ESRC CENTRE ON
SKILLS, KNOWLEDGE
AND ORGANISATIONAL
PERFORMANCE

Are University Admissions Academically Fair?

SKOPE Research Paper No. 115 January 2013

Debopam Bhattacharya, Shin Kanaya, Margaret Stevens

Department of Economics, University of Oxford

Editor's Foreword

SKOPE Publications

This series publishes the work of the members and associates of SKOPE. A formal editorial process ensures that standards of quality and objectivity are maintained.

**Orders for publications should be addressed to the SKOPE Secretary,
School of Social Sciences, Cardiff University, Glamorgan Building,
King Edward VII Avenue, Cardiff CF10 3WT
Research papers can be downloaded from the website:
www.skope.ox.ac.uk**

ISSN 1466-1535

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 admit students with the highest academic potential. If so, then the expected performance of the marginal admitted candidates — the admission threshold — should be equalized across socioeconomic groups. We show that such thresholds are nonparametrically identified from standard admissions data if unobserved officers' 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 these methods to admissions data for a large undergraduate programme at Oxford and using blindly-marked, first-year exam-performance as the outcome of interest, we find that the admission-threshold is about 3.7 percentage-points (0.6 standard-deviations) higher for males than females and about 1.7 percentage-points (0.3 standard-deviations) higher for private-school than state-school applicants. In contrast, average admission-rates are equal across gender and school-type, both before and after controlling for applicants' background characteristics.

Keywords: University admissions, academic fairness, economic efficiency, marginal admit, conditional median restriction, nonparametric identification.

1 Introduction

Background: Selective universities are frequently accused of biased admission practices which favour applicants from socially advantaged backgrounds and thus contribute to the perpetuation of socioeconomic inequality. 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.[1] Universities usually respond to such allegations by claiming to practice academically fair admissions, i.e., to admit students with the best academic potential, irrespective of their socioeconomic status. For example, Oxford claims to be "...committed to recruiting the academically 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".[2] [3] 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 applicant-level admission 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, in the tradition of Becker (1957) and closely corresponds to the notion of economic efficiency. 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 expectations are 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 constraints, c.f., Bhattacharya and Dupas (2012). 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 admits" is intertwined with the assignment process and often depends on variables not observable to an analyst (c.f., Heckman, 1998). Secondly, it is sometimes difficult to observe the relevant outcomes or calculate their expected values. An ex-

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

[2] 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/

[3] 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. The closest US equivalent would be admission to post-graduate academic programs.

ample is the case of hiring workers, where it is difficult for an analyst 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 tutors 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 the applicant covariates. Defining and identifying the "overall" marginal candidates in presence of such *unobserved* treater-heterogeneity is a nontrivial problem – an issue that seems not to have been discussed previously in the literature.

In the university-application case, however, the first problem is mitigated to a large extent when the analyst can access the same application forms and standardized test-scores as those used by the admission tutors. For example, an economist studying admissions in her own university can easily access these data, especially if she herself is involved in conducting admissions.[4] 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.[5] Therefore in this case, access to applicant records largely eliminates the unobserved applicant characteristics issue that plagues studies of unfair protocols in some other situations, such as medical treatment, where patients are treated sequentially and individually and different *criteria* may be 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 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, Nelson and Vytlacil, 2011, discussed below) and the issuance of insurance coverage.[6]

A second advantage of the admission case is that one can easily match pre-application records

[4] For instance, the first and third authors of the present paper have served as admission tutors at Oxford. During their tenure, they could access the entire admission data for all subjects at the undergraduate level. Such access is also known to be feasible in other universities, c.f., Arcidiacono et al. (2011), Bertrand, Hanna and Mullainathan (2010), etc.

[5] For example, in our empirical application reported below, the regression of getting an admission offer on the set of commonly observed covariates yields a value of 50% for McFadden's pseudo-R^2 for a probit model and an R^2 of 45% for a linear probability model. These magnitudes are about ten to hundred times higher than goodness-of-fit measures typically reported by applied researchers for cross-sectional regressions – either linear or probit/logit.

