{\gamma}_g$evaluated at $p(X_i, g) = 1/2$, its distribution follows from standard conditions and takes the form:

$$\sqrt{nh}\left[\tilde{\gamma}_g - \gamma_g - h^2 \mathbf{B}(g)\right] \xrightarrow{d} N\left(0, V(g)\right),$$

where $\mathbf{B}(g)$ and $V(g)$ denote bias and variance terms. Under appropriate undersmoothing – leading to the asymptotic disappearance of the bias – one can construct confidence intervals for γ_g. The forms of the bias and variance together with the sufficient technical conditions are formally stated as theorem 1 in the appendix.

Remark 3 *Note that the convergence rate of $\hat{\gamma}_g$ does not depend on the dimension of X since the asymptotic distribution of $\hat{\gamma}_g$ and the infeasible $\tilde{\gamma}_g$ are identical (shown in theorem 1 in the appendix). This result is generic and also obtains even when $p(x, g)$ and $\mu(x, g)$ are nonparametrically estimated. The latter is shown in theorem 2 in the appendix subsection 8.3.*

Choosing Bandwidths: Note that our parameter of interest, γ_g is essentially the conditional mean $E\left(\mu(X, G) | p(X, G) = 1/2, G = g\right)$. So we recommend the standard method based on cross-

Y's for the past years for individuals whose X values are close to x_i. In order to be robust to this situation, we use $\hat{\mu}(X_i, G_i)$ instead of Y_i.

validation, which uses a global goodness of fit criterion for the conditional mean $E(\mu(X,G)|p(X,G),G=g)$. In the present context, cross-validation is achieved by minimizing the leave-one-out criterion

$$\begin{aligned} R(h) &= \sum_{j=1}^{n} 1(G_j = g) \times [\hat{\mu}(X_j, g) - m_{-j}(\hat{p}(X_j, G_j), g); h]^2, \text{ where} \\ m_{-j}(a, g; h) &= \frac{\sum_{\substack{i=1 \\ i \neq j}}^{n} K_h(\hat{p}(X_i, G_i) - a)\hat{\mu}(X_i, G_i)\mathbf{1}\{G_i = g\}}{\sum_{\substack{i=1 \\ i \neq j}}^{n} K_h(\hat{p}(X_i, G_i) - a)\mathbf{1}\{G_i = g\}} \end{aligned}$$

is an estimator of $E(\mu(X,G)|p(X,G) = a, G = g)$, calculated using the bandwidth h. The minimizer $\hat{h}_{CV}$ of the CV criterion is optimal in that it converges to the minimizer of the true mean squared error of the estimator. However, if we let $h = \hat{h}_{CV}$, then we incur the asymptotic bias, since the order of $\hat{h}_{CV}$ is $n^{-1/5}$. To remove the bias, we use $h = \hat{h}_{CV}/(\ln n)$ in our implementation. This undersmoothing, as is well-known, serves to reduce the asymptotic bias and makes it possible to construct confidence intervals for γ_g without explicitly estimating the bias component.[9]

5 Productivity Loss from Unequal Thresholds

Many universities deliberately apply a lower threshold for applicants from historically disadvantaged backgrounds – such as women or racial minorities. From a policy perspective, it is important to cast the social benefits of such a policy against the short-term loss in productive efficiency – measured as the academic achievement foregone by the university – as a consequence of the policy. In this section of the paper, we outline a method of calculating such productivity shortfalls. Note that these calculations do not capture the full, general equilibrium consequences of equating thresholds across demographic groups, starting from a situation of unequal thresholds. The latter would require a more comprehensive modelling of the action of various types of colleges as well as application decisions by youths which would depend on, among other things, the financial returns to college education. This type of analysis is outside the scope of the present paper.[10] Instead, our motivation for studying the productivity effects is that it can be used as a metric of how far

[9] Note that the need for the undersmoothing is not a problem unique to our estimator, but is shared by any kernel-based estimators (see, e.g., pp. 41-43 in Pagan and Ullah, 1999). Alternatively, we might be able to estimate the bias component. However, it is not easy since $\mathbf{B}(g)$ involves derivatives of relevant functions, whose nonparametric estimation requires some other bandwidth choice.

[10] Arcidiacono (2005) conducts a structural, simulation-based analysis of the general equilibrium effects of changing finacial aid policies on black enrolment at elite US colleges and on black college attendance in general. However, he is not concerned with detecting unfair treatment or quantifying its effect.

the actual admission process deviates from the (expected) outcome-maximizing ideal. Indeed, for males and females, say, the difference $\gamma_m - \gamma_f$ is itself one such metric. However, such a difference does not capture the efficiency loss per se because the latter depends not just on the thresholds but also on the distribution of productivities $\mu(X, m)$ and $\mu(X, f)$ as well as the distributions of the officer errors ε.[11]

Toward that end, suppose that progressive motives lead a university to choose a lower threshold for female candidates, i.e., $\gamma_f < \gamma_m$. Then, the efficiency cost of this policy may be computed as the achievement foregone by replacing high ability males with low ability females relative to the efficient situation where a common threshold $\tilde{\gamma} \in (\gamma_f, \gamma_m)$ is used. This intermediate $\tilde{\gamma}$ can be chosen so that the aggregate number of admits, if the common $\tilde{\gamma}$ is used, will remain identical to current number of admits in expectation.

Under the new threshold, the model of admission is given by: admit a candidate of type $X = x$ and $G = g$ (where g is either male (m) or female (f)), according to the rule

$$D = \begin{cases} \mathbf{1}\{\mu(x, g) \geq \tilde{\gamma} + \varepsilon\}, \text{ if } x \in \mathcal{X}_g \\ 0, \text{ if } x \notin \mathcal{X}_g. \end{cases}.$$

Note that under this change in policy, the set $\mathcal{X}_g$ remains the same; so those characteristics for which the probability of admission was zero before the change will still lead to outright rejection. It's only individuals with characteristics in $\mathcal{X}_g$ who will now be treated differently, corresponding to the use of a new threshold $\tilde{\gamma}$. This restriction is needed because even under the new university-specified threshold, admission officers do not have any information on the outcomes of individuals whose X lie outside $\mathcal{X}_g$ (except perhaps that their conditional expectation are at most $\inf_{x \in \mathcal{X}_g} \mu(x, g)$). We assume therefore that under the new thresholds, they continue to use the set $\mathcal{X}_g$. The assumption does not seem overtly restrictive because individuals with X outside $\mathcal{X}_g$ are likely to be extremely weak and even under $\tilde{\gamma}$, their probability of getting an offer will be very small, so that they will contribute very little to the expected outcome anyway. We also maintain the stationarity assumption throughout this section, so that we will drop the P superscript everywhere and to fix ideas, we will use gender to be the covariate of interest.

Before turning to the actual derivation, it may be worth noting that we can, in general, only bound such efficiency costs. The intuitive reason for this is that we cannot exactly extrapolate probabilities of getting an offer under the counterfactual $\tilde{\gamma} \in (\gamma_f, \gamma_m)$ for male candidates with

[11] Fryer and Loury (1996), while simulating efficiency loss from "color-blind" admissions", measure it relative to the loss from ignoring standardized test-scores. This is of course another alternative.

high enough values of $\mu(x,m)$ and females with low enough values of $\mu(x,f)$. All we can say is that their probabilities under $\tilde{\gamma}$ will be higher than the highest and lower than the lowest probabilities respectively, under the current thresholds.

By way of notation, let π_m denote the fraction of applicants who are male. Given the new threshold $\tilde{\gamma}$, define the two sets

$$\begin{aligned} M &= \left\{ x \in \mathcal{X}_m : \mu(x,m) + \gamma_m - \tilde{\gamma} < \sup_{z \in \mathcal{X}_m} \mu(z,m) \right\}, \\ F &= \left\{ x \in \mathcal{X}_f : \mu(x,f) + \gamma_f - \tilde{\gamma} < \inf_{z \in \mathcal{X}_f} \mu(z,f) \right\}. \end{aligned}$$

These sets represent individuals for whom we can directly compute the probability of admission offer corresponding to the new threshold and these will be used in our calculations below.

We will now strengthen some of our previous assumptions, which is needed for calculating the change in expected average outcome resulting from the new threshold.

Assumption 6 *ε is independent of X, given G; i.e., for all g, the conditional c.d.f. of ε satisfies $F_{\varepsilon|X=x,G=g}(\cdot) = F_{\varepsilon|G=g}(\cdot)$.*

Assumption 7 *Conditional on each g in the support of G, the random variable $\mu(X,G)$ assumes every value within $\max_{z \in \mathcal{X}_g} \mu(z,g)$ and $\min_{z \in \mathcal{X}_g} \mu(z,g)$ with positive probability. That is, for any $a \in \left(\inf_{z \in \mathcal{X}_g} \mu(z,g), \sup_{z \in \mathcal{X}_g} \mu(z,g) \right)$, there exists $x \in \Omega_g$ such that $\mu(x,g) = a$.*

Assumption 8 *Acceptance of an offer and the outcome are independent of ε, given X and G, i.e., corresponding to any pair of offer thresholds $\gamma, \tilde{\gamma}$, we have*

$$\begin{aligned} & E(Y|A=1, \varepsilon \leq \mu(X,G) - \gamma, X=x, G=g) \\ = \; & E(Y|A=1, \varepsilon \leq \mu(X,G) - \tilde{\gamma}, X=x, G=g), \end{aligned}$$

and

$$\begin{aligned} & \Pr(A=1|\varepsilon \leq \mu(X,G) - \gamma, X=x, G=g) \\ = \; & \Pr(A=1|\varepsilon \leq \mu(X,G) - \tilde{\gamma}, X=x, G=g). \end{aligned}$$

We denote the probability of accepting an offer (conditional on having got an offer) by male and female applicants of type $X=x$ by $\alpha_m(x)$ and $\alpha_f(x)$, respectively.

Discussion: Assumption 6 strengthens the conditional median independence of assumption 2 (ii) to conditional distributional independence. Note, however, that this assumption still allows

the variance of ε to depend on G, which is a useful generalization, as noted above. Assumption 7 strengthens the support condition above to require that there are no "holes" in the (conditional on $G = g$) support of $\mu(X, g)$. These two assumptions enable us to compute the (counterfactual) probabilities $\Pr_{\varepsilon|X=x,G=g}\left[\tilde{\gamma} \leq \mu(x, m) - \varepsilon < \gamma_g\right]$ for $G = m$ and $G = f$, which are ingredients of the productivity change calculations below. Assumption 8 says that two candidates with identical values of X and G who are offered admission have the same expected outcome and the same probability of acceptance, irrespective of the offer generating mechanism. This assumption is implied by the stronger condition that ε is a purely officer-specific error and applicant files are randomly allocated among admission officers. Note that assumption 8 still allows acceptance to be correlated with (X, G) and with unobserved determinants of Y and thus includes cases where more able students (who also have higher test-scores on average) get outside offers and do not accept the university's offer or less able students cannot satisfy the conditions of the offer and drop out.

Bounds on productivity loss: For $x \in M$, define $\tilde{x}(x; \tilde{\gamma})$ as a solution to: $\mu(\tilde{x}, m) = \mu(x, m) + \gamma_m - \tilde{\gamma}$. For $x \in F$ define $\bar{x}(x; \tilde{\gamma})$ as a solution to $\mu(\bar{x}, f) = \mu(x, f) + \gamma_f - \tilde{\gamma}$. Since $\gamma_f \leq \tilde{\gamma} \leq \gamma_m$, by assumption 7 , such an $\tilde{x}(x, \tilde{\gamma})$ must necessarily exist in $\mathcal{X}_m$ and $\bar{x}(x; \tilde{\gamma})$ in $\mathcal{X}_f$. Under assumptions 6, 7 and 8, the productivity gain achieved by raising the threshold from γ_f to $\tilde{\gamma}$ for women and lowering it from γ_m to $\tilde{\gamma}$ for men is given by the difference $\omega_m(\tilde{\gamma}) - \omega_f(\tilde{\gamma})$, where the constituent terms can be bounded as follows. The derivations appear in the appendix.

- $\omega_m(\tilde{\gamma}) \in [\omega_{ml}(\tilde{\gamma}), \omega_{mu}(\tilde{\gamma})]$ where

$$\begin{aligned}
\omega_{mu}(\tilde{\gamma}) &= \pi_m \int_{x \in M} \mu_m(x)\, \alpha_m(x) \left\{p_m\left(\tilde{x}(x; \tilde{\gamma})\right) - p_m(x)\right\} dF(x|G = m) \\
&\quad + \pi_m \int_{x \in \mathcal{X}_m \cap M^c} \mu_m(x)\, \alpha_m(x) \left\{1 - p_m(x)\right\} dF(x|G = m), \\
\omega_{ml}(\tilde{\gamma}) &= \pi_m \int_{x \in M} \mu_m(x)\, \alpha_m(x) \left\{p_m\left(\tilde{x}(x; \tilde{\gamma})\right) - p_m(x)\right\} dF(x|G = m) \\
&\quad + \pi_m \int_{x \in \mathcal{X}_m \cap M^c} \mu_m(x)\, \alpha_m(x) \left\{\sup_{z \in \mathcal{X}_m} p_m(z) - p_m(x)\right\} dF(x|G = m). \quad (6)
\end{aligned}$$

- $\omega_f(\tilde{\gamma}) \in [\omega_{fl}(\tilde{\gamma}), \omega_{fu}(\tilde{\gamma})]$, where

$$\begin{aligned}
\omega_{fl}(\tilde{\gamma}) &= (1-\pi_m)\int_{x\in F}\mu_f(x)\,\alpha_f(x)\,\{p_f(x)-p_f(\bar{x}(x,\tilde{\gamma}))\}\,dF(x|G=f) \\
&\quad +(1-\pi_m)\int_{x\in\mathcal{X}_f\cap F^c}\mu_f(x)\,\alpha_f(x)\left\{p_f(x)-\inf_{z\in\mathcal{X}_f}p_f(z)\right\}dF(x|G=f), \\
\omega_{fu}(\tilde{\gamma}) &= (1-\pi_m)\int_{x\in F}\mu_f(x)\,\alpha_f(x)\,\{p_f(x)-p_f(\bar{x}(x,\tilde{\gamma}))\}\,dF(x|G=f) \\
&\quad +(1-\pi_m)\int_{x\in\mathcal{X}_f\cap F^c}\mu_f(x)\,\alpha_f(x)\,p_f(x)\,dF(x|G=f). \qquad (7)
\end{aligned}$$

- If

$$\inf_{z\in\mathcal{X}_f}p_f(z)=0 \text{ and } \sup_{z\in\mathcal{X}_m}p_m(z)=1, \qquad (8)$$

then the bounds shrink to singletons

$$\begin{aligned}
\omega_m(\tilde{\gamma}) &= \omega_{mu}(\tilde{\gamma})=\omega_{mu}(\tilde{\gamma}) \\
\omega_f(\tilde{\gamma}) &= \omega_{fl}(\tilde{\gamma})=\omega_{fu}(\tilde{\gamma}).
\end{aligned}$$

We summarize the above results into a proposition.

Proposition 2 *Under assumptions 6, 7 and 8, the change in expected overall outcome resulting from the use of a common threshold $\tilde{\gamma}$ is given by the difference $\omega_m(\tilde{\gamma})-\omega_f(\tilde{\gamma})$, where bounds for $\omega_m(\tilde{\gamma})$ and $\omega_f(\tilde{\gamma})$ are provided in (6) and (7) respectively.*

Proof: See Appendix

An interesting special case is where $\tilde{\gamma}$ is chosen such that the net enrolment in expectation remains the same as before, i.e., $\tilde{\gamma}$ satisfies with $p_m(x;\tilde{\gamma})$ and $p_f(x;\tilde{\gamma})$ denoting the probabilities of offer under the new threshold $\tilde{\gamma}$:

$$\begin{aligned}
&\pi_m\int_{x\in\mathcal{X}_m}p_m(x;\tilde{\gamma})\,\alpha_m(x)\,dF_{X|G=m}(x)+(1-\pi_m)\int_{x\in\mathcal{X}_f}p_f(x;\tilde{\gamma})\,\alpha_f(x)\,dF_{X|G=f}(x) \\
= \;&\pi_m\int_{x\in\mathcal{X}_m}p_m(x)\,\alpha_m(x)\,dF_{X|G=m}(x)+(1-\pi_m)\int_{x\in\mathcal{X}_f}p_f(x)\,\alpha_f(x)\,dF_{X|G=f}(x). \qquad (9)
\end{aligned}$$

The probabilities on the LHS of the above equation can be calculated in the same way as above. In particular, assuming $\inf_{z\in\mathcal{X}_f}p_f(z)=0$ and $\sup_{z\in\mathcal{X}_m}p_m(z)=1$ (as in our application), the LHS equals

$$\begin{aligned}
&\pi_m\int_{x\in M}\alpha_m(x)\,p_m(\tilde{x}(x;\tilde{\gamma}))\,dF(x|G=m) \\
&+\pi_m\int_{x\in\mathcal{X}_m\cap M^c}\alpha_m(x)\,dF(x|G=m) \\
&+(1-\pi_m)\int_{x\in F}\alpha_f(x)\,p_f(\bar{x}(x;\tilde{\gamma}))\,dF(x|G=f).
\end{aligned}$$

Plugging into (9), one can solve for the budget neutral $\tilde{\gamma}$.

Developing the distribution theory for this functional is complicated. The complication arises from the possible non-uniqueness of $\tilde{x}(x;\tilde{\gamma})$ and $\bar{x}(x;\tilde{\gamma})$ and the need to estimate the sets M and F which involve extrema of estimated functions. In such cases, naive bootstrap and/or subsampling are now well-known to provide inconsistent inference in a uniform sense. So we report the point estimates in the application.