[6] One other scenario where the analyst and the decision-makers observe the same set of applicant characteristics is the experimental set-up in Bertrand and Mullainathan (2004). They, however, focus on a notion of fairness which is different from our outcome-oriented approach (see the discussion just before Assumption 2).

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 tutors, 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 tutors 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 which is held after the admission process. 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 efficiency can then be carried out by checking equality of the identified thresholds across the groups. Our key behavioral assumptions are that (a) admission tutors 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 tutors' subjective expectations and the true mathematical expectations, have zero median – i.e., the errors are equally likely to be positive or negative.[7] It is important to note that the latter, "rational expectations"-type 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 a tutor's subjective beliefs regarding those types of individuals with whom the tutor has had less experience.

As a final step in our analysis, we apply our identification (and corresponding inferential) methods 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 large at about 0.6 standard deviations and that for school-types about 0.3 standard deviations of the outcome. This finding is suggestive of some degree of affirmative action – either explicit or implicit – within the admission process, which is not apparent from the equal

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

success rates, thereby illustrating the usefulness of our approach.

Related literature: The present paper is substantively related to three broad research areas – (i) the econometric literature on treatment effects and treatment assignment, (ii) the evaluation of university admission procedures in education and educational sociology, especially with regard to social mobility and (iii) the economic analysis of affirmative-action in university admissions.

In regards to treatment effect analysis, our paper complements a recent literature in econometrics – pioneered by Manski (2004) – on the reverse problem of how treatment should be targeted for future populations, using information from past treatment outcomes.[8] Much of this literature assumes existence of trial data on treatment efficacy, which is difficult to obtain in the admissions context. But the more important distinction is that here we are trying to evaluate the current admission practice rather than proposing an "optimal" admission protocol. The latter is the goal of the treatment assignment literature.

In the education literature, a large number of papers have been written about various aspects of admission to elite colleges and universities, largely focusing on the United States. For a broad, historical perspective on selectivity in US college admission, see Hoxby (2009). However, we are not aware of any previous attempt in the academic literature in education, economics or applied statistics to formally test *outcome-based* efficiency of such admissions. Some prior studies by educational sociologists attempt to test fairness by comparing the aggregate or covariate-conditioned fraction of applicants who were offered admission in each socio-demographic group. See for example Zimdars (2010) or Zimdars et al. (2009) and the references therein. A key contribution of the present paper is to shift the focus of analysis to the eventual outcomes of the students and thereby show that equal success rate in admissions across demographic groups can be consistent with very different admission standards across these different groups. Indeed, that is precisely what we find in our empirical application. A further point is that here our analysis focuses on the expected outcome of the *marginal* admits in different demographic groups. This is in contrast to many other studies – both academic and policy-oriented – which compare the *average* pre-admission test-scores (c.f., Kane, 1998) or *average* post-admission performance of admitted students (c.f., Keith et al., 1985) among different socioeconomic groups. The need to focus on the "marginal" rather than the average treatment recipient in a discussion of fair treatment was previously emphasized by Heckman (1998) and that is the approach we take in the present paper.

Given our focus on the marginal admits, the substantively closest work to ours is Bertrand, Hanna and Mullainathan (2010), who examine the consequences of affirmative action in admission to Indian engineering colleges on the marginal graduates' earnings. In their context, admission is based on a single exam score and admission thresholds differ by applicants' social caste. These thresholds are fixed and publicly known, thereby removing a key empirical challenge – that of defining and identifying the marginal admits and rejects – arising in general admissions contexts

[8] Stoye (2010), Hirano and Porter (2009) and Tetenov (2011) have more recently extended this line of research.

where entrance is based on several background variables and there is heterogeneity across admission tutors. Our methodology is designed to deal with these more general scenarios.

It may also be noted that our work is complementary to a large volume of research in the education literature on the usefulness of standardized test scores such as the SAT in the US 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 Institute of Education in the UK and the College Board in the USA. Rothstein (2004) provides a critical review of this line of research. Indeed, the purported aim of this literature is to inform the question of how to select applicants – i.e., the reverse of the question addressed in the present paper which is related to how students are, in fact, currently being admitted.

On the economic front, our paper complements an existing literature on analyzing the *consequences* of affirmative actions in college admissions. Fryer and Loury (2005) provide a critical review of this theoretical literature and a comprehensive bibliography. A survey of the theoretical literature on profiling in more general situations is Fang and Moro (2008). On the empirical side, Arcidiacono (2005) uses a structural model of admissions to simulate the potential, counterfactual consequences of removing affirmative action in US college admission and financial aid. In a different project, Arcidiacono, Aucejo, Fang and Spenner (2011) use admissions data from Duke University to empirically investigate the possibility that intended beneficiaries of affirmative action are on average hurt by its presence due to quality mismatch. In a related paper, Arcidiacono, Aucejo and Spenner (2011) investigate the consequence of affirmative action for major choice at Duke. Card and Krueger (2005) investigate the realized impact of eliminating affirmative action on minority students' application behavior. In contrast to these works, the present paper may be viewed as one that attempts to detect the *presence* of affirmative action type policies from admissions-related data. In section 3.2 below, we contrast our identification strategy with those that have been used to detect unfair treatment in law enforcement and healthcare where, however, the empirical settings are in fact quite different from the college admissions scenario.

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 and contrasts our approach with alternative identification strategies in the empirical microeconomics literature. Section 4 deals with inference. Section 5 contains the substantive application of our methods to the case of admission to a large undergraduate programme at Oxford University. Section 6 concludes. All technical proofs are collected in the appendix.

2 Benchmark Model

We start our analysis by laying out a benchmark economic model of admissions to help fix ideas. Based on this economic model, in the next section we develop a corresponding econometric model incorporating unobserved heterogeneity, which can be used to analyze admissions data.

Let W denote an applicant's pre-admission characteristics, observed by the university. We let $W := (X, G)$, where G denotes 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 commonly alleged mistreatment. The variables in X are the applicant's other characteristics observed prior to admission which include one or more continuously distributed components like standardized test-scores. Also, let Y denote the applicant's future academic performance if admitted to the university (assumed to take on non-negative values, e.g., GPA), and the binary indicator D denote whether the applicant received an admission offer and the binary indicator A denote whether the admission offer was accepted by the applicant.

Let $\mathcal{W}$ denote the support of the random vector W, $F_W(\cdot)$ denote the marginal cumulative distribution function (C.D.F.) of W and $\mu^*(w)$ denote a w-type student's expected performance ($w \in \mathcal{W}$) if he/she enrols, and let $\alpha(w)$ denote the probability that a w-type student upon being offered admission eventually enrols.

Let $c \in (0,1)$ be a constant denoting the maximum fraction of applicants who can be admitted, given the number of available spaces.

Admission protocols: We can define an admission protocol as a probability $p(\cdot) : \mathcal{W} \to [0,1]$ such that an applicant with characteristics w is offered admission with probability $p(w)$. A generic objective of the university may be described as

$$\sup_{p(\cdot)\in\mathcal{F}} \int_{w\in\mathcal{W}} p(w)\, h(w)\, \alpha(w)\, \mu^*(w)\, dF_W(w), \quad \text{subject to} \quad \int_{w\in\mathcal{W}} p(w)\, \alpha(w)\, dF_W(w) \leq c.$$