6 Application to Oxford Admissions

Background: Our application is based on admissions data from a large undergraduate programme at Oxford University. The data pertain to three recent cohorts of applicants. We focus on UK-based applicants who have (i) written a substantive essay (a requirement for entry), (ii) had taken a standardized aptitude test (comparable to the SAT for US colleges), (iii) had taken the standardized school-leaving examination in the UK, viz., the GCSE, and (iv) have either taken or will take the advanced school qualifications – A-levels – before college begins. Almost all UK-based applicants would normally satisfy these four criteria.

The application process consists of an initial stage whereby a standardized "UCAS" form is filled by the applicant and submitted to the university. The form records the applicant's name, gender, school type, prior academic performance, a personal statement and a letter of reference from the school. The aptitude-test and essay-assessment scores are separately entered into the central database.

About one-third of all applicants are selected for interview on the basis of UCAS information, aptitude test and essay, and the rest rejected. Selected candidates are then assessed via a face-to-face interview and the interview scores are recorded in the central database. This sub-group of applicants who have been called to interview will constitute our sample of interest. Therefore, we are in effect testing the academic fairness of the second round of the selection process, taking the first round as given. Accordingly, from now on, we will refer to those summoned for interview as the applicants. The final admission decision is made by considering all the above information from among the candidates called for interviews. Whenever a student has not yet taken the A-level exams, the schools' prediction of their A-level performance is taken into account. In such cases, admission offers are made conditional on the applicant securing the predicted grades. For the application, we use anonymized data for three cohorts of applicants from their records held

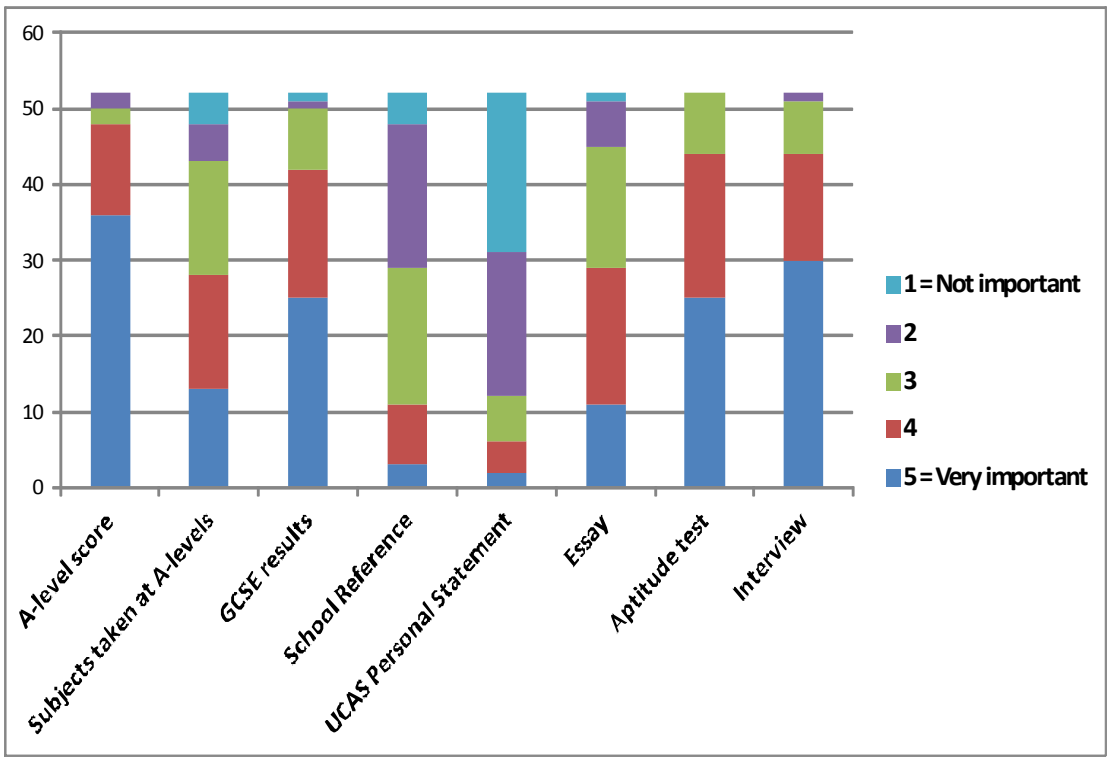


Figure 3:

at the central admissions database at Oxford. For the admitted students, we merged these with their performance in the first year examinations, in which students take three papers. The scores across the three papers are averaged to calculate the overall performance, which we take to be the outcome of interest.

In table 0, we provide explanation of the labels used in the subsequent tables.

Choice of covariates: We chose a preliminary set of potential covariates, based mainly on intuition and personal anecdotal experiences with colleagues. To confirm our choice, we conducted an anonymized online survey of the subject-tutors in Oxford, who participate in the admission process. The survey asked the tutors to state how much weight they attach during admissions to each of these potential covariates with "1" representing no weight and "5" denoting maximum weight. The results, based on 52 responses, are summarized in figure 4. One may count the fraction of "important (score=4)" and "very important (score=5)" for each category (equivalently the sum of heights of the bottom two sections of the bars in figure 4) to gauge its perceived importance in the admissions process. The A-levels appear to be the most important criterion, followed by the aptitude test and interview scores and then GCSE performance. The choice of subjects at

A-level (two specific subjects , referred to as subjects 1 and 2, are recommended by Oxford for this particular programme of study) are given medium weights and the personal statement and school reference are given fairly low weights. We therefore settle on using scores from the GCSE, A-levels, aptitude test scores and the interview for our analysis. We also use dummies for whether the applicant studied the two recommended subjects at A-level.

Group Identities G: We consider fairness of admissions with regards to two different group identities, viz., gender and type of school attended by the applicant. Oxford University is frequently criticized for the relatively high proportion of privately-educated students admitted overall (c.f., footnote 1 above). The implication is that applicants from independent (private) schools, where spending per student is very much higher than in state schools (Graddy and Stevens, 2005), have an unfair advantage in the admissions process. As regards gender, in the UK, as in most OECD countries, the higher education participation rate is higher for women, having overtaken the participation rate for men in 1993. However, Oxford University appears to have lagged behind the trend: in 2010/11, 55% of undergraduates in UK universities were female, but 56% of students admitted to Oxford were male.[12] Typically, gender imbalances are more pronounced in certain programmes and includes the one we study, where male enrolment is nearly twice the female enrolment.

Given our focus on these group-identities, we separately asked tutors in our survey whether they took into account gender and school-type of the applicants in making their decision. This question is more politically sensitive than the previous ones and an affirmative answer is likely more trustable than a negative one. The responses are plotted in figure 5 where we see that tutors claim to use both characteristics in making their decision and school-type is paid more attention in general than applicant gender. Given these findings, we include school-type as an explanatory variable when calculating thresholds by gender and vice versa.

Outcome: After entering university, the candidates take examinations at the end of their first year. There are three papers, and each script is marked blindly, i.e., the marking tutors do not know anything about the candidate's background. We use the average score over the three papers as our outcome – labelled prelim_tot – which can range from 0 to 100. Obviously, this variable is available for admitted candidates only. There are two advantages of using the preliminary year score as the relevant outcome measure. One, every admit sits the same preliminary exam in any given year, so that there is no confounding from the difference in score distributions across different optional

[12] See, for instance, a recent Guardian newspaper report at:
http://www.guardian.co.uk/education/2009/aug/19/oxford-university-men-places-women

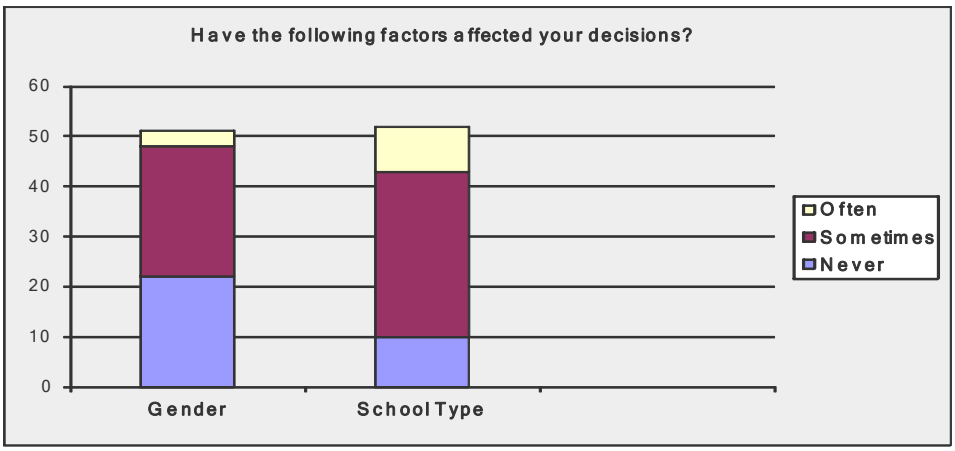


Figure 4:

subjects, as often happens in the final examinations at the end of the 3-year course. Secondly, the first year is closer to the time of entry; between the first and the third year, students have widely different academic experiences which likely play a significant role in determining their performance in the final exam. In other words, pre-admission characteristics, on which admissions are based, are likely to play a comparatively larger role in first year performance.

Summary statistics and success rates: We provide summary statistics for the data in Tables 1A and 1B.We first focus on differences in admission patterns by gender. Table 1A shows that male applicants have better aptitude test scores and interview averages and male admits score an average of about 2 points (out of 100) higher in the first year exams. They perform worse on average in their GCSE and A-levels. These differences are statistically significant at 5%. Note that there is no significant difference in offer rates between male and female candidates.

In table 2 we report the results of (i) a probit regression of receiving an offer as a function of various characteristics among all applicants and (ii) a linear regression of first year average outcome among the admitted candidates, as a function of the same characteristics. Table 2A strengthens the findings from table 1A and 1B by showing that even after controlling for covariates, gender and school-type do not affect the *average* success rate among applicants. On a more minor note, table 2A and 2B further show that the aptitude test and interview scores have the largest impact upon receiving an offer for the applicant population and a relatively smaller impact on first year performance among the admitted candidates. But since the underlying samples are different, the effects are not directly comparable. It is conceivable that among the sample selected to receive an admission offer, those with lower aptitude-test score are better along other dimensions than those

with low aptitude test-scores among the general applicant pool. This would serve to mitigate the effect of the aptitude test scores on first year performance among the admitted students relative to their impact on the potential outcomes of all applicants.

Threshold results: We now turn to the key results from applying the ideas of the present paper –viz., a test of whether the marginal admitted male and the marginal admitted female student have identical expected first year scores. To do this test, for each gender, we compute the expected score as a linear function of age, GCSE score, A-level scores, dummies for whether the candidate took the recommended subjects at A-level, aptitude-test scores, the interview score and whether the applicant came from an independent school. Using the zero conditional median restriction on errors, as explained above, we calculate the threshold faced by each gender as the average of expected first year scores for admitted applicants whose probability of being admitted is predicted (through a probit) to be close to 0.5. To choose the bandwidth for defining "closeness", we use the leave-one-out cross-validation. The CV criterion is plotted in figure 5 for the four cases of (clockwise from top left) male, female, state-school and independent-school. The numerical minimum of this criterion over a grid of 0.01 to 0.1 is taken to be the optimal bandwidth.

In table 3A, we show the difference in estimated admission thresholds for a range of bandwidths (which define "closeness to 0.5") and an Epanechnikov kernel $K(u) = \frac{3}{4}\left(1-u^2\right) \times 1(|u| \leq 1)$. The middle bandwidth(shaded row) is the optimal one, according to the cross validation criterion calculated above divided by log of the number of g-type applicants where g is either male or female. The other rows correspond to bandwidths that are 0.5 times the optimal one and 2 times the optimal one respectively. The very last row corresponds to a fully parametric analysis where the parametrically estimated $\hat{\mu}(x,g)$ is regressed on the parametrically estimated $\hat{p}(x,g)$ and its square and the predicted value at $\hat{p} = 1/2$ is taken to be the estimate of $\hat{\gamma}$. The second row in table 3A, for instance, may be read as follows. The first column specifies the scale by which the optimal bandwidth is multiplied (in this case 1), the second column reports the male threshold: we see that the marginal male admits are expected to score 60.9 percent in their first year examination. The third column shows that the marginal female admits can be expected to score 57 percent, implying a difference of 3.9 percent (reported in the 4th column). This difference has a 1-sided p-value of 0.004 under the null of equal thresholds, reported in the 5th column. Finally, the 6th column reports that if we were to use a common budget neutral threshold, the first year exam score of the resulting admitted cohort would be about 1.8 percentage points higher per applicant. Since the overall applicant success rate is 0.36, this amounts to 1.8/0.36=5 percentage points, i.e., about 1

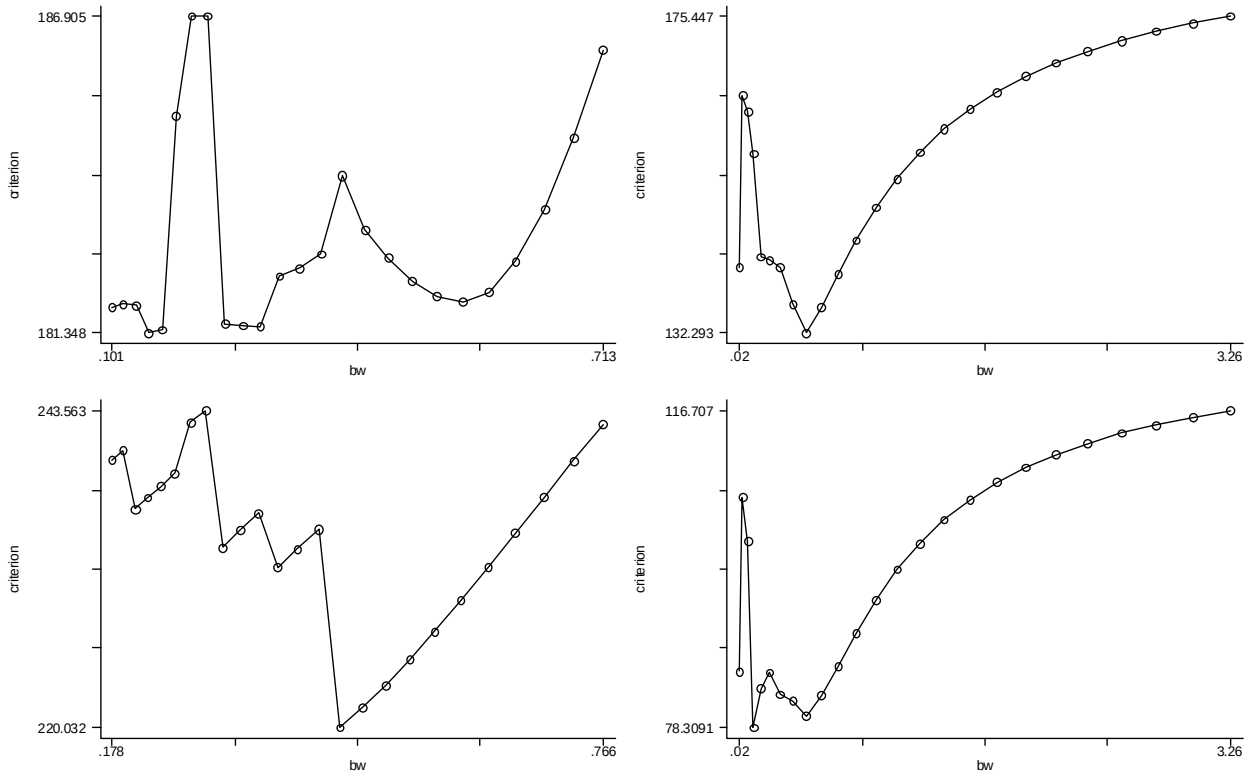


Figure 5:

standard deviation of the outcome among the admits.

From table 3A, it becomes clear that male candidates have to cross about a 4 percentage points higher barrier of expected performance than females in order to get admitted. It is interesting to contrast this finding with table 1A where we found that application success rates were almost identical across gender and table 2A where we found that gender was not a significant predictor of the *average* application success, conditional on other covariates. This highlights the usefulness of our approach which, by focusing on the *marginal* admits, reveals a stark difference between the treatments of male and female candidates not apparent from the conditional or the unconditional (on covariates) *average* success rates by gender. It is also interesting to note that the gender-difference in expected outcomes for the average admit is about 0.92 percentage points which much smaller than the 3.9 points difference among the marginal candidates.

Outcome variants and productivity loss: In table 4A, we consider slightly different forms of the outcome, viz., (i) the chances of scoring 70 or above, (ii) securing at least 60 and (iii) securing

at least 55. These correspond roughly to the 95th, the 50th and the 20th percentile of the overall score distribution. In particular, the 55+ criterion corresponds to an admission process designed to maximize the probability of securing a minimum benchmark. As such, it can be interpreted as the university acting in a risk-averse way. In all of these cases, estimates of the male threshold are significantly higher, confirming the previous findings. The difference is marginally significant for the outcome of 60+.

Results for school-type: Finally, we repeat the analysis reversing the roles of gender and school background, i.e., we use gender as an explanatory variable and test if applicants from independent school face a higher threshold than their counterparts who apply from state-funded schools. The results are reported in the lower panels (marked B) of tables 3 and 4. Now, we see a difference of about 1.7 percentage points for the average first year score suggesting that students from independent schools are held to a higher threshold of expected first year performance. The magnitude of difference and corresponding productivity loss are less than half the corresponding numbers for gender. In addition, table 4B reveals that for certain variants of the outcome, estimated thresholds are slightly higher for state-school applicants; however, these differences are statistically insignificant.