Here, $\mathcal{F}(= \mathcal{F}(\mathcal{W}, [0,1]))$ denotes a set of $[0,1]$-valued functions on $\mathcal{W}$, and $h(w)$ denotes a non-negative welfare weight, capturing how much the outcome of a w-type applicant is worth to the university. For affirmative action policies, $h(\cdot)$ will be larger for applicants from disadvantaged socioeconomic backgrounds or under-represented demographic groups. The overall objective is thus to maximize mean welfare-weighted outcome among the admitted applicants, subject to a budget constraint. The solution to the above problem takes the form described below in Proposition 1, which holds under the following condition:

Condition (C): $h(w) > 0$ and $\alpha(w) > 0$ for any $w \in \mathcal{W}$.[9] Further, for some $\delta > 0$,

$$\int_{w\in\mathcal{W}} \alpha(w)\, \mathbf{1}\{\mu^*(w) \geq 0\}\, dF_W(w) \geq c + \delta,$$

[9] This assumption is innocuous in the sense that those w for which $\alpha(w)$ is zero will not contribute to either the objective function or the constraint. We can simply redefine $\mathcal{W}$ to be the subset of the support of W with $\alpha(w) > 0$. On the other hand, $h(w)$ has a "welfare" weight interpretation and is thus positive by construction.

i.e., admitting everyone with $\mu^*(w) \geq 0$ will exceed the budget in expectation.

Proposition 1 *Under Condition (C), the solution to the problem:*

$$\sup_{p(\cdot)\in\mathcal{F}} \int_{w\in\mathcal{W}} p(w)\, h(w)\, \alpha(w)\, \mu^*(w)\, dF_W(w), \quad \text{subject to} \quad \int_{w\in\mathcal{W}} p(w)\, \alpha(w)\, dF_W(w) \leq c$$

takes the form:

$$p^{opt}(w) = \begin{cases} 1 \text{ if } \beta(w) > \gamma; \\ q \text{ if } \beta(w) = \gamma; \\ 0 \text{ if } \beta(w) < \gamma, \end{cases} \tag{1}$$

where

$$\beta(w) := h(w)\, \mu^*(w); \quad \gamma := \inf\{r : \int_{w\in\mathcal{W}} \alpha(w)\, \mathbf{1}\{\beta(w) > r\}\, dF_W(w) \leq c\};$$

and $q \in [0,1]$ *satisfies*

$$\int_{w\in\mathcal{W}} \alpha(w) \left[\mathbf{1}\{\beta(w) > \gamma\} + q\mathbf{1}\{\beta(w) = \gamma\}\right] dF_W(w) = c.$$

The solution (1) is unique in the F_W-almost-everywhere sense (i.e., if there is another solution, it differs from (1) only on sets whose probabilities are zero with respect to F_W).

The result basically says that the planner should order individuals by their values of $\beta(W)$ and first admit applicants with those values of W for which $\beta(W)$ is the largest, then to those for whom it is the next largest and so on till the budget is exhausted. If the distribution of $\beta(W)$ has point masses, then there could be a tie at the margin, which is then broken by randomization (hence the probability q). In the absence of any point masses in the distribution of $\beta(W)$, the optimal protocol is of a simple threshold-crossing form $p^{opt}(w) = \mathbf{1}\{\beta(w) \geq \gamma\}$. For the rest of the paper, we will assume that this is the case.

Academically efficient admissions: We define an academically efficient admission protocol as one which maximizes expected performance of the incoming cohort subject to the restriction on the number of vacant places. Such an objective is also "academically fair" in the sense that the expected performance criterion gives equal weight to the *outcomes* of all applicants, regardless of their value of W, i.e., $h(w)$ is a constant. In this case, the previous solution takes the form $p^{opt}(w) = \mathbf{1}\{\mu^*(w) \geq \gamma\}$, where γ solves

$$c = \int_{w\in\mathcal{W}} \alpha(w)\, \mathbf{1}\{\mu^*(w) \geq \gamma\}\, dF_W(w).$$

The key feature of the above rule is that γ does not depend on W and so the value of an applicant's W affects the decision on his/her application only through its effect on $\mu^*(W)$. To get some intuition on this, consider the case where one of the covariates in W is gender and assume that the admission threshold for women, γ_f, is strictly lower than that for men, γ_m. Then the marginal female,

admitted with $w = (x, female)$, contributes $\gamma_f \times \alpha(x, female)$ to the expected aggregate outcome and takes up $\alpha(x, female)$ places, implying a contribution of γ_f $(= \alpha(x, female)\gamma_f/\alpha(x, female))$ to the objective of average realized outcome. Similarly, the marginal rejected male, if admitted, would contribute γ_m to the average outcome. Since $\gamma_m > \gamma_f$ we can increase the average outcome if we replaced the marginal female admit with the marginal male reject. Thus different thresholds cannot be consistent with the objective of maximizing the mean outcome.

3 Econometric Model

The economic model above takes the entire university as a single decision-making entity whereas in reality, admission decisions are made by individual officials who apply at least some personal discretion and/or display heterogeneity in taste or knowledge in making the decision. This heterogeneity is likely to introduce idiosyncratic variation in the individual decisions around a baseline university-wide policy. In view of this, we extend the previous economic model into an econometric one, which incorporates heterogeneity across admission officers and forms the basis of our empirical analysis.

To set up the empirical framework, we assume that we (i.e., the analysts) observe W and D for applicants in the current year, drawn in an independent and identically distributed (I.I.D.) fashion from a distribution of potential applicants. In addition, we have data on one or more cohorts of applicants in past years who had enrolled in the university. For each such enrolled applicant, we observe W and the outcome of interest Y (e.g., examination score after the first year of university). When referring to variables from past years or expectations calculated on the basis of past variables, we will use the superscript "^{P}". We may or may not observe the outcomes of current year applicants, depending on the timing of data collection. Our methodology does not depend on the availability of outcome data for current applicants. Our aim is to evaluate academic efficiency of current year's admission, given data on (X, G, D) for all current year applicants and $(Y^P, X^P, G^P \mid A^P = 1)$ for past years' (successful) applicants.

Let

$$\mu^P(x, g) = E\left[Y^P | X^P = x, G^P = g, A^P = 1\right] \tag{2}$$

denote the conditional expectation of outcome Y^P for a past enrolled applicant given his/her characteristics $(X^P, G^P) = (x, g)$. We assume that when admission tutors 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 had enrolled in previous years. Note that $\mu^P(x, g)$ is in general different from $E[Y^P | X^P = x, G^P = g]$ which is typically unknown to admission tutors in universities (or loan tutors in banks in our loan application example above).[10] Indeed, a large literature in educational statistics on so-called "validation studies" use predicted

[10] If there existed trial data where admissions were randomized, then the latter can be calculated and used instead

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 admission tutors are likely to do – rather than what one could have done under ideal circumstances like having experimental data – using $\mu^P(x,g)$ rather than $E[Y^P|X^P=x,G^P=g]$ is the correct approach here.

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

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

This is the set of the values of X^P which occur among the admits of type g in past years and so one can, in principle, calculate (i.e., estimate) the values of $\mu^P(x,g)$ when $x\in\mathcal{X}_g$. We assume that a current year applicant $i(\in\{1,\ldots,n\})$ with $G_i=g$ and $X_i=x\in\mathcal{X}_g$ is offered admission if and only if $\mu_i^{*P}(x,g)\geq\gamma_g$, where $\mu_i^{*P}(x,g)$ denotes the subjective conditional expectation of the admission-tutor handling applicant i's file and γ_g denotes the university-wide baseline threshold for applicants of demographic type g.[11] We specify that $\mu_i^{*P}(X_i,G_i)=\mu^P(X_i,G_i)-\varepsilon_i$ where $\mu^P(\cdot,\cdot)$ denotes the true mathematical expectation defined in (2) and ε_i is a "friction" or "slippage" term capturing, for instance, a deviation of the admission tutor's subjective expectation from the true mathematical expectation.