In order to gain some visual insight into how the threshold discrepancies arise, in figure 6, we plot the empirical marginal c.d.f. of the estimated $\mu(X, male)$ and $\mu(X, female)$ (left panel) and those of the estimated $\mu(X, indep\_school)$ and $\mu(X, state\_school)$ (right panel). It is clear that the male distribution first-order stochastic dominates the female distribution and the independent school distribution nearly first-order stochastic dominates the state-school distribution. This means that even if admissions are centrally conducted and are deterministic conditional on μ (i.e., there is no unobserved heterogeneity across admission officers), *any* common acceptance rate across gender will result in a higher μ for the marginal accepted male than the marginal accepted female. This can be see in figure 6, by looking along any fixed cutoff on the vertical axis. Any such horizontal cut-off line will intersect the female c.d.f. at a point that will lie strictly to the left of the point of intersection with the male c.d.f. We conjecture that the presence of unobserved heterogeneity across admission officers does not alter this fundamental dominance situation and produces the results reported above. A similar, albeit relatively weaker, dominance situation occurs for school-type, as can be seen in the right-hand graph in figure 6.

Interpretation of the Empirical Findings: It would be natural to conjecture that the observed threshold-differences arise primarily from the implicit or explicit practice of affirmative

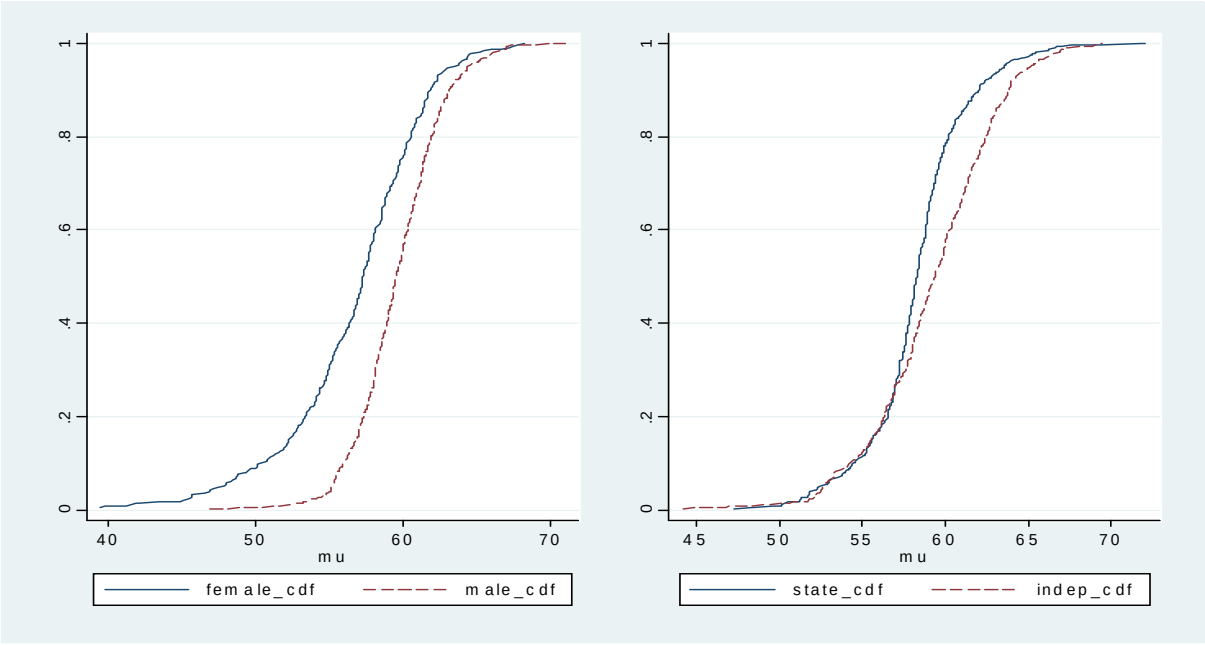


Figure 6:

action, viz., the overweighting of outcomes for historically disadvantaged groups. A second possibility is that, in face of political and/or media pressure, admissions tutors try to equate application success rate for, say, male with female applicants, which is also consistent with our empirical findings (see tables 1A and 1B and the last paragraph of the previous section). This would make the effective male threshold higher if, say, the conditional male outcome distribution has a thicker right tail (see figure 6) and officer perception errors are identically distributed. Regardless of what the underlying determinants of the officers' behavior are, we can conclude from our analysis that the admission practice under study deviates from the outcome-oriented benchmark and makes male or independent school applicants face effectively higher admission thresholds.[13]

[13]This conclusion is subject to the obvious caveat that if we use a different outcome, such as performance on the final examinations, the conclusion may be quite different. In that sense, our analysis is consistent with all the papers in the empirical microeconomics literature, including those cited above, which test "taste-based" treatment motives by focusing on a single outcome. It would be interesting to repeat our empirical analysis with finals performance data; however, data on final year scores are unfortunately not currently available for the relevant years, as of date. Furthermore, as discussed above, the preliminary year examination papers are identical across candidates, unlike finals where different students write exams in different subjects, depending on which areas they chose to specialize in.

7 Summary and Conclusion

This paper has proposed a general empirical methodology for testing whether an existing treatment protocol is economically fair in the sense of equalizing expected returns from treating marginal candidates across demographic groups, and calculating the productivity loss in situations where it is not. The focus is on the specific context of admissions to selective universities where allegations of unfairness are frequently made. Specifically, we consider the situation where a university bases admissions on the applicants' background data obtained through application forms and on standardized-test and/or interview-performance. We assume that a researcher can access this background information by acquiring the application form and the performance scores. Such admission procedures and data situations are extremely common across universities in the world, making our methodology fairly generally applicable. Furthermore, academic researchers can easily obtain these information, possibly in anonymized form, from their own institutions.

Applying our methods to admissions data for a large undergraduate programme of study at Oxford University and focusing on first-year examination performance as the outcome of interest, we found that the admission threshold faced by applicants who are male or from independent schools are systematically higher than those for female or state-school applicants. This contrasts sharply with the average admission rates, which are seen to be identical across gender and across school-type, whether or not we control for other covariates. This finding highlights the usefulness of our approach which, by focusing on the expected outcome of the marginal admits, rather than the aggregate admissions rate, reveals how applicants of different types face effectively different admission standards.

Our paper has left several substantive issues to future research. One, we do not consider peer-effects in our analysis; so we ignore scenarios where a student with relatively weaker predicted performance can, nonetheless, create positive externalities for other students and may therefore be preferred over someone with higher predicted individual performance but a negative externality on peers. However, in real settings, it is a bit unclear if admission tutors have enough information regarding peer effects to base their admission decisions on it.[14] Secondly, we do not consider a formal analysis of risk-aversion for the university and only provide a brief illustration in the empirical section. Indeed, for *binary outcomes*, like those reported in table 4, risk cannot play a

[14] In the somewhat different but related context of room-mate assignment policies that explicitly take into account peer effects, see recent papers by Bhattacharya (2009) and Carrell, Sacerdote and West (2011).

separate role and we see qualitatively similar results to those obtained when using the continuous outcome.[15] Nonetheless, for the sake of completeness and for use in other applications, this may be a direction worth further explorations in future.[16] Third, it may be useful to perform the empirical analysis using other types of outcome measures – such as wage upon graduation – as and when such data are available. However, we suspect that college performance data are much more readily available in general than wage data because the latter requires costly follow-up of alumni and can entail non-ignorable non-response. Finally, in ongoing work, we are (i) investigating the related but reverse question of how individual characteristics should be weighed in admission decisions and (ii) expanding on the idea outlined in section 3.2 regarding how median independence and/or symmetry conditions can be used to detect inefficient treatment allocation in medical-type settings.

[15] The literature on outcome-based analysis of fair treatment, cited above, either considers binary outcomes or assumes risk-neutrality when outcomes are continuous.

[16] For example, one can consider a family of utility functions for the university, indexed by a risk-aversion parameter, and ask what range of values of this parameter would rationalize the observed admissions data as the consequence of average utility maximization. This would naturally lend itself to a partial identification analysis.

Table 0: Variable labels

Label	Explanation
GCSE-score	Overall score in GCSE, 0-4
Alevel-score	Average A-level scores 80-120
Subject 1	Whether studied 1st recommended subject at A-level
Subject 2	Whether studied 2nd recommended subject at A-level
Aptitude-test	Overall score in Aptitude Test 0-100
Essay	Score on Substantive Essay 0-100
Interview	Performance score in interview 0-100
Prelim_tot	Average score in first year university exam; 0-100
Offer	Whether offered admission
Accept	Whether accepted admission offer

Note: The alevelscore is an average of the A-levels achieved by or predicted for the candidate by his/her school, excluding general studies.Scores are calculated on the scale A=120, A/B = 113, B/A = 107, B = 100, C = 80, D = 60, E = 40, as per England-wide UCAS norm.

Note: gcsescore is an average of the GCSE grades achieved by the candidate for eight subjects, where A* = 4, A = 3, B = 2, C = 1, D or below =0. The grades used are mathematics plus the other seven best grades.

Note: Oxford recommends that candidates study two specific subjects at A-levels for entry into the undergraduate programme under study. Subject 1 and Subject 2 are dummies for whether an applicant did study them at A-level.

1A. Summary stats by Gender

Variable	Obs	Mean	Obs	Mean	Difference	p-value
	Female		**Male**			
gcsescore	499	3.83	980	3.75	0.08	0
took subject 1	499	0.69	980	0.68	0.01	0.54
took subject 2	499	0.48	980	0.52	0.04	0.27
alevelscore	499	119.73	980	119.44	0.29	0.01
aptitude test	499	62.53	980	65.24	-2.71	0
essay	499	63.23	980	64.49	-1.26	0
interview	499	64.68	980	65.29	-0.61	0.04
prelim_tot	165	60.98	306	61.89	-0.92	0.04
offer	499	0.363	980	0.357	0.01	0.41
accept	499	0.34	980	0.34	0.00	0.5

1B. Summary stats by School-Type

Variable	Obs	Mean	Obs	Mean	Difference	p_value
	State		**Indep**			
gcsescore	818	3.70	661	3.87	-0.17	0
took subject 1	818	119.43	661	119.68	-0.24	0.02
took subject 2	818	0.65	661	0.73	-0.08	0.004
alevelscore	818	0.53	661	0.46	0.07	0.02
aptitude test	818	63.82	661	64.94	-1.12	0.0015
essay	818	64.06	661	64.07	-0.01	0.5
interview	818	65.02	661	65.17	-0.15	0.65
prelim_tot	260	61.15	211	62.10	-0.95	0.03
offer	818	0.361	661	0.357	0.00	0.5
accept	818	0.33	661	0.35	-0.01	0.46

2A. Probit of receiving offer

	Coef.	Std. Err.	z	pvalue
gcsescore	0.26	0.25	1.04	0.30
alevelscore	0.08	0.06	1.26	0.21
took subject 1	-0.06	0.17	-0.33	0.74
took subject 2	-0.25	0.15	-1.65	0.10
aptitude test	0.09	0.01	7.01	0.00
essay	0.01	0.01	0.44	0.66
interview	0.23	0.02	10.59	0.00
indep	-0.13	0.15	-0.88	0.38
male	-0.18	0.16	-1.13	0.26
_cons	-31.50	7.76	-4.06	0.00

N=1479, Pseudo-R-squared=0.5

2B. Regression of prelim_score

	Coefficient	Std. Err.	t	pvalue
gcsescore	4.19	2.42	1.73	0.09
alevelscore	0.79	0.40	1.96	0.05
took subject 1	0.24	1.11	0.22	0.83
took subject 2	-1.25	0.86	-1.45	0.15
aptitude test	0.28	0.07	4.15	0.00
essay	-0.02	0.07	-0.30	0.76
interview	0.17	0.10	1.76	0.08
indep	-0.01	0.92	-0.01	0.99
male	1.56	0.89	1.75	0.08
_cons	-79.84	49.54	-1.61	0.11

N=1479, R-squared=0.26

3A. Thresholds by Gender

Outcome mean=61.54, std dev=5.2

Method	Male-thld	Fem-thld	Male-Fem	p-value	Loss
Scale=0.5	60.92	56.44	4.48	0	1.78
Scale=1.00	60.9	57	3.9	0.004	1.86
Scale=2	60.88	56.86	4.02	0.01	1.84
Parametric	60.51	57.86	2.65	0.04	1.36

3B. Thresholds by School-type

Outcome mean=61.54, std dev=5.2

Method	Indep thld	State thld	Ind-State	p-value	Loss
Scale=0.5	60.67	59.02	1.65	0.16	1.15
Scale=1.00	60.74	59.11	1.63	0.05	1.02
Scale=2.00	60.62	58.96	1.66	0.04	0.9
Parametric	60.34	58.7	1.64	0.05	1.04

4A. Other outcomes and Productivity Loss by Gender

Outcome	Male-thld	Fem-thld	Male-Fem	p-value	Loss
70+ (mean 0.06)	0.07	0.013	0.057	0.02	0.012
60+ (mean 0.52)	0.44	0.37	0.07	0.15	0.015
55+ (mean 0.78)	0.8	0.63	0.17	0.05	0.05
Avg (mean 61.54)	60.9	57	3.9	0	1.86

4B. Other outcomes and Productivity Loss by School-type

Outcome	Indep-thld	State-thld	Indep-State	p-value	Loss
70+ (mean 0.06)	0.007	0.023	-0.016	0.34	0.07
60+ (mean 0.52)	0.421	0.416	0.005	0.48	0.02
55+ (mean 0.78)	0.75	0.79	-0.04	0.65	0.008
Avg (mean 61.54)	60.74	59.11	1.63	0.05	1.02

References

[1] Antonovics, Kate L. and Brian G. Knight, (2009): "A New Look at Racial Profiling: Evidence from the Boston Police Department," Review of Economics and Statistics, 2009, 91, 163–177.

[2] Anwar, Shamena and Hanming Fang (2006): "An Alternative Test of Racial Profiling in Motor Vehicle Searches: Theory and Evidence," American Economic Review, 96, 127–151.

[3] Anwar, S and H. Fang (2011): Testing for the role of prejudice in emergency departments using bounceback rates, NBER WP Number 16888.

[4] Arcidiacono, P (2005): "Affirmative Action in Higher Education: How do Admission and Financial Aid Rules Affect Future Earnings?" Econometrica, Vol. 73, No. 5, 1477-1524

[5] Becker, Gary (1957). The economics of discrimination, University of Chicago Press.

[6] Bertrand, Marianne and Sendhil Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," American Economic Review, 94 (4), 991–1013.

[7] Bertrand, M., R. Hanna and S. Mullainathan (2010): Affirmative action in education: Evidence from engineering college admissions in India, Journal of Public Economics, v. 94, iss. 1-2, pp. 16-29.

[8] Bhattacharya, D. (2009). Inferring Optimal Peer Assignment from Experimental Data. Journal of the American Statistical Association, Jun 2009, Vol. 104, No. 486: pages, 486–500.

[9] Bhattacharya, D. & Dupas, P. (2010). Inferring Efficient Treatment Assignment under Budget Constraints", forthcoming, Journal of Econometrics, also, NBER working paper number 14447.

[10] Bhattacharya, D. (2011): Evaluating Treatment Protocols Using Data Combination, mimeo. University of Oxford.

[11] Bouezmarni, T. & O. Scaillet (2005) Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data, *Econometric Theory* 21, 390-412.

[12] Brock, William A., J Cooley, S. Durlauf and S. Navarro (2011): "On the Observational Implications of Taste-Based Discrimination in Racial Profiling,", forthcoming, Journal of Econometrics.

[13] Carneiro, P., James J. Heckman and Edward J. Vytlacil (2011): Evaluating marginal policy changes and the average effect treatment for individuals at the margin, NBER working paper, w15211.

[14] Carrell, S., Bruce I. Sacerdote & James E. West, 2011. "From Natural Variation to Optimal Policy? The Lucas Critique Meets Peer Effects," NBER Working Papers 16865, National Bureau of Economic Research, Inc.

[15] Chandra, A. & D. Staiger (2009): Identifying provider prejudice in medical care, mimeo., Harvard University.

[16] Fryer Jr., Roland G., and Glenn C. Loury 2005. "Affirmative Action and Its Mythology." Journal of Economic Perspectives, 19(3): 147–162.

[17] Fryer, Roland G., Glenn C. Loury and Tolga Yuret (1996): "Color-Blind Affirmative Action," NBER WP. No. 10103.

[18] Gospodinov, N. & M. Hirukawa (2010) Nonparametric estimation of scalar diffusion processes of interest rates using asymmetric kernels, Working Paper 08-011, Department of Economics, Concordia University.

[19] Graddy, K. and M. Stevens (2005): "The Impact of School Inputs on Student Performance: An Empirical Study of Private Schools in the United Kingdom", Industrial and Labor Relations Review, 58(3), pp.435-451.

[20] Grogger, Jeffrey and Greg Ridgeway, "Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness," Journal of the American Statistical Association, 2006, 101, 878–887.

[21] Heckman, J. (1998). Detecting discrimination, Journal of Economic Perspectives-Volume 12, Number 2, 101-116.

[22] Jiang, W., R. Nelson and E. Vytlacil (2011): "Nonparametric Identification and Estimation of a Binary Choice Model of Loan Approval Using Only Approved Loans," ongoing work.