Thus the admission process for an applicant i satisfies:

Assumption 1

$$D_i=\mathbf{1}\left\{\mu^P(X_i,G_i)\geq\gamma_{G_i}+\varepsilon_i\right\}\quad \textit{if } X_i\in\mathcal{X}_{G_i}, \tag{3}$$

where γ_{G_i} and $\mathcal{X}_{G_i}$ are defined for each individual i, analogously to γ_g and $\mathcal{X}_g$. For $x\notin\mathcal{X}_{G_i}$, the probability of an offer $\Pr\left[D_i=1|\ X_i=x\right]$ *is bounded away from* $1/2$.[12]

Academically Efficient Admissions: In this setting, we define an admission practice to be academically efficient/fair at the university level if and only if γ_g is identical across g. The

of $\mu^P(x,g)$. Alternatively, if Y^P were independent of A^P, given $\left(X^P,G^P\right)$ (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 typically act on the basis of $\mu^P(x,g)$, whether or not it equals $E[Y^P|X^P=x,G^P=g]$.

[11] We will hereafter write a random variable/vector with a subscript i, e.g., Z_i, to indicate that it is associated with an individual applicant i, while we often suppress the subscript (as heretofore), e.g., Z, to denote one for a generic applicant.

[12] It is not necessary for our analysis to specify how g-type applicants with values of X outside $\mathcal{X}_g$ are treated in the current year, since all the information regarding the parameter of interest γ_g will come from those g-type applicants whose predicted probability of getting an offer is one-half. Unless the admission process changes drastically between the two periods, it is reasonable to expect that characteristics which do not occur at all among past admits will be admitted in the current year either with very low probability (if they have lower test scores than anyone admitted in any previous year) or with very high probability (if they have higher test scores than anyone admitted in any previous year). In either case, the probability will be bounded away from 1/2.

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_g for, say, females than males implies that a male applicant with the same expected outcome as a female applicant is more likely to be admitted. Conversely, under affirmative action type policies, γ_g will be lower for those gs which represent historically disadvantaged groups. Therefore, we are interested in identifying the value of the threshold γ_g for various values of g and testing if they are identical across g. We will call γ_g the "admission threshold" for group g. Further, among g type applicants, those whose X makes $\mu^P(X, g) = \gamma_g$ will be referred to as the *marginal* g type candidates. It is important to note that our definition of the marginal does not involve ε. The justification for this is that no matter what the university's baseline policy, it has to allow for slippages arising from individual tutors guessing the academic potential of an applicant based on subjective beliefs. As long as these slippages are not systematic – as captured by a zero median restriction (see below) – the university can be regarded as practising academically efficient admissions when γ_g does not vary by g.[13]

It is also important to note that here we are not making any assumption about whether or not G affects the distribution of the outcome, 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 and Mullainathan (2004, BM) 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 BM's finding to imply inefficiency according to our criterion, one needs to assume that, conditional on the information in CVs, race has no impact on average worker productivity.

A third point is that our requirement of economic *efficiency* can also be interpreted as a requirement of academic *fairness* in the following sense. Suppose G denotes socioeconomic status and X denotes score on the admission test. Then it seems that "fairness" gives more credit to an applicant from underprivileged backgrounds who studied in schools with lower resources but has the same score on the entrance test as an applicant who had studied in a fee-paying school with abundant resources. The underlying assumption, of course, is that the former student is more "meritorious." Conditioning the expected outcome μ^P on both X and G can reveal whether this judgement is appropriate precisely by predicting a higher eventual outcome for the first student if

[13] Our use of the term "marginal" is also different from the notion of marginal individuals in Carneiro, Heckman and Vytlacil (2009, CHV). Firstly, their paper's set-up involves an instrumental variable (IV), satisfying an exclusion restriction and a large support condition, 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 satisfies $\mu^P(x, g) = \gamma_g$.

the assumption is true and a lower eventual outcome for him/her otherwise. A G-blind admission process where μ^P (and γ) is not conditioned on G will not reveal this difference and is therefore both inefficient and as academically "unfair" in this sense.