[23] Knowles, J., Persico, N. and Todd, P. (2001) "Racial bias in motor vehicle searches: theory and evidence", Journal of Political Economy, 109, (1), 203-232.

[24] Kobrin et al (2008). Validity of the SAT for Predicting First-year College Grade Point Average, College Board, New York.

[25] Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. Psychological Bulletin, 127, 162-181.

[26] Li, Q. & J.S. Racine (2007) *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[27] Mack, Y.P. (1981) Local Properties of k-NN Regression Estimates, *SIAM Journal on Algebraic and Discrete Methods* 2-3, 311-323.

[28] Marron, J. S. & W. Härdle (1986) Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation, *Journal of Multivariate Analysis* 20, 91-113.

[29] Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," Journal of Econometrics, Vol. 3, No. 3, 1975, pp. 205–228.

[30] Manski, C. (1988): "Identification of Binary Response Models," Journal of the American Statistical Association, Vol. 83, No. 403, pp. 729-738.

[31] Manski, C. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, vol. 72, no. 4, pp. 1221-46.

[32] Masry, E. (1996) Multivariate local polynomial regression for time series: uniform strong consistency and rates, *Journal of Time Series Analysis* 17, 571-599.

[33] Ogg , T., Anna Zimdars & Anthony Heath (2009): Schooling effects on degree performance: a comparison of the predictive validity of aptitude testing and secondary school grades at Oxford University, British Educational Research Journal, Vol. 35, Issue 5.

[34] Pagan, A. & A. Ullah (1999) *Nonparametric Econometrics*. Cambridge University Press.

[35] Parks, G. (2011). Academic Performance of International baccalaureate students at Cambridge, by school, http://www.cam.ac.uk/admissions/undergraduate/research/docs/ib_performance.pdf

[36] Persico, N (2009). Racial Profiling? Detecting Bias Using Statistical Evidence. Annual Review of Economics, volume 1.

[37] Rothstein, J. (2004). College Performance Predictions and the SAT, Journal of Econometrics 121(1-2), July-August 2004, 297-317.

[38] Sackett, P., Kuncel, N., Arneson, J., Cooper, G., Waters, S. (2009). Socioeconomic Status and the Relationship Between the SAT and Freshman GPA - An Analysis of Data from 41 Colleges and Universities, available online at: http://professionals.collegeboard.com/data-reports-research/cb/SES-SAT-FreshmanGPA

[39] Sawyer, R. (2010). Usefulness of High School Average and ACT Scores in Making College Admission Decisions, available online at: http://www.act.org/research/researchers/reports/pdf/ACT_RR2010-2.pdf

8 Technical Appendix

The appendix contains three subsections – subsection 8.1 formally states and derives the asymptotic distribution of the semiparametric estimator of γ_g, on which the application is based; subsection 8.2 outlines the proof of proposition 2; subsection 8.3 states and derives the distribution theory for the fully nonparametric estimator of γ_g.

8.1 Asymptotic distribution theory for semiparametric estimator

$$\hat{\gamma}_g = \frac{\sum_{i=1}^{n} K_h\left(\hat{p}\left(X_i, G_i\right) - 1/2\right) \hat{\mu}\left(X_i, G_i\right) \mathbf{1}\left\{G_i = g\right\}}{\sum_{l=1}^{n} K_h\left(\hat{p}\left(X_l, G_l\right) - 1/2\right) \mathbf{1}\left\{G_l = g\right\}},$$

where $K_h(z) = K(z/h)/h$; $K(\cdot)$ is a kernel function; h is a smoothing parameter (bandwidth); $\hat{p}(x,g)$ and $\hat{\mu}(x,g)$ are first-step estimators of $p(x,g)$ $(:= \Pr[D_i = 1 | (X_i, G_i) = (x,g)])$ and $\mu(x,g)$, respectively. For notational simplicity, we write $W_i := (X_i, G_i)$. We suppose that W_i consists of W_i^c and W_i^d, i.e., $W_i = \left(W_i^c, W_i^d\right)$, where the d_1-dimensional random (row) vector W_i^c is continuously distributed with its support $S^c (\subset \mathbb{R}^{d_1})$ compact; and the d_2-dimensional random (row) vector W_i^d takes discrete values with the support S^d (the number of points of support of S^d is finite). We let the last element of the vector W_i^d be G_i, the variable for gender. In what follows, we often write $(x,g) = w$ or $\left(w^c, w^d\right)$; $p(x,g) = p(w)$ or $p\left(w^c, w^d\right)$; and $\mu(x,g) = \mu(w)$ or $\mu\left(w^c, w^d\right)$.

In what follows and in the proofs, we will use the notations: $A := B$ and $C =: D$, where the former means that A is defined by B, and the latter means that D is defined by C. For a vector/matrix E whose elements are $\{E_{i,j} : 1 \leq i \leq I;\ 1 \leq j \leq J\}$ with I and J some positive integers, $||E|| = \max_{1 \leq i \leq I;\ 1 \leq j \leq J} |E_{i.j}|$.

Our semiparametric estimator can be written as follows:

$$\hat{\gamma}_{sp}(g) := \frac{\sum_{i=1}^{n} K_h\left(\bar{p}\left(X_i, G_i; \hat{\theta}_p\right) - 1/2\right) \bar{\mu}\left(X_i, G_i; \hat{\theta}_\mu\right) \mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(\bar{p}\left(X_l, G_l; \hat{\theta}_p\right) - 1/2\right) \mathbf{1}\{G_l = g\}}, \tag{10}$$

where $\bar{p}(x, g; \theta_p) = \bar{p}(w; \theta_p)$ is a consistent (semi) parametric estimator of $p(w)$ with a finite dimensional parameter θ_p; $\hat{\theta}_p$ is a consistent estimator for a (pseudo) true parameter θ_p^0; and $\bar{\mu}(x, g; \theta_\mu) = \bar{\mu}(w; \theta_\mu)$, θ_μ, $\hat{\theta}_\mu$ and θ_μ^0 are defined analogously. We may use various parametric models, e.g., a probit or logit model for $p(w)$; and a linear (regression) model for $\mu(w)$, whose requirements presented in Assumptions 11 and 12 are quite mild.

Asymptotic Distribution Theory:

Let $\tilde{\gamma}_g$ be defined as

$$\tilde{\gamma}_g := \frac{\sum_{i=1}^{n} K_h\left(p(X_i, g) - 1/2\right) \mu(X_i, g)}{\sum_{l=1}^{n} K_h\left(p(X_l, g) - 1/2\right)} = \frac{\sum_{i=1}^{n} K_h\left(p(W_i) - 1/2\right) \mu(W_i) \mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(p(W_l) - 1/2\right) \mathbf{1}\{G_l = g\}},$$

for each $g \in \{f, m\}$. This is not an feasible estimator, requiring true objects $p(x, g)$ and $\mu(x, g)$. However, we below show that our feasible semiparametric estimator $\hat{\gamma}_{sp}(g)$ and $\tilde{\gamma}_g$, under the assumption of correct specification, share the same asymptotic distribution.

To derive the asymptotic distribution of the infeasible estimator $\tilde{\gamma}_g$, we work with the following conditions:

Assumption 9 *Let*

$$m(z, g) := E[\mu(W_i) \mid p(W_i) = z,\ G_i = g] = E[\mu(X_i, g) \mid p(X_i, g) = z].$$

For each $g \in \{f, m\}$, $m(\cdot, g)$ is twice continuously differentiable on $[0, 1]$. The probability function $\nu(z, g)$ of random variables $p(W_i) (= \Pr[D_i = 1 | W_i])$ and G_i exists (ν: $[0, 1] \times \{f, m\} \to [0, \infty)$ and $\nu(z, g)\, dz = \Pr[p(W_i) \in dz, G_i = g]$); for each $g \in \{f, m\}$, $\nu(\cdot, g)$ is twice continuously differentiable on $[0, 1]$ and; there exists some constant $\epsilon > 0$ such that

$$\inf_{(z,g) \in [1/2-\epsilon, 1/2+\epsilon] \times \{f,m\}} \nu(z, g) > 0.$$

Assumption 10 *The kernel function $K(\cdot)$ ($\mathbb{R} \to [0, \infty)$) is of bounded variation and satisfies the following conditions: $\int_{\mathbb{R}} K(u)\, du = 1$; $\int_{\mathbb{R}} uK(u)\, du = 0$; there exists some constant $\bar{K} \in (0, \infty)$ such that $\sup_{u \in \mathbb{R}} K(u) \leq \bar{K}$ and $\int_{\mathbb{R}} u^2 |K(u)|\, du \leq \bar{K}$.*

Lemma 1 *Suppose that Assumptions 9-10 hold. Then, as $n \to \infty$ and $h \to 0$ with $nh \to \infty$ and $nh^5 = O(1)$,*

$$\sqrt{nh}\left[\tilde{\gamma}_g - \gamma_g - h^2 \mathbf{B}(g)\right] \xrightarrow{d} N(0, V(g)),$$

for each $g \in \{f, m\}$, where

$$\begin{aligned}
\mathbf{B}(g) &:= \int_{\mathbb{R}} u^2 K(u)\, du \left[\frac{(\partial/\partial z)\, m(z,g) \times (\partial/\partial z)\, \nu(z,g)}{\nu(z,g)} + \frac{1}{2}\left(\partial^2/\partial z^2\right) m(z,g)\right]\Big|_{z=1/2}; \\
\mathbf{V}(g) &:= \int_{\mathbb{R}} K^2(u)\, du \frac{\mathrm{Var}\left[\mu(X_i, g) \,|\, p(X_i, g) = z\right]}{\nu(z,g)}\Big|_{z=1/2}.
\end{aligned}$$

The conditions in Assumptions 9-10 and the result of this lemma is quite standard (see, e.g., Ch. 3 of Li and Racine, 2007), and therefore we omit the proof. We now analyze our semiparametric estimator $\hat{\gamma}_{sp}(g)$ under the following conditions:

Assumption 11 *(i) The estimator $\hat{\theta}_p$ is consistent for the (pseudo) true parameter θ_p^0 with the following expression:*

$$\hat{\theta}_p - \theta_p^0 = n^{-1} \sum\nolimits_{j=1}^{n} Z_{n,j} + O_p\left(n^{-1}\right), \tag{11}$$

where $\{Z_{n,j}\}$ is a vector-valued independent (triangular) array with $E[Z_{n,j}] = 0$ and $E[|Z_{n.j}|^2] < \infty$ (the dimension of $Z_{n,j}$ is the same as that of $\hat{\theta}_p$); and $Z_{n,j}$ is independent of (X_i, G_i, D_i) for $i \neq j$. (ii) There exists some compact set Θ_p such that $\bar{p}(w; \theta_p)$ is twice continuously differentiable with respect to θ_p in the interior of Θ_p; θ_p^0 is in the interior of Θ_p;

$$\sup_{w \in S^c \times S^d;\ \theta_p \in \Theta_p} \left|\left|(\partial/\partial\theta_p)\, \bar{p}\left(w; \theta_p^0\right)\right|\right| < \infty; \quad \textit{and} \quad \sup_{w \in S^c \times S^d;\ \theta_p \in \Theta_p} \left|\left|\left(\partial^2/\partial\theta_p \partial\theta_p'\right) \bar{p}\left(w; \theta_p^0\right)\right|\right| < \infty.$$

Assumption 12 *The estimator $\hat{\theta}_\mu$ is consistent for the (pseudo) true parameter θ_μ^0 with*

$$\sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|\bar{\mu}\left(w; \hat{\theta}_\mu\right) - \bar{\mu}\left(w; \theta_\mu^0\right)\right| = O_P\left(n^{-1/2}\right).$$

The condition on $\hat{\theta}_\mu$ in Assumption 12 is fairly weak and should be satisfied by many parametric models and estimators. The conditions on $\hat{\theta}_p$ in Assumption 11 are slightly stronger, but are also satisfied in many cases. The requirement of (11) can be usually fulfilled if $\hat{\theta}_p$ is an estimator of the extremum type. To see this point, let $\hat{\theta}_p$ be a (quasi) maximum-likelihood or (nonlinear) least-square type estimator with the $\sqrt{n}$-consistency. Then, it can be written as

$$\hat{\theta}_p - \theta_p^0 = H_n^{-1}\left(\tilde{\theta}_p\right) n^{-1} \sum\nolimits_{j=1}^{n} s_{n,j}\left(\theta_p^0\right), \tag{12}$$

where $\tilde{\theta}_p$ is on the line segment connecting $\hat{\theta}_p$ to θ_p; $\left\{s_{n,j}\left(\theta_p^0\right)\right\}$ is a zero-mean independent (score) array (with finite second moments); H_n is the Hessian matrix with $H_n\left(\theta_p\right) := n^{-1}\sum_{j=1}^n q_{n,j}\left(\theta\right)$; and $\{q_{n,j}\left(\theta\right)\}$ is an independent array (with finite second moments). Now, for notational simplicity, let the dimension of θ_p be 1 (a multi-dimension case can be treated analogously). In this case,

$$\begin{aligned}
& H_n^{-1}\left(\tilde{\theta}_p\right) \\
= & \left[E\left[H_n\left(\theta_p^0\right)\right] + n^{-1}\sum\nolimits_{j=1}^n \left(q_{n,j}\left(\theta_p^0\right) - E\left[q_{n,j}\left(\theta_p^0\right)\right]\right) - \left(\partial/\partial\theta_p\right) H_n\left(\check{\theta}_p\right)\left[\tilde{\theta}_p - \theta_p^0\right]\right]^{-1} \\
= & E\left[H_n\left(\theta_p^0\right)\right]^{-1} + O_p\left(1\right) \times \left\{n^{-1}\sum\nolimits_{j=1}^n \left(q_{n,j}\left(\theta_p^0\right) - E\left[q_{n,j}\left(\theta_p^0\right)\right]\right) - \left(\partial/\partial\theta_p\right) H_n\left(\check{\theta}_p\right)\left[\tilde{\theta}_p - \theta_p^0\right]\right\} \\
= & E\left[H_n\left(\theta_p^0\right)\right]^{-1} + O_p\left(n^{-1/2}\right), \qquad (13)
\end{aligned}$$

where $\check{\theta}_p$ is on the line segment connecting $\tilde{\theta}_p$ to θ_p; the second equality uses the Taylor expansion (for $f\left(x\right) = 1/x$), as well as some suitable differentiability and boundedness conditions on $q_{n,j}\left(\theta\right)$; and the last equality holds since $\{\left(q_{n,j}\left(\theta\right) - E\left[q_{n,j}\left(\theta\right)\right]\right)\}$ is a zero-mean independent array and $\tilde{\theta}_p - \theta_p^0 = O_p\left(n^{-1/2}\right)$. By (12) and (13), we can write

$$\begin{aligned}
\hat{\theta}_p - \theta_p^0 &= \left\{E\left[H_n\left(\theta_p^0\right)\right]^{-1} + O_p\left(n^{-1/2}\right)\right\} n^{-1}\sum\nolimits_{j=1}^n s_{n,j}\left(\theta_p^0\right) \\
&= n^{-1}\sum\nolimits_{j=1}^n E\left[H_n\left(\theta_p^0\right)\right]^{-1} s_{n,j}\left(\theta_p^0\right) + O_p\left(n^{-1/2}\right) \times n^{-1}\sum\nolimits_{j=1}^n s_{n,j}\left(\theta_p^0\right).
\end{aligned}$$

Noting that $n^{-1}\sum_{j=1}^n s_{n,j}\left(\theta_p^0\right) = O_p\left(n^{-1/2}\right)$ and letting $Z_{n,j} = E\left[H_n\left(\theta_p^0\right)\right]^{-1} s_{n,j}\left(\theta_p^0\right)$, we can check that $\hat{\theta}_p - \theta_p^0$ has the expression as in (11).

We also impose the following condition on the kernel function $K\left(\cdot\right)$.

Assumption 13 *The kernel function $K(: \mathbb{R} \to [0,\infty))$ is twice continuously differentiable whose support is a compact interval in $\mathbb{R}$.*

This assumption rules out some class of kernel functions, e.g., the normal kernel. While we might be able to relax the compactness condition by imposing some other explicit condition on the tail decay (say, Assumption 3 in Hansen, 2008), we maintain this for the sake of simplicity in our proof.

Theorem 1 *Let the kernel function K satisfy the conditions in Assumptions 9 and 13. Suppose that Assumptions 11 and 12 hold. If $\bar{p}\left(w;\theta_p^0\right) = p\left(w\right)$ and $\bar{\mu}\left(w;\theta_\mu^0\right) = \mu\left(w\right)$, then it holds that as $n \to \infty$ and $h \to 0$ with $nh^3 \to \infty$,*

$$\sqrt{nh}\left[\hat{\gamma}_{sp}\left(g\right) - \tilde{\gamma}_g\right] = o_P\left(1\right).$$

Therefore, additionally if $nh \to \infty$ *and* $nh^5 = O(1)$*, the asymptotic bias and distribution of* $\hat{\gamma}_{sp}(g)$ *are the same as those for* $\tilde{\gamma}_g$ *given in Lemma 1.*

o. f of theorem 1]: First, consider the convergence of the numerator of (10). We have the following decomposition:

$$\begin{aligned} & (\sqrt{nh}/n)\sum_{i=1}^{n}\left[K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)\bar{\mu}\left(W_i;\hat{\theta}_\mu\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\bar{\mu}\left(W_i;\theta_\mu^0\right)\right] \\ = \;& \bar{A}_n+\bar{B}_n+\bar{C}_n, \end{aligned}$$

where

$$\begin{aligned} \bar{A}_n \;:\; & = (\sqrt{nh}/n)\sum_{i=1}^{n}K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\left[\bar{\mu}\left(W_i;\hat{\theta}_\mu\right)-\bar{\mu}\left(W_i;\theta_\mu^0\right)\right]\mathbf{1}\left\{G_i=g\right\}; \\ \bar{B}_n \;:\; & = (\sqrt{nh}/n)\sum_{i=1}^{n}\left[K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\right]\bar{\mu}\left(W_i;\theta_\mu^0\right)\mathbf{1}\left\{G_i=g\right\}; \\ \bar{C}_n \;:\; & = (\sqrt{nh}/n)\sum_{i=1}^{n}\left[K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\right] \\ & \quad\times\left[\bar{\mu}\left(W_i;\hat{\theta}_\mu\right)-\bar{\mu}\left(W_i;\theta_\mu^0\right)\right]\mathbf{1}\left\{G_i=g\right\}. \end{aligned}$$

By Assumptions 11 and 13, we can easily see that

$$\bar{A}_n = O_P(\sqrt{h}) \;\text{ and }\; \bar{C}_n = O_P(1/\sqrt{nh}).$$

To find the convergence rate of $\bar{B}_n$, observe that

$$\begin{aligned} & K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right) \\ = \;& \frac{1}{h^2}K'\left(\frac{\bar{p}\left(W_i;\theta_p^0\right)-z_0}{h}\right)\left(\partial/\partial\theta'\right)\bar{p}\left(W_i;\theta_p^0\right)\left[\hat{\theta}_p-\theta_p^0\right] \\ & +\frac{1}{2h^3}K''\left(\frac{\bar{p}\left(W_i;\theta_p^0\right)-z_0}{h}\right)\left[\hat{\theta}_p-\theta_p^0\right]'\left(\partial^2/\partial\theta'\partial\theta\right)\bar{p}\left(W_i;\tilde{\theta}_p\right)\left[\hat{\theta}_p-\theta_p^0\right], \end{aligned}$$

where $\tilde{\theta}_p$ is on the line segment connecting $\hat{\theta}_p$ to θ_p^0. Then, since $\hat{\theta}_p-\theta_p^0 = n^{-1}\sum_{i=1}^{n}Z_{n,j}+O_p\left(n^{-1}\right)$ (Assumption 11),

$$\begin{aligned} \bar{B}_{n,1} \;=\; & \frac{\sqrt{nh}}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K'\left(\frac{\bar{p}\left(W_i;\theta_p^0\right)-z_0}{h}\right)\left(\partial/\partial\theta'\right)\bar{p}\left(W_i;\theta_p^0\right)Z_j\right]\bar{\mu}\left(W_i;\theta_p^0\right)\mathbf{1}\left\{G_i=g\right\} \\ & +\frac{\sqrt{nh}}{n^3h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K'\left(\frac{\bar{p}\left(W_i;\theta_p^0\right)-z_0}{h}\right)\left(\partial/\partial\theta'\right)\bar{p}\left(W_i;\theta_p^0\right)\right]\bar{\mu}\left(W_i;\theta_p^0\right)\mathbf{1}\left\{G_i=g\right\}\times O_P(1) \\ = \;& :\bar{B}_{n,1,1}+\bar{B}_{n,1,2}. \end{aligned}$$

We have

$$\begin{aligned} E\left[|\bar{B}_{n,1,1}|^2\right] \;&=\; \left(nh/n^4h^4\right)\times O(nh) = O(1/n^2h^2); \\ E\left[|\bar{B}_{n,1,2}|\right] \;&=\; (\sqrt{nh}/n^3h^2)\times O_p\left(n^2\right) = O_p(1/n^{1/2}h^{3/2}), \end{aligned}$$

where the former follows from the zero-mean and independence proprieties of $\{Z_{n,j}\}$. From these arguments, we can see that

$$\begin{aligned}\text{the numerator of (10)} &= (\sqrt{nh}/n)\sum_{i=1}^{n} K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\bar{\mu}\left(W_i;\theta_\mu^0\right) \\ &\quad +O_P(\sqrt{h})+O_p(1/n^{1/2}h^{3/2})+O_P(1/\sqrt{nh}). \qquad (14)\end{aligned}$$

by almost the same arguments as for $\bar{B}_{n,1}$, we can also show that

$$\text{the denominator of (10)} = n^{-1}\sum_{l=1}^{n} K_h\left(\bar{p}\left(W_l\right)-z_0;\theta_p^0\right)\mathbf{1}\left\{G_l=g\right\}+O_P(1/n^{3/2}h^{3/2})+O_p(1/nh^2),$$

which, together with (14), leads to the desired result. ■

8.2 Derivation of Bounds on Productivity Change (Proposition 2)

Let $\tilde{A}$ denote acceptance of an offer which would be made using the new threshold $\tilde{\gamma}$; analogously let $\tilde{\mu}(x,m)$ denote $E\left(Y|X=x,G=m,\tilde{A}=1\right)$.

The magnitude of the productivity loss can then be calculated as the difference between the gain from admitting more men less the loss from admitting fewer women: $\omega(\tilde{\gamma})=\omega_m(\tilde{\gamma})-\omega_f(\tilde{\gamma})$ where

$$\begin{aligned}\omega_m(\tilde{\gamma}) &= E_{X|G=m}\left\{E\left(\tilde{A}Y|X,m\right)\Pr{}_{\varepsilon|X,G=m}\left[\tilde{\gamma}\le\mu(X,m)-\varepsilon<\gamma_m\right]\right\}\times\pi_m \\ \omega_f(\tilde{\gamma}) &= E_{X|G=f}\left\{E\left(\tilde{A}Y|X,f\right)\Pr{}_{\varepsilon|X,G=f}\left[\gamma_f\le\mu(X,f)-\varepsilon<\tilde{\gamma}\right]\right\}\times(1-\pi_m), \qquad (15)\end{aligned}$$

where π_m denotes the fraction of males among the applicants, $\mu(x,m)$ denotes as before $E(Y|X=x,G=m,A=1)$

First, note that since $\gamma_m>\tilde{\gamma}$, if $\mu(x,m)+\gamma_m-\tilde{\gamma}\le\sup_{z\in\mathcal{X}_m}\mu(z,m)$, then there will exist $\tilde{x}(x)\in\mathcal{X}_g$ satisfying $\mu(\tilde{x}(x),m)=\mu(x,m)+\gamma_m-\tilde{\gamma}$. Then

$$\begin{aligned}&\Pr{}_{\varepsilon|X=x,G=m}\left[\tilde{\gamma}\le\mu(X,m)-\varepsilon<\gamma_m\right] \\ &\overset{\text{by (6)}}{=} F_{\varepsilon|G=m}\left(\mu(x,m)-\tilde{\gamma}\right)-F_{\varepsilon|G=m}\left(\mu(x,m)-\gamma_m\right) \\ &= F_{\varepsilon|G=m}\left(\mu(\tilde{x}(x),m)-\gamma_m\right)-F_{\varepsilon|G=m}\left(\mu(x,m)-\gamma_m\right) \\ &= \Pr\left(D=1|X=\tilde{x}(x),G=m\right)-\Pr\left(D=1|X=x,G=m\right).\end{aligned}$$

Therefore, we can identify $\mu(\tilde{x}(x),m)$ as the observed $E(Y|X=\tilde{x}(x),G=m,A=1)$ and thus the probability in the previous display is identifiable from the existing data, as well. Now suppose $\mu(x,m)+\gamma_m-\tilde{\gamma}>\max_{x\in\mathcal{X}_m}\mu(x,m)$. This would imply that

$$F_{\varepsilon|G=m}\left(\mu(x,m)-\tilde{\gamma}\right)>F_{\varepsilon|G=m}\left(\max_{x\in\mathcal{X}_m}\mu(x,m)-\gamma_m\right)=\max_{x\in\mathcal{X}_m}\Pr\left(D=1|X=x,G=m\right)$$

so that we can bound $F_{\varepsilon|G=m}\left(\mu\left(x,m\right)-\tilde{\gamma}\right)$ by the interval $\left(\max_{x\in\mathcal{X}_m}\Pr\left(D=1|X=x,G=m\right),1\right]$. If there exists $x\in\mathcal{X}_m$ such that $\Pr\left(D=1|X=x,G=m\right)=1$, then the interval shrinks to the singleton 1.

The remaining step is to calculate $E\left(\tilde{A}Y|X,m\right)$ which, by assumption 8 is given by $E\left(Y|X,G=m,A=1\right)$, which is identified from the current data.

Putting all of this together, we get the following bounds for $\omega_m\left(\tilde{\gamma}\right)$: with $p\left(x,m\right)$ denoting $\Pr\left(D=1|X=x,G=m\right)$ and stationarity implying $\mu^P\left(x,g\right)\equiv\mu\left(x,g\right)\equiv E\left(Y|X=x,G=g,A=1\right)$,

$$\begin{aligned}
& \omega_m\left(\tilde{\gamma}\right) \\
\geq\ & \pi_m\int_{x\in M}E\left(AY|D=1,x,m\right)\left\{p\left(\tilde{x}\left(x\right),m\right)-p\left(x,m\right)\right\}dF\left(x|G=m\right) \\
& +\pi_m\int_{x\in\mathcal{X}_m\cap M^c}E\left(AY|D=1,x,m\right)\left\{\sup_{z\in\mathcal{X}_m}p\left(z,m\right)-p\left(x,m\right)\right\}dF\left(x|G=m\right) \\
:\ & =\omega_{ml}\left(\tilde{\gamma}\right),
\end{aligned}$$

and

$$\begin{aligned}
& \omega_m\left(\tilde{\gamma}\right) \\
\leq\ & \pi_m\int_{x\in M}E\left(AY|D=1,x,m\right)\left\{p\left(\tilde{x}\left(x\right),m\right)-p\left(x,m\right)\right\}dF\left(x|G=m\right) \\
& +\pi_m\int_{x\in\mathcal{X}_m\cap M^c}E\left(AY|D=1,x,m\right)\left\{1-p\left(x,m\right)\right\}dF\left(x|G=m\right) \\
:\ & =\omega_{mu}\left(\tilde{\gamma}\right),
\end{aligned}$$

and in case $\sup_{z\in\mathcal{X}_m}p_m\left(z\right)=1$, we have that

$$\begin{aligned}
& \omega_m\left(\tilde{\gamma}\right) \\
=\ & \pi_m\int_{x\in M}E\left(AY|D=1,x,m\right)\left\{p\left(\tilde{x}\left(x\right),m\right)-p\left(x,m\right)\right\}dF\left(x|G=m\right) \\
& +\pi_m\int_{x\in\mathcal{X}_m\cap M^c}E\left(AY|D=1,x,m\right)\left\{1-p\left(x,m\right)\right\}dF\left(x|G=m\right).
\end{aligned}$$

Lastly, consider

$$\omega_f\left(\tilde{\gamma}\right)=E_{X|G=f}\left\{\Pr\left(\tilde{A}Y|\tilde{D}=1,X,f\right)\Pr{}_{\varepsilon|X,G=f}\left[\gamma_f\leq\mu\left(X,f\right)-\varepsilon<\tilde{\gamma}\right]\right\}\times\left(1-\pi_m\right)$$

For any $x\in\mathcal{X}_f$, if $\mu\left(x,f\right)+\gamma_f-\tilde{\gamma}\geq\min_{z\in\mathcal{X}_f}\mu\left(z,f\right)$, then there will exist $\bar{x}\left(x\right)\in\mathcal{X}_f$ such that $\mu\left(\bar{x}\left(x\right),f\right)=\mu\left(x,f\right)+\gamma_f-\tilde{\gamma}$. Then for any such x,

$$\begin{aligned}
& \Pr{}_{\varepsilon|X=x,G=f}\left[\gamma_f\leq\mu\left(X,f\right)-\varepsilon<\tilde{\gamma}\right] \\
\overset{\text{by }(6)}{=}\ & F_{\varepsilon|G=f}\left(\mu\left(x,f\right)-\gamma_f\right)-F_{\varepsilon|G=f}\left(\mu\left(x,f\right)-\tilde{\gamma}\right) \\
=\ & \Pr\left(D=1|X=x,G=f\right)-\Pr\left(D=1|X=\bar{x}\left(x\right),G=f\right).
\end{aligned}$$

On the other hand, for any $x \in \mathcal{X}_f$ such that $\mu(x,f) + \gamma_f - \tilde{\gamma} < \min_{z \in \mathcal{X}_f} \mu(z,f)$, by an analogous argument to the one for males, we will have that

$$0 \leq F_{\varepsilon|G=f}\left(\mu(x,f) - \tilde{\gamma}\right) < F_{\varepsilon|G=f}\left(\min_{z \in \mathcal{X}_f} \mu(z,f) - \gamma_f\right) = \min_{z \in \mathcal{X}_f} \Pr\left(D=1|X=z, G=f\right).$$

And, as before, if $\min_{z \in \mathcal{X}_f} \Pr\left(D=1|X=z, G=f\right) = 0$, then the bounds collapse to the singleton zero. The independence assumption 8 then gives us the probability of acceptance under the new threshold. Finally, since $\tilde{\gamma} > \gamma_f$, for any $x \in \mathcal{X}_f$ we must have that $\Pr\left(\tilde{D}=1|X=x, G=f\right) > 0$. Therefore, $\mu(x,f)$ is directly identified for all x for which the probability of admission is positive under the new threshold.

8.3 Distribution theory in fully nonparametric case

Consider a nonparametric estimator of γ_g as a solution to the following estimating equation:

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} K_h\left(\hat{p}_{-i}(X_i, g) - 1/2\right)\left[\hat{\mu}_{-i}(X_i, g) - \gamma_g\right] = 0 \quad \text{for each } g, \tag{16}$$

where $K_h(z) = K(z/h)/h$; $K(\cdot)$ is a kernel function ($\mathbb{R} \to [0,\infty)$); h is a smoothing parameter (bandwidth); $\hat{p}_{-i}(x,g)$ and $\hat{\mu}_{-i}(x,g)$ are leave-one-out nonparametric estimators of $p(x,g)(:= \Pr\left[D_i = 1|\left(X_i, G_i\right) = (x,g)\right])$ and $\mu(x,g)$, respectively, whose forms are provided in (18) and (19).

The estimator of γ_g has the following closed-form expression:

$$\hat{\gamma}_{np}(g) = \frac{\sum_{i=1}^{n} K_h\left(\hat{p}_{-i}(X_i, G_i) - 1/2\right)\hat{\mu}_{-i}(X_i, G_i)\,\mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(\hat{p}_{-l}(X_l, G_l) - 1/2\right)\mathbf{1}\{G_l = g\}}. \tag{17}$$

For notational simplicity, we write $W_i := (X_i, G_i)$. We suppose that W_i consists of W_i^c and W_i^d, i.e., $W_i = \left(W_i^c, W_i^d\right)$, where the d_1-dimensional random (row) vector W_i^c is continuously distributed with its support $S^c (\subset \mathbb{R}^{d_1})$ compact; and the d_2-dimensional random (row) vector W_i^d takes discrete values with the support S^d (the number of points of support of S^d is finite). We let the last element of the vector W_i^d be G_i, the variable for gender. In what follows, we often write $(x,g) = w$ or $\left(w^c, w^d\right)$; $p(x,g) = p(w)$ or $p\left(w^c, w^d\right)$; and $\mu(x,g) = \mu(w)$ or $\mu\left(w^c, w^d\right)$.

For estimating $p(x,g)$ and $\mu(x,g)$, consider the following nonparametric estimators:

$$\hat{p}_{-i}(x,g) = \hat{p}_{-i}(w) := \frac{\sum_{1 \leq j \leq n;\ j \neq i} L_{\xi_p}\left(W_j^c - w^c\right)\mathbf{1}\left\{W_j^d = w^d\right\} D_j}{\sum_{1 \leq k \leq n;\ k \neq i} L_{\xi_p}\left(W_k^c - w^c\right)\mathbf{1}\left\{W_k^d = w^d\right\}}; \quad \text{and} \tag{18}$$

$$\hat{\mu}_{-i}(x,g) = \hat{\mu}_{-i}(w) := \frac{\sum_{1 \leq j \leq n;\ j \neq i} M_{\xi_\mu}\left(W_j^c - w^c\right)\mathbf{1}\left\{W_j^d = w^d\right\} Y_j D_j}{\sum_{1 \leq k \leq n;\ k \neq i} M_{\xi_\mu}\left(W_k^c - w^c\right)\mathbf{1}\left\{W_k^d = w^d\right\} D_j}, \tag{19}$$

where $L_a(z) := L(z/a)/a^{d_1} = L(z_1/a, \ldots, z_{d_1}/a)/a^{d_1}$ for $z \in \mathbb{R}^{d_1}$ and $a > 0$; $L(\cdot)$ is a kernel function ($\mathbb{R}^{d_1} \to \mathbb{R}$); $\xi_p\,(>0)$ is a smoothing parameter/bandwidth; $M_a(z)$ is defined analogously; $M(\cdot)$ is another kernel function and ξ_μ is another bandwidth.

Remark 4 *We let bandwidths, ξ_p and ξ_μ, be common for all components of continuously distributed variables. This is mainly for (notational) simplicity, and we may use bandwidth matrices (as long as the rate conditions provided below are satisfied), Ξ_p, $\Xi_\mu \in \mathbb{R}^{d_1 \times d_1}$, allowing for different bandwidths for different components. In this case, $L_{\xi_p}\left(W_j^c - w^c\right)$ in (18) is replaced by $L\left(\Xi_p^{-1}\left(W_j^c - w^c\right)\right)/\det(H)$ ($\det(H)$ is the determinant of H).*

Remark 5 *The suggested estimators (17), (18) and (19) are of the form of so-called frequency estimators (see, e.g., Ch. 3 of Li and Racine, 2007), which do not use any smoothing for discrete variables. The use of these estimators is only for simplicity, and we can instead think of estimators smoothing discrete variables, as found in Ch. 4 of Li and Racine (2007).*

Asymptotic behavior of the nonparametric estimator: We here show that the asymptotic distribution of $\hat{\gamma}_{np}(g)$ is determined by that of $\tilde{\gamma}_g$. For this purpose, we work with the following conditions:

Assumption 14 *For each $w^d \in S^d$, the functions $p(\cdot, w^d)$, $\mu(\cdot, w^d)$ and $f(\cdot, w^d)$ are compactly supported on S^c. Let κ_p, κ_μ be some positive integers with $\kappa_p, \kappa_\mu \geq 2$. For each $w^d \in S^d$, $p(\cdot, w^d)$ is κ_p-times continuously differentiable on $\mathbb{R}^{d_1}$ except for boundary points of S^c, $\mu(\cdot, w^d)$ is κ_μ-times continuously differentiable on $\mathbb{R}^{d_1}$ except for the boundary points of S^c, and $f(\cdot, w^d)$ is $\max\{\kappa_p, \kappa_\mu\}$-times continuously differentiable on $\mathbb{R}^{d_1}$ except for the boundary points of S^c.*

Assumption 15 *The kernel function $L(\cdot)$ ($\mathbb{R}^{d_1} \to \mathbb{R}$) satisfies the following conditions: , the support $S^L(\subseteq \mathbb{R}^{d_1})$ of L is bounded; $L(\cdot)$ is continuously differentiable on $\mathbb{R}^{d_1}$; $\int_{\mathbb{R}^{d_1}} L(u)\,du = 1$; and $L(\cdot)$ is the κ_pth-order kernel, i.e., $\int_{\mathbb{R}^{d_1}} \left[\bigotimes_{l=1}^k u\right] L(u)\,du = 0$ for $k = 1, \ldots, (\kappa_p - 1)$.*

Assumption 16 *The kernel function $M(\cdot)$ ($\mathbb{R}^{d_1} \to \mathbb{R}$) satisfies the following conditions: the support $S^M(\subseteq \mathbb{R}^{d_1})$ of M is bounded; $M(\cdot)$ is continuously differentiable on $\mathbb{R}^{d_1}$; $\int_{\mathbb{R}^{d_1}} M(u)\,du = 1$ and $M(\cdot)$ is the κ_μth-order kernel, i.e., $\int_{\mathbb{R}^{d_1}} \left[\bigotimes_{l=1}^k u\right] M(u)\,du = 0$ for $k = 1, \ldots, (\kappa_\mu - 1)$.*

Assumption 17 *(i) The probability function of W_i (i.e., a function $f\left(w^c, w^d\right)$ satisfying $f\left(w^c, w^d\right)dw^c = \Pr\left[W_i^c \in dw^c, W_i^d = w^d\right]$) exists, and there exists some constant $C_1 \in (0, \infty)$ such that*

$$C_1 \leq \inf_{\left(w^c, w^d\right) \in S^c \times S^d} f\left(w^c, w^d\right).$$

(ii) There exists some set $S_\circ^c$ such that if $D_i = 1$, then

$$W_i^c \in S_\circ^c \subsetneq S^c,$$

and that any boundary points of $S_\circ^c$ are in the interior of S^c, and there exist some constant $C_2 \in (0, \infty)$ such that

$$C_2 \leq \inf_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} p\left(w^c, w^d\right).$$

Assumptions 14-16 are standard for establishing uniform convergence results for $\hat{p}(w)$ and $\hat{\mu}(w)$ (see Lemmas 2 and 3 in the Appendix). The last condition in Assumption 15 and that in Assumption 16 require that the kernels are of higher order (bias reducing) of orders κ_p and κ_μ, respectively. These conditions are used to guarantee that the estimation errors due to the first step are negligible in the second step. We impose Assumption 17 to avoid the so-called boundary-bias problem. Our nonparametric estimators $\hat{p}(w)$ and $\hat{\mu}(w)$ are of the Nadaraya-Watson type (with symmetric kernel functions), and have slower uniform convergence rates around the boundary points of S^c (see, e.g., arguments in Bouezmarni and Scaillet, 2005).[17] The condition (ii) of Assumption 17 is similar to that imposed in Ahn and Powell (1993), called "exogenous trimming," which, together with the condition (i), is useful to allow us to avoid the so-called random-denominator problem. Note that these two conditions are imposed only for simplicity. We may be able to proceed without (i) and/or (ii) of Assumption 17. However, to do so, we will require a trimming device and more intricate conditions on the bandwidths and trimming parameters.

Given these conditions, we obtain the following result:

Theorem 2 *Let the kernel function K satisfy the conditions in Assumptions 10 and 13. Suppose that Assumptions 9, 14, 15, 16 and 17 hold. Let*

$$\begin{aligned}
\Delta_{n,a} &: = nh\xi_\mu^{2\kappa_\mu} + (\log n)\, h\xi_\mu^{2\kappa_\mu - d_1} + n^{-1}(\log n)^2 h\xi_\mu^{-2d_1}; \\
\Delta_{n,b} &: = h^{-1}\xi_p^2 + n\xi_p^{2\kappa_p} + nh^{-3}\xi_p^{4\kappa_p} + n^{-1}(\log n)^2 h^{-3}\xi_p^{-2d_1} + n^{-1}(\log n)\,\xi_p^{2\kappa_p - d_1}; \\
\Delta_{n,c} &: = nh^{-1}\xi_\mu^{2\kappa_\mu}\xi_p^{2\kappa_p} + n^{-1}(\log n)^2 h^{-1}\xi_\mu^{-d_1}\xi_p^{-d_1} + (\log n)\, h^{-1}\xi_\mu^{2\kappa_\mu}\xi_p^{-d_1} + (\log n)\, h^{-1}\xi_\mu^{-d_1}\xi_p^{2\kappa_p}.
\end{aligned}$$

Let $n \to \infty$, and h, ξ_p and $\xi_\mu \to 0$ with $[\log(\log n)]^4 [\log n]^2 / n\xi_p^{d_1} \to 0$, $(\log n)/n\xi_\mu^{d_1} \to 0$, and $\Delta_{n,a}, \Delta_{n,b}, \Delta_{n,c} \to 0$. Then, for each $g \in \{f, m\}$,

$$\sqrt{nh}\left[\hat{\gamma}_{np}(g) - \tilde{\gamma}_g\right] = O_P(\sqrt{\Delta_{n,a} + \Delta_{n,b} + \Delta_{n,c}}) = o_P(1).$$

[17] The boundary bias may be avoided by using asymmetric kernels as in Bouezmarni and Scaillet (2005) and Gospodinov and Hirukawa (2010), or by usimg the local polynomial method as in Masry (1996).

Therefore, additionally if $nh \to \infty$ *and* $nh^5 = O(1)$*, the asymptotic bias and distribution of* $\hat{\gamma}_{np}(g)$ *are the same as those for* $\tilde{\gamma}_g$ *given in Lemma 1.*

To prove the above theorem, we will utilize two lemmas.

Lemma 2 *Suppose that Assumptions 14, 15 and 17 hold. Let* $n \to \infty$ *and* $\xi_p \to 0$ *with* $[\log(\log n)]^4 (\log n)^2 / n\xi_p^{d_1} \to 0$*. Then, it holds that*

$$\begin{aligned} \hat{p}_{-i}(w) - p(w) &= n^{-1} \sum_{1 \le j \le n;\ j \ne i} L_{\xi_p}\left(W_j^c - w^c\right) \mathbf{1}\left\{W_j^d = w^d\right\} \left[p(W_j) - p(w)\right] / f(w) \\ &\quad + O_{a.s.}([\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2); \end{aligned} \tag{20}$$

uniformly over $i \in \{1, \ldots, n\}$ *and* $w \in S_\circ^c \times S^d$*; and that*

$$\hat{p}_{-i}(w) - p(w) = O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}), \tag{21}$$

uniformly over $i \in \{1, \ldots, n\}$ *and* $w \in S^c \times S^d$*.*

Lemma 3 *Suppose that Assumptions 14, 16 and 17 hold. Let* $n \to \infty$ *and* $\xi_\mu \to 0$ *with with* $(\log n)/n\xi_\mu^{d_1} \to 0$*. Then, it holds that*

$$\begin{aligned} \hat{\mu}_{-i}(w) - \mu(w) &= n^{-1} \sum_{1 \le j \le n;\ j \ne i} M_{\xi_\mu}\left(W_j^c - w^c\right) \mathbf{1}\left\{W_j^d = w^d\right\} D_j \left[\mu(W_j) - \mu(w)\right] / f(w)\, p(w) \\ &\quad + O_P([\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}]^2); \end{aligned} \tag{22}$$

$$\hat{\mu}_{-i}(w) - \mu(w) = O_P(\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}), \tag{23}$$

uniformly over $i \in \{1, \ldots, n\}$ *and* $w \in S_\circ^c \times S^d$*.*

The proofs of these lemmas are provided below.[18] Now, we start the proof of Theorem 2. We write $z_0 = 1/2$ throughout this proof.

[18] To establish the almost sure convergence result, we impose a slightly stronger condition on the bandwidth in Lemma 2 than in Lemma 3. The almost sure result might not be necessarily required, but it turns out to be very useful. In particular, it allows us to obtain a sharp convergence rate between $K_h(p(X_i, g) - 1/2)$ and $K_h(\hat{p}(X_i, g) - 1/2)$ without some extra rate loss due to h. The almost sure result is also useful for us to avoid the boundary bias problem under a simple compact-support condition on K (imposed below). For these technical points, see the arguments in deriving (25). Note also that except for (21), the uniform rates are established over the set $S_\circ^c \times S^d$, where $S_\circ^c$ is some subset of S^c given in Assumption 17. We may be able to derive uniform rates over $S^c \times S^d$. However, under the compact-support condition of K, our current results are sufficient for our purpose.

s. t look at the denominator of (17).

$$\frac{1}{n}\sum_{l=1}^{n} K_h\left(\hat{p}_{-l}\left(W_l\right)-z_0\right)\mathbf{1}\left\{G_l=g\right\} = \underbrace{\frac{1}{nh}\sum_{l=1}^{n} K\left(\frac{p\left(W_l\right)-z_0}{h}\right)\mathbf{1}\left\{G_l=g\right\}}_{=:J_{n,1}}$$

$$+\underbrace{\frac{1}{nh^2}\sum_{l=1}^{n} K'\left(\frac{\check{p}_{-l}\left(W_l\right)-z_0}{h}\right)\mathbf{1}\left\{G_l=g\right\}\left[\hat{p}_{-l}\left(W_l\right)-p\left(W_l\right)\right]}_{=:J_{n,2}}, \tag{24}$$

where $\check{p}_{-l}\left(W_l\right)$ is on the line segment connecting $\hat{p}_{-l}\left(W_l\right)$ to $p\left(W_l\right)$, and the first equality holds by the mean-value theorem. We find the probability bounds of $J_{n,1}$ and $J_{n,2}$. Now, by standard arguments for kernel-based estimators (see, e.g., Ch. 3 of Li and Racine, 2007), we can show that $J_{n,1} \xrightarrow{p} \nu\left(z_0, g\right)$. To find the bound of $J_{n,2}$, we note that by Assumptions 10 and 13, there exist some function $\mathcal{K}^*(\cdot)$ and some positive constant $\bar{\varepsilon}\,(>0)$ such that $\sup_{|\varepsilon|\leq\bar{\varepsilon}}\left|K'\left(u+\varepsilon\right)\right| \leq \mathcal{K}^*\left(u\right)$ for any $u \in \mathbb{R}$, $\sup_{u\in\mathbb{R}}\left|\mathcal{K}^*\left(u\right)\right| < \infty$ and $\int_{\mathbb{R}}\left|\mathcal{K}^*\left(u\right)\right|du < \infty$. Note also that $\left[\check{p}_{-l}\left(W_l\right)-z\right]/h = \left[p_{-l}\left(W_l\right)-z\right]/h + o_{a.s.}\left(1\right)$ uniformly over $l \in \{1,\ldots,n\}$ (by the result (21) of Lemma 2 and the stated condition on the shrinking rates of ξ_p and h). Therefore, for each $\omega \in \Omega^*$ (Ω^* is an event with $\Pr\left(\Omega^*\right)=1$), there exists some $\bar{n}$ such that for any $n \geq \bar{n}$,

$$\left|\frac{\check{p}_{-l}\left(W_l\right)-z_0}{h}-\frac{p_{-l}\left(W_l\right)-z_0}{h}\right| = \left|\frac{\check{p}_{-l}\left(W_l\right)-p_{-l}\left(W_l\right)}{h}\right| \leq \bar{\varepsilon}.$$

Therefore, for n large enough, it holds that

$$\begin{aligned} J_{n,2} &\leq \frac{1}{nh^2}\sum_{l=1}^{n}\mathcal{K}^*\left(\frac{p_{-l}\left(W_l\right)-z_0}{h}\right)\times\max_{1\leq l\leq n}\sup_{w\in S^c\times S^d}\left|\hat{p}_{-l}\left(w\right)-p\left(w\right)\right| \\ &= O_P\left(1/h\right)\times O_{a.s.}(\xi_p^{\kappa_p}+\sqrt{\left(\log n\right)/n\xi_p^{d_1}}), \end{aligned} \tag{25}$$

where the equality holds since

$$\frac{1}{nh}\sum_{l=1}^{n}\mathcal{K}^*\left(\frac{p_{-l}\left(W_l\right)-z_0}{h}\right) = O_p\left(1\right), \tag{26}$$

which follows from standard arguments for kernel-based estimators. Now, we have shown that

$$\frac{1}{n}\sum_{l=1}^{n} K_h\left(\hat{p}_{-l}\left(X_l,g_l\right)-z_0\right)\mathbf{1}\left\{g_l=g\right\} = \nu\left(z_0,g\right)+o_P\left(1\right).$$

Next, we look at a scaled version of the numerator of (17). Look at the following decomposition:

$$\left(\sqrt{nh}/n\right)\sum_{i=1}^{n} K_h\left(\hat{p}\left(W_i\right)-z_0\right)\hat{\mu}\left(W_i\right) = A_n+B_n+C_n, \tag{27}$$

where

$$\begin{aligned} A_n &: = \left(\sqrt{nh}/n\right)\sum_{i=1}^{n} K_h\left(p\left(W_i\right)-z_0\right)\left[\hat{\mu}_{-i}\left(W_i\right)-\mu\left(W_i\right)\right]\mathbf{1}\left\{G_i=g\right\}; \\ B_n &: = \left(\sqrt{nh}/n\right)\sum_{i=1}^{n}\left[K_h\left(\hat{p}_{-i}\left(W_i\right)-z_0\right)-K_h\left(p\left(W_i\right)-z_0\right)\right]\mu\left(W_i\right)\mathbf{1}\left\{G_i=g\right\}; \\ C_n &: = \left(\sqrt{nh}/n\right)\sum_{i=1}^{n}\left[K_h\left(\hat{p}_{-i}\left(W_i\right)-z_0\right)-K_h\left(p\left(W_i\right)-z_0\right)\right]\left[\hat{\mu}_{-i}\left(W_i\right)-\mu\left(W_i\right)\right]\mathbf{1}\left\{G_i=g\right\}. \end{aligned}$$

For these terms, we have the following convergence rates:

$$A_n = O_P(\sqrt{n^{-1}\xi_\mu^{-d_1+2} + \xi_\mu^2 + nh\xi_\mu^{2\kappa_\mu}}) + O_P(\sqrt{nh}[\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}]^2); \quad (28)$$

$$B_n = O_P(\sqrt{n^{-1}h^{-1}\xi_p^{-d_1+2} + h^{-1}\xi_p^2 + n\xi_p^{2\kappa_p}}) + O_P(n^{1/2}h^{-3/2}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2) \quad (29)$$

$$C_n = O_P(n^{1/2}h^{-1/2}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}][\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}]), \quad (30)$$

which are proved below. By expanding the right-hand sides of these, we can obtain the conclusion of the theorem.

The convergence rates of the term A_n in (28). Recall that if $W_i^c \notin S_\circ^c$, then $D_i = 0$ and $p(W_i) = 0$ (by Assumption 17). In this case, we have

$$K_h(p(W_i) - z_0) = h^{-1}K(-z_0/h) = 0 \text{ for } h(>0) \text{ small enough}, \quad (31)$$

since the support of K is bounded and $-z_0/h(= -1/2h)$ is large enough. Therefore, for h small enough, we can only restrict our attention to the case $W_i^c \in S_\circ^c$, and thus obtain

$$\begin{aligned} A_n &= (\sqrt{nh}/n)\sum_{i=1}^n K_h(p(W_i) - z_0)\left[\hat{\mu}_{-i}(W_i) - \mu(W_i)\right]\mathbf{1}\{G_i = g\}\mathbf{1}\{W_i^c \in S_\circ^c\} \\ &= (\sqrt{nh}/n^2)\sum_{i=1}^n \sum_{1\le j\le n;\, j\neq i} a_n(i,j) \\ &\quad +(\sqrt{nh}/n)\sum_{i=1}^n K_h(p(W_i) - z_0)\mathbf{1}\{W_i \in S_\circ^c\}\mathbf{1}\{G_i = g\} \\ &\quad \times O_P([\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}]^2), \end{aligned} \quad (32)$$

where the first equality holds by (22) of Lemma 3 with

$$\begin{aligned} a_n(i,j) \;:\; &= K_h(p(W_i) - z_0)M_{\xi_\mu}\left(W_j^c - W_i^c\right)\mathbf{1}\left\{W_j^d = W_i^d\right\}D_j\left[\mu(W_j) - \mu(W_i)\right] \\ &\quad \times f^{-1}(W_i)p^{-1}(W_i)\mathbf{1}\{G_i = g\}\mathbf{1}\{W_i^c \in S_\circ^c\}. \end{aligned}$$

We can show that

$$E\left[\left|\sum_{i=1}^n \sum_{1\le j\le n;\, j\neq i} a_n(i,j)\right|^2\right] = O\left(n^2h^{-1}\xi_\mu^{-d_1+2} + n^3h^{-1}\xi_\mu^2 + n^4\xi_\mu^{2\kappa_\mu}\right); \quad (33)$$

$$n^{-1}\sum_{i=1}^n K_h(p(W_i) - z_0)\mathbf{1}\{W_i \in S_\circ^c\} = O_p(1), \quad (34)$$

where the former result is proved below and the latter follows from standard arguments for kernel-based estimators. By (32), (33) and (34), we obtain (28) as desired. It remains to show (33).

Proof of (33). To derive the moment bound in (33). Look at

$$\begin{aligned}
&E\left[\left|\sum_{i=1}^{n}\sum_{1\leq j\leq n;\ j\neq i}a_n(i,j)\right|^2\right] = n(n-1)\left\{E\left[a_n^2(1,2)\right]+E\left[a_n(1,2)\,a_n(2,1)\right]\right\}\\
&+n(n-1)(n-2)\left\{E\left[a_n(1,2)\,a_n(1,3)\right]+E\left[a_n(1,2)\,a_n(3,1)\right]+E\left[a_n(1,2)\,a_n(2,3)\right]\right.\\
&\left.+E\left[a_n(1,2)\,a_n(3,2)\right]\right\}\\
&+n(n-1)(n-2)(n-3)\,E\left[a_n(1,2)\right]E\left[a_n(3,4)\right]. \qquad (35)
\end{aligned}$$

For the components in the first terms on the RHS of (35), we have

$$E\left[a_n^2(1,2)\right]=O(h^{-1}\xi_\mu^{-d_1+2}) \ \text{ and } \ E\left[a_n(1,2)\,a_n(2,1)\right]=O(h^{-1}\xi_\mu^{-d_1+2}), \qquad (36)$$

These two results can be shown analogously, and we only consider the former:

$$\begin{aligned}
&E\left[\left|a_n(1,2)\right|^2\right]\\
\leq\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{2d_1}}E\left[K^2\left(\tfrac{p(W_1)-z_0}{h}\right)M^2\left(\tfrac{W_2^c-W_1^c}{\xi_\mu}\right)\mathbf{1}\left\{W_2^d=W_1^d\right\}\left[\mu(W_2)-\mu(W_1)\right]^2\right]\\
=\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{2d_1}}\sum_{w_1^d\in S^d}\int_{w_1^c\in S^c}\int_{w_2^c\in S^c}K^2\left(\tfrac{p\left(w_1^c,w_1^d\right)-z_0}{h}\right)M^2\left(\tfrac{w_2^c-w_1^c}{\xi_\mu}\right)\\
&\times\left[\mu\left(w_2^c,w_1^d\right)-\mu\left(w_1^c,w_1^d\right)\right]^2\times f\left(w_1^c,w_1^d\right)f\left(w_2^c,w_1^d\right)dw_1^cdw_2^c\\
\leq\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{d_1}}\sum_{w_1^d\in S^d}\int_{w_1^c\in S^c}\int_{r_2\in\left\{r|\ w_1^c+\xi_p r\in S^c\right\}}K^2\left(\tfrac{p\left(w_1^c,w_1^d\right)-z_0}{h}\right)C_3^2M^2(r_2)\left|\left|r_2\right|\right|^2\xi_\mu^2\\
&\times f\left(w_1^c,w_1^d\right)f\left(w_1^c+\xi_\mu r_2,w_2^d\right)dw_1^cdr_2\\
\leq\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{d_1-2}}\sum_{w_1^d\in S^d}\int_{w_1^c\in S^c}K^2\left(\tfrac{p\left(w_1^c,w_1^d\right)-z_0}{h}\right)f\left(w_1^c,w_1^d\right)dw_1^c\times C_3^2C_4\int_{S_L}M^2(r_2)\left|\left|r_2\right|\right|^2dr_2\\
=\ & O(h^{-1}\xi_\mu^{-d_1+2}), \qquad (37)
\end{aligned}$$

where the first inequality holds since $\inf_{\left(w^c,w^d\right)\in S^c\times S^d}f(w)\,p(w)\geq C_1C_2$ and $|D_2|\leq 1$ (C_1 and C_2 are constants given in Assumption 17); the inequality in the fifth line uses the change-of-variable argument and the mean-value theorem with $r_2=\left(w_2^c-w_1^c\right)/\xi_\mu$,

$$\left|\mu\left(w_1^c+r\xi_\mu,w_2^d\right)-\mu\left(w_1^c,w_1^d\right)\right|=\left|\left\langle\xi_p r,\left(\partial/\partial w^c\right)f\left(\tilde{w}_1^c,w_1^d\right)\right\rangle\right|\leq C_3\left|\left|r\right|\right|\xi_p,$$

$C_3:=\sup_{\left(w^c,w^d\right)\in S^c\times S^d}\left|\left|\left(\partial/\partial w^c\right)f\left(w_1^c,w_1^d\right)\right|\right|$; and the inequality in the seventh line holds with $C_4:=\sup_{\left(w^c,w^d\right)\in S^c\times S^d}f\left(w_1^c,w_1^d\right)$ (C_3 and C_4 are finite by Assumption 14); and the last equality holds since

$$\sum_{w_1^d\in S^d}\int_{w_1^c\in S^c}K^2\left(\frac{p\left(w_1^c,w_1^d\right)-z_0}{h}\right)f\left(w_1^c,w_1^d\right)dw_1^c=\int_0^1K^2\left(\frac{q-z_0}{h}\right)f_p(q)\,dq=O(h),$$

where $f_p(q) := \nu(q, f) + \nu(q, m)$.

For the components in the second term on the RHS of (35), it holds that

$$\begin{aligned} &E[a_n(1,2)\, a_n(1,3)] = O\left(h^{-1}\xi_\mu^2\right); \quad E[a_n(1,2)\, a_n(3,1)] = O\left(h^{-1}\xi_\mu^2\right); \\ &E[a_n(1,2)\, a_n(2,3)] = O\left(h^{-1}\xi_\mu^2\right); \quad E[a_n(1,2)\, a_n(3,2)] = O\left(h^{-1}\xi_\mu^2\right). \end{aligned} \tag{38}$$

These four results can be shown in almost the same manner, and we only prove the last one. Similarly to (37),

$$\begin{aligned} & E[a_n(1,2)\, a_n(3,2)] \\ \leq\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{2d_1}} E\left[K\left(\tfrac{p(W_1)-z_0}{h}\right)\left|M\left(\tfrac{W_2^c-W_1^c}{\xi_\mu}\right)\right| \mathbf{1}\left\{W_2^d = W_1^d = W_2^d\right\} |\mu(W_2) - \mu(W_1)| \right. \\ & \left. \times K\left(\tfrac{p(W_3)-z_0}{h}\right)\left|M\left(\tfrac{W_2^c-W_3^c}{\xi_\mu}\right)\right| |\mu(W_2) - \mu(W_3)|\right] \\ =\ & \frac{(C_1C_2)^{-2}}{h^2\xi_\mu^{2d_1}} \sum_{w_1^d\in S^d} \int_{w_1^c\in S^c} \int_{w_2^c\in S^c} \int_{w_3^c\in S^c} K\left(\tfrac{p(w_1^c)-z_0}{h}\right)\left|M\left(\tfrac{w_2^c-w_3^c}{\xi_\mu} - \tfrac{w_1^c-w_3^c}{\xi_\mu}\right)\right| \left|\mu\left(w_2^c, w_1^d\right) - \mu\left(w_1^c, w_1^d\right)\right| \\ & \times K\left(\tfrac{p(w_3^c)-z_0}{h}\right)\left|M\left(\tfrac{w_2^c-w_3^c}{\xi_\mu}\right)\right| \left|\mu\left(w_2^c, w_1^d\right) - \mu\left(w_3^c, w_1^d\right)\right| f\left(w_1^c, w_1^d\right) f\left(w_2^c, w_1^d\right) f\left(w_3^c, w_1^d\right) dw_1^c dw_2^c dw_3^c \\ \leq\ & \frac{(C_1C_2)^{-2}\,\bar{K}}{h^2} \sum_{w_1^d\in S^d} \int_{r_1\in\left\{r|w_3^c+\xi_p r\in S^c\right\}} \int_{r_2\in\left\{r|w_3^c+\xi_p r\in S^c\right\}} \int_{w_3^c\in S^c} M(r_2 - r_1) \\ & \times \left|\mu\left(w_3^c + \xi_\mu r_2, w_1^d\right) - \mu\left(w_3^c + \xi_\mu r_1, w_1^d\right)\right| K\left(\tfrac{p(w_3^c)-z_0}{h}\right) |M(r_2)| \left|\mu\left(w_3^c + \xi_\mu r_2, w_1^d\right) - \mu\left(w_3, w_1^d\right)\right| \\ & \times f\left(w_3^c + \xi_\mu r_1, w_1^d\right) f\left(w_3^c + \xi_\mu r_2, w_1^d\right) f\left(w_3^c, w_1^d\right) dw_1^c dw_2^c dw_3^c \\ \leq\ & \frac{(C_1C_2)^{-2}\,\bar{K}\xi_\mu^2}{h^2} \sum_{w_3^d\in S^d} \int_{w_3^c\in S^c} K\left(\tfrac{p(w_3^c)-z_0}{h}\right) f\left(w_3^c, w_1^d\right) dw_3^c \\ & \times C_3^2 C_4^2 \int_{S_L}\int_{S_L} (||r_2|| + ||r_1||)\, ||r_2||\, |M(r_2 - r_1)|\, |M(r_2)|\, dr_2 dr_3 \\ =\ & O\left(\xi_\mu^2 h^{-1}\right). \end{aligned}$$

For the component of the third term on the RHS of (35), we have

$$E[a_n(1,2)\, a_n(3,4)] = \{E[a_n(1,2)]\}^2 = O(\xi_\mu^{2\kappa_\mu}), \tag{39}$$

This can be shown by arguments similar to those for (37), and we only outline main points. Letting $S_g^d := \left\{w^d \in S^d |\ d_2\text{-th element of } w^d \text{ is } g\right\}$, it holds that

$$\begin{aligned} & E\left[a_n\left(1,2\right)\right] = \frac{1}{h\xi_\mu^{d_1}} \sum_{w_1^d \in S_g^d} \int_{w_1^c \in S_\circ^c} \int_{w_2^c \in S^c} K\left(\frac{p\left(w_1^c, w_1^d\right) - z_0}{h}\right) M\left(\frac{w_2^c - w_1^c}{\xi_\mu}\right) \\ & \times \frac{p\left(w_2^c, w_1^d\right)\left[\mu\left(w_2^c, w_1^d\right) - \mu\left(w_1^c, w_1^d\right)\right]}{p\left(w_1^c, w_1^d\right)} f\left(w_2^c, w_1^d\right) dw_1^c dw_2^c \\ = & \frac{1}{h} \sum_{w_1^d \in S_g^d} \int_{w_1^c \in S_\circ^c} \int_{r_2 \in \left\{r |\ w_1^c + \xi_p r \in S^c\right\}} K\left(\frac{p\left(w_1^c, w_1^d\right) - z_0}{h}\right) M\left(r_2\right) \\ & \times \frac{p\left(w_1^c + \xi_\mu r_2, w_1^d\right)}{p\left(w_1^c, w_1^d\right)} \left[\mu\left(w_1^c + \xi_\mu r_2, w_1^d\right) - \mu\left(w_1^c, w_1^d\right)\right] f\left(w_1^c + \xi_\mu r_2, w_1^d\right) dw_1^c dr_2. \end{aligned}$$

Then, by standard arguments for the kernel estimators (with applying Taylor expansion to $\mu\left(w_1^c + \xi_\mu r_2, w_1^d\right)$ and $f\left(w_1^c + \xi_\mu r_2, w_1^d\right)$ and by changing variables involving h), we have $E\left[Ea_n\left(1,2\right)\right] = \xi_\mu^{\kappa_\mu}$. Putting (35), (36), (38) and (39) together, we now obtain (33).

The convergence rate of B_n in (29). Applying the Taylor expansion to $K_h\left(\hat{p}\left(X_i, g\right) - z\right) - K_h\left(p\left(X_i, g\right) - z\right)$, we have

$$B_n = B_{n,1} + B_{n,2},$$

where

$$\begin{aligned} B_{n,1} &= n^{1/2} h^{-1/2} \times \frac{1}{n} \sum\nolimits_{i=1}^n \frac{1}{h} K'\left(\frac{p\left(W_i\right) - z}{h}\right)\left[\hat{p}\left(W_i\right) - p\left(W_i\right)\right] \mu\left(W_i\right); \\ B_{n,2} &= n^{1/2} h^{-3/2} \times \frac{1}{2n} \sum\nolimits_{i=1}^n \frac{1}{h} K''\left(\frac{\check{p}_{-i}\left(W_i\right) - z}{h}\right)\left[\hat{p}\left(W_i\right) - p\left(W_i\right)\right]^2 \mu\left(W_i\right); \end{aligned}$$

and $\check{p}_{-i}\left(W_i\right)$ is on the line segment connecting $\hat{p}_{-i}\left(W_i\right)$ to $p\left(W_i\right)$. We derive the convergence rates of these terms.

To derive the convergence rate of $B_{n,1}$, we use (20) in Lemma 2. Then, we have

$$\begin{aligned} B_{n,1} &= n^{1/2} h^{-1/2} \times \frac{1}{n^2} \sum\nolimits_{i=1}^n \sum\nolimits_{1 \le j \le n;\ j \ne i} b_n\left(i,j\right) \\ & \underbrace{+ n^{1/2} h^{-1/2} \times \frac{1}{n} \sum\nolimits_{i=1}^n \frac{1}{h} K'\left(\frac{p\left(W_i\right) - z}{h}\right) \mu\left(W_i\right) \times O_{a.s.}(\left[\xi_p^{\kappa_p} + \sqrt{\left(\log n\right)/n\xi_p^{d_1}}\right]^2),}_{= O_P(n^{1/2} h^{-1/2}\left[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}\right]^2)} \end{aligned}$$

where

$$b_n\left(i,j\right) := \frac{1}{h} K'\left(\frac{p\left(W_i\right) - z}{h}\right) L_{\xi_p}\left(W_j^c - W_i^c\right) \mathbf{1}\left\{W_j^d = W_i^d\right\}\left[p\left(W_j\right) - p\left(W_i\right)\right] \mu\left(W_i\right) / f\left(W_i\right),$$

and the rate of the second term on the RHS follows from the same arguments as before. As for the first term, we have

$$E\left[\left|\sum_{i=1}^{n}\sum_{1\leq j\leq n;\ j\neq i} b_n(i,j)\right|^2\right] = O(n^2h^{-1}\xi_p^{-d_1+2} + n^3h^{-1}\xi_p^2 + n^4\xi_p^{2\kappa_p}),$$

whose proof is almost the same as that of (33), and is omitted. Given this, we have

$$B_{n,1} = O_p(\sqrt{n^{-1}h^{-1}\xi_p^{-d_1+2} + h^{-1}\xi_p^2 + n\xi_p^{2\kappa_p}}) + O_P(n^{1/2}h^{-1/2}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2). \quad (40)$$

The convergence of $B_{n,2}$ can be shown in the same manner as for $J_{n,2}$. That is, we can find some function $\mathcal{K}^{**}(\cdot)$ and some positive constant $\bar{\epsilon}$ such that $\sup_{|\epsilon|\leq\bar{\epsilon}} K''(u+\epsilon) \leq \mathcal{K}^{**}(u)$ for any $u \in \mathbb{R}$, $\sup_{u\in\mathbb{R}} |\mathcal{K}^{**}(u)| < \infty$ and $\int_{\mathbb{R}} |\mathcal{K}^{**}(u)|\, du < \infty$. Then,

$$\begin{aligned} |B_{n,2}| &\leq n^{1/2}h^{-3/2} \times \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h}\mathcal{K}^{**}\left(\frac{p(W_i)-z}{h}\right) \\ &\quad \times \max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(w) - p(w)|^2 \times \sup_{w\in S^c\times S^d}|\mu(w)| \\ &= O_P(n^{1/2}h^{-3/2}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2), \end{aligned} \quad (41)$$

where the first equality holds almost surely for n large enough; and the equality uses the boundedness of μ and (21) of Lemma 2. Now, (40) and (41) imply the desired result (29).

The convergence rate of C_n in (30). Let $\varepsilon\,(>0)$ any (small) constant. Then, by (21) of Lemma 2, for $\omega \in \Omega^*$ such that $\Pr(\Omega^*) = 1$, there exists some $\bar{n}$ such that for any $n \geq \bar{n}$,

$$\max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(w) - p(w)| \leq \varepsilon.$$

In this case, if $W_i^c \notin S_\circ^c$, $p(W_i) = 0$ and thus $\max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(W_i)| \leq \varepsilon$. Therefore, for h small enough, $[\hat{p}_{-i}(W_i) - z_0]/h$ is large enough and $K_h(\hat{p}_{-i}(W_i) - z_0) = 0$. This, together with (31), implies that if $W_i \notin S_\circ^c$ and n is large enough (with h small enough),

$$[K_h(\hat{p}_{-i}(W_i) - z_0) - K_h(p(W_i) - z_0)]\left[\hat{\mu}_{-i}(W_i) - \mu(W_i)\right] = 0.$$

Thus, for deriving the upper bound of C_n, it is sufficient to consider only the case where $W_i \in S_\circ^c$ and the following expression is valid:

$$\begin{aligned} |C_n| &\leq \left(n^{1/2}h^{-1/2}\right) \times \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathcal{K}^*\left(\frac{p(W_i)-z_0}{h}\right) \\ &\quad \times \max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(w) - p(w)| \times \max_{i\in\{1,\ldots,n\}}\sup_{w\in S_\circ^c\times S^d}\left|\hat{\mu}_{-i}(w) - \mu(w)\right|, \end{aligned}$$

where $\mathcal{K}^*$ is the function used for deriving the bound of $J_{n,2}$ in (25). By (21), (23) and (26), we now obtain the desired result (30). ■

It remains to prove two auxiliary lemmas:

o. f of Lemma (2) Let

$$\begin{aligned}
\hat{f}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq k\leq n;\ k\neq i} L_{\xi_p}(W_k^c - w^c)\,\mathbf{1}\left\{W_k^d = w^d\right\};\\
\hat{\Gamma}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq j\leq n;\ j\neq i} L_{\xi_p}(W_j^c - w^c)\,\mathbf{1}\left\{W_j^d = w^d\right\}[p(W_j) - p(w)].\\
\hat{H}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq j\leq n;\ j\neq i} L_{\xi_p}(W_j^c - w^c)\,\mathbf{1}\left\{W_j^d = w^d\right\}[D_j - p(W_j)].
\end{aligned}$$

Then, for each i, we can write

$$\hat{p}_{-i}(w) - p(w) = \left[\hat{\Gamma}_{-i}(w) + \hat{H}_{-i}(w)\right] \times \left[\frac{1}{f(w)} + \frac{f(w) - \hat{f}_{-i}(w)}{\hat{f}_{-i}(w) f(w)}\right]. \tag{42}$$

For the components on the RHS of (42), we can show the following convergence results:

$$\begin{aligned}
\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S_\circ^c\times S^d} \left|\hat{f}_{-i}(w) - f(w)\right| &= O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}); &(43)\\
\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S^c\times S^d} \left|\hat{f}_{-i}(w) - f(w)\right| &= O_{a.s.}(\xi_p + \sqrt{(\log n)/n\xi_p^{d_1}}); &(44)\\
\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S^c\times S^d} \left|\hat{\Gamma}_{-i}(w)\right| &= O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}); &(45)\\
\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S^c\times S^d} \left|\hat{H}_{-i}(w)\right| &= O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}}), &(46)
\end{aligned}$$

whose proofs are provided below. Now, fix any $\omega\in\Omega^*$, where Ω^* is an event with $\Pr(\Omega^*) = 1$. Then, (44) implies that as $n\to\infty$, $\max_{i\in\{1,\ldots,n\}} \sup_{w\in S^c\times S^d} \left|\hat{f}_{-i}(w) - f(w)\right| < C_1/2$. In this case, since $\inf_{(w^c,w^d)\in S^c\times S^d} f(w) \geq C_1$ (C_1 is the constant given in Assumption 17), it holds that

$$\begin{aligned}
&\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S^c\times S^d} 1/\hat{f}_{-i}(w)\\
= \ &\max\nolimits_{i\in\{1,\ldots,n\}} \sup\nolimits_{w\in S^c\times S^d} 1/\left[f(w) + \hat{f}_{-i}(w) - f(w)\right] \leq 1/(C_1/2),
\end{aligned}$$

implying that

$$1/\hat{f}_{-i}(w) = O_{a.s.}(1), \ \text{ uniformly over } i\in\{1,\ldots,n\} \text{ and } w\in S^c\times S^d. \tag{47}$$

Given these, we have the following expression:

$$\hat{p}_{-i}(w) - p(w) = \hat{\Gamma}_{-i}(w)/f(w) + O_{a.s.}([\xi_p + \sqrt{(\log n)/n\xi_p^{d_1}}]^2) + O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}})$$

Then, this leads to the first result (20). By (43), we have the second result (21). It remains to show show (43), (44), (45) and (46).

Proofs of (43) and (44). Letting

$$\hat{f}(w) := n^{-1}\sum\nolimits_{1\leq k\leq n} L_{\xi_p}(W_k^c - w^c)\,\mathbf{1}\left\{W_k^d = w^d\right\},$$

we have the following decomposition:

$$\begin{aligned}\hat{f}_{-i}(w) - f(w) &= \hat{f}(w) - f(w) - \left(n\xi_p^{d_1}\right)^{-1} L\left(\frac{W_i^c - w^c}{\xi_p}\right)\mathbf{1}\left\{W_i^d = w^d\right\} \\ &= \left\{\hat{f}(w) - E[\hat{f}(w)]\right\} + \left\{E[\hat{f}(w)] - f(w)\right\} + O(1/n\xi_p), \end{aligned} \tag{48}$$

where the last equality holds uniformly over $i \in \{1, \ldots, n\}$ and $w \in S^c \times S^d$ by the boundedness of the kernel function L. By applying Theorem 3 of Hansen (2008), we can show that the first term on the RHS of (48) is $O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}})$ uniformly over $w \in S^c \times S^d$. As for the second term, noting that $S_\circ^c$ is strictly in the interior of S^c, we can also show that

$$\sup_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} \left|E[\hat{f}(w)] - f(w)\right| = O(\xi_p^{\kappa_p}),$$

which follows from standard arguments for biases of kernel-based estimators, say change-of-variable and Taylor-approximation arguments with Assumption 15 (see, e.g., proofs of Theorems 6 and 8 in Hansen, 2008). This implies the desired result (43). Next, if we let the domain of w^c as the whole set S^c, we have

$$\begin{aligned} & \sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|E[\hat{f}(w)] - f(w)\right| \\ = & \sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|\sum\nolimits_{u^d \in S^d} \xi_p^{-d_1} \int_{u^c \in S^c} L\left(\frac{u^c - w^c}{\xi_p}\right)\mathbf{1}\left\{u^d = w^d\right\} f\left(u^c, u^d\right) du^c - f(w)\right| \\ = & \sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|\int_{v^c \in T^c\left(w^c, \xi_p\right)} L(v^c)\left[f\left(w^c + \xi_p v^c, w^d\right) - f\left(w^c, w^d\right)\right] dv^c\right| \\ = & \sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|\int_{v^c \in T^c\left(w^c, \xi_p\right)} L(v^c)\left\langle \xi_p v^c, \left(\partial/\partial w^c\right) f\left(\tilde{w}^c, w^d\right)\right\rangle dv^c\right| \\ \leq & \xi_p \int_{v^c \in S^L} |L(v^c)|\, ||v^c||\, dv^c \times \sup_{\left(w^c, w^d\right) \in S^c \times S^d} \left|\left|\left(\partial/\partial w^c\right) f\left(w^c, w^d\right)\right|\right| = O\left(\xi_p\right) \end{aligned} \tag{49}$$

where $T^c\left(w^c, \xi_p\right) := \left\{v |\ w^c + \xi_p v \in S^c\right\}$; $\langle a, b\rangle$ is the inner product of vectors a and b; $\tilde{w}^c$ is on the line segment connecting w^c and $w^c + \xi_p v^c$; the second equality holds by changing variables with $\left(u^c - w^c\right)/\xi_p = v^c$; and the third equality uses the mean-value theorem; and the inequality uses the fact that $T^c\left(w^c, \xi_p\right) \supset S^L$ (uniformly) over any $w^c \in S^c$ for ξ_p is small enough (note S_L is the support of L, and S^L and S^c are compact). Now, we can see that the above arguments and (49) implies the desired result (44).

Proof of (45). We write

$$\hat{\Gamma}(w) := n^{-1} \sum\nolimits_{1 \leq j \leq n} L_{\xi_p}\left(W_j^c - w^c\right)\mathbf{1}\left\{W_j^d = w^d\right\}\left[D_j - p(w)\right].$$

By the same arguments as for (48), it holds that

$$\hat{\Gamma}_{-i}(w) = O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}}) + E[\hat{\Gamma}(w)],$$

uniformly over $i \in \{1,\ldots,n\}$ and $w \in S^c \times S^d$. To derive the bound of $E[\hat{\Gamma}(w)]$, find a set $S_+^c \left(\subset \mathbb{R}^{d_1}\right)$ satisfying the following conditions: (1) $S_\circ^c \subsetneq S_+^c \subsetneq S^c$; (2) all the boundary points of $S_\circ^c$ are in the interior of S_+^c, and all the boundary points of S_+^c are in the interior of S^c. By Assumption (17), such S_+^c exists. Let $N_+^c := \left\{u \in S^c |\ u \notin S_+^c\right\}$. Then, we look at the following bound:

$$\sup_{\left(w^c,w^d\right)\in S^c\times S^d} E[\hat{\Gamma}(w)] \leq \sup_{\left(w^c,w^d\right)\in S_+^c\times S^d} \left|E[\hat{\Gamma}(w)]\right| + \sup_{\left(w^c,w^d\right)\in N_+^c\times S^d} \left|E[\hat{\Gamma}(w)]\right|. \quad (50)$$

The first term on the RHS of (50) is $O\left(\xi_p^2\right)$. This can be shown by the standard arguments for the biases of kernel-based estimator (note that all the points of S_+^c are strictly in the interior of S^c). As for the second term on the RHS of (50), we see

$$\begin{aligned}
& \sup_{\left(w^c,w^d\right)\in N_+^c\times S^d} \left|E[\hat{\Gamma}(w)]\right| \\
= & \sup_{\left(w^c,w^d\right)\in N_+^c\times S^d} \left|\int_{u^c\in S^c} \frac{1}{\xi_p^{d_1}} L\left(\frac{u^c-w^c}{\xi_p}\right)\left[p\left(u^c,w^d\right)-p\left(w^c,w^d\right)\right] f\left(u^c,w^d\right) du^c\right| \\
= & \sup_{\left(w^c,w^d\right)\in N_+^c\times S^d} \left|\int_{v^c\in\left\{v|\ w^c+\xi_p v\in S^c;\ v\in S^L\right\}} L(v^c)\, p\left(w^c+\xi_p v^c,w^d\right) f\left(w^c+\xi_p v^c,w^d\right) dv^c\right| = 0
\end{aligned}$$

where the second equality holds since $p\left(w^c,w^d\right)=0$ for $\left(w^c,w^d\right)\in N_+^c\times S^d$, and the last equality holds for ξ_p small enough, since $p\left(w^c+\xi_p v^c,w^d\right)=0$ for such ξ_p, which follows from the fact that $w^c+\xi_p v^c \notin S_\circ^c$ for $w_c \in N_+^c$ and for any v^c, if ξ_p is small enough (we note that the support of L, S_L, is supposed to be bounded and $||v^c|| < C$ for some positive constant). Therefore, we have the desired result of (45).

Proof of (46). This result follows from arguments analogous above, and we omit details (note that each summand is a zero-mean random variable, and we apply Theorem 3 of Hansen, 2008). Now, the proof is completed. ■

o. of of Lemma 3] Let

$$\begin{aligned}
\hat{q}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq k\leq n;\ k\neq i} M_{\xi_\mu}\left(W_k^c-w^c\right)\mathbf{1}\left\{W_k^d=w^d\right\}D_k; \\
q(w) &: = f(w)\,p(w); \\
\hat{\Pi}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq j\leq n;\ j\neq i} M_{\xi_\mu}\left(W_j^c-w^c\right)\mathbf{1}\left\{W_j^d=w^d\right\}\left[\mu(W_j)\,p(W_j)-\mu(w)\,p(w)\right] \\
\hat{\Theta}_{-i}(w) &: = n^{-1}\sum\nolimits_{1\leq j\leq n;\ j\neq i} M_{\xi_\mu}\left(W_j^c-w^c\right)\mathbf{1}\left\{W_j^d=w^d\right\}D_j u_j
\end{aligned}$$

Then, for each $w^c \in S_\circ^c$, we can write

$$\hat{\mu}_{-i}(w) - \mu(w) = \left[\hat{\Pi}_{-i}(w) + \hat{\Theta}_{-i}(w)\right] \times \left[\frac{\mathbf{1}}{q(w)} + \frac{q(w) - \hat{q}_{-i}(w)}{\hat{q}_{-i}(w)\, q(w)}\right], \tag{51}$$

where we note $q(w) \geq C_1 C_3 (> 0)$ for any $w^c \in S_\circ^c$ (by Assumption 17).

We can show that

$$\begin{aligned}
\hat{q}_{-i}(w) - q(w) &= O_P(\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}); & (52)\\
\hat{\Pi}_{-i}(w) &= O_P(\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}); & (53)\\
\hat{\Theta}_{-i}(w) &= O_P(\sqrt{(\log n)/n\xi_\mu^{d_1}}); & (54)\\
1/\hat{q}_{-i}(w) &= O_P(1), & (55)
\end{aligned}$$

uniformly over $i \in \{1, \ldots, n\}$ and $w \in S_\circ^c \times S^d$, whose proofs are provided below. Analogously to the proof of Lemma 2, by (51)-(55), we obtain the desired results (22) and (22).

Proofs of (52), (53) and (54). The proofs of these three results use almost the same arguments, and they are analogous to the proof of (45) in Lemma 2). Thus, we only outline main points for the proof of (53). By letting

$$\hat{\Pi}(w) := n^{-1} \sum\nolimits_{1 \leq j \leq n} M_{\xi_\mu}\left(W_j^c - w^c\right) \mathbf{1}\left\{W_j^d = w^d\right\} D_j \left[\mu(W_j) - \mu(w)\right],$$

we can write

$$\hat{\Pi}_{-i}(w) = \left\{\hat{\Pi}(w) - E[\hat{\Pi}(w)]\right\} + E[\hat{\Pi}(w)] + O(1/n\xi_u),$$

uniformly over $i \in \{1, \ldots, n\}$ and $w \in S_\circ^c \times S^d$. Using Theorem 2 of Hansen (2008), we have

$$\sup_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} \left|\hat{\Pi}(w) - E[\hat{\Pi}(w)]\right| = O_P(\sqrt{(\log n)/n\xi_\mu^{d_1}}).$$

We can also show that $\sup_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} \left|E[\hat{\Pi}(w)]\right| = O\left(\xi_\mu^{\kappa_\mu}\right)$ by the same arguments as for $\sup_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} E[\hat{\Gamma}(w)]$, the first term on the RHS of (50) in Lemma 2). From these, we obtain the desired result (53).

Proof of (55). Let J be an arbitrary positive constant, and define an event E_n as

$$E_n := \left\{\max_{i \in \{1, \ldots, n\}} \sup_{w \in S_\circ^c \times S^d} |\hat{q}_{-i}(w) - q(w)| \leq C_1 C_2 / 2\right\},$$

where C_1 and C_2 are constants given in Assumption 17. Denote by E_n^C the complement event of E_n. Then,

$$\begin{aligned}
\Pr\left[\max_{i \in \{1, \ldots, n\}} \sup_{w \in S_\circ^c \times S^d} |1/\hat{q}_{-i}(w)| > J\right] &\leq \Pr\left[\max_{i \in \{1, \ldots, n\}} 1/\inf_{w \in S_\circ^c \times S^d} |\hat{q}_{-i}(w)| > J \,\middle|\, E_n\right] \Pr[E_n] + \Pr\left[E_n^C\right] \\
&\leq \Pr\left[C_1 C_3 / 2 > J \mid E_n^C\right] + \Pr\left[E_n^C\right], \qquad (56)
\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}
\inf_{w\in S_{\circ}^{c}\times S^{d}}\left|\hat{q}_{-i}\left(w\right)\right| &= \inf_{w\in S_{\circ}^{c}\times S^{d}}\left|q\left(w\right)+\left[\hat{q}_{-i}\left(w\right)-q\left(w\right)\right]\right| \\
&\geq \min\left\{\inf_{w\in S_{\circ}^{c}\times S^{d}}\left|q\left(w\right)\right|,\ \left(\inf_{w\in S_{\circ}^{c}\times S^{d}}\left|q\left(w\right)\right|-\sup_{w\in S_{\circ}^{c}\times S^{d}}\left|\hat{q}_{-i}\left(w\right)-q\left(w\right)\right|\right)\right\} \\
&> \min\left\{C_{1}C_{2},\ C_{1}C_{2}/2\right\}=C_{1}C_{2}/2,
\end{aligned}$$

uniformly over $i\in\{1,\ldots,n\}$, under the condition that $\sup_{w\in S_{\circ}^{c}\times S^{d}}\left|\hat{q}\left(w\right)-q\left(w\right)\right|\leq C_{1}C_{2}/2$. The first term on the RHS of (56) is zero for J large enough, and the second term $\Pr\left[E_{n}^{C}\right]$ converges to zero for any $J>0$ as $n\to\infty$ (by (52)). Therefore, $\Pr\left[\max_{i\in\{1,\ldots,n\}}\sup_{w\in S_{\circ}^{c}\times S^{d}}\left|1/\hat{q}_{-i}\left(w\right)\right|>J\right]\to 0$ for J large enough, which imply the desired result of (55). The proof of Lemma 3 is now completed.

■