Identifying assumptions: For the identification/estimation of γ_g, we impose the following conditions for current year applicants $i = 1, \ldots, n$.

Assumption 2 *(i)* $\{(X_i, G_i)\}_{i=1}^n$ *is an I.I.D. sequence of random vectors and* $\{\varepsilon_i\}_{i=1}^n$ *is a sequence of random variables which is first-order stationary (i.e., the marginal distribution of* ε_i *is the same as that of* ε_j *for any* $i \neq j$*) and is* α*-mixing (strong mixing) with the mixing coefficients* $a_m \leq Am^{-b}$ *for some constants* $A > 0$ *and* $b > 2$.[14] *(ii)* $\text{median}[\varepsilon_i | X_i, G_i] = 0$ *almost surely. (iii) The distribution of* ε_i *has a strictly positive density (with respect to the Lebesgue measure) around* 0*, given* (X_i, G_i)*, almost surely.*

Discussion: The presence of ε_i in (3) allows admission to be non-deterministic, given X_i and G_i. We allow the *friction* sequence $\{\varepsilon_i\}_{i=1}^n$ to be non-I.I.D. and (weakly) dependent. As discussed above, we interpret the friction ε_i as the expectational error made by the admission tutor handling applicant i's file. If several candidate files are handled by the same tutor, then it is possible that a tutor-specific effect leads to correlations within some of the ε_is. Our α-mixing condition in part (i), which is one of the weakest conditions for the weak dependence used in the literature, will capture this sort of situation (the degree of dependence is controlled by the mixing coefficients). The condition says that ε_i and ε_{i+l} are almost independent when l is large enough (asymptotically independent as $l \to \infty$). In particular, if subjective errors of different tutors are independent and only a small number of applicants are allotted to each tutor (which means that under the hypothetical situation when the number of applicants n tends to ∞, the number of tutors also tends to ∞ with the same order as n), the mixing condition in part (i) of Assumption 2 should be satisfied.[15] [16]

Part (ii) of Assumption 2 is a now-familiar median restriction assumption, first used in discrete choice settings by Manski (1975). In the admissions context, it will hold when systematic determinants of admission, such as past test scores, interview grades and demographic characteristics

[14] The α-mixing coefficients are defined as follows (see, e.g., Bradley, 2005):

$$a_m := \sup_{1 \leq k \leq n-m} \sup\{|\Pr[A \cap B] - \Pr[A]\Pr[B]| : A \in \mathbb{F}_{m+k}^n,\ B \in \mathbb{F}_k^1\},$$

for $m (= 1, 2, \ldots)$, where $\mathbb{F}_k^j (\in \mathbb{F})$ denotes the σ-algebra generated by $\varepsilon_j, \varepsilon_{j+1}, \ldots, \varepsilon_k$ (with $(\mathbf{\Omega}, \mathbb{F}, \Pr)$ denoting the probability space where our econometric model is defined).

[15] We note that our mixing condition still allows for some cases when different tutors have (weakly) correlated beliefs.

[16] Note also that by the first-order stationarity condition, together with the I.I.D. condition on covariates $\{(X_i, G_i)\}_{i=1}^n$, we have $p(x, g) = \Pr[D_i = 1 \mid X_i = x, G_i = g]$ well-defined as a function independent of i, since $\{(D_i, X_i, G_i)\}_{i=1}^n$ is also first-order stationary. While the I.I.D. condition of $\{(X_i, G_i)\}_{i=1}^n$ can be easily relaxed to being first-order stationary and α-mixing (as $\{\varepsilon_i\}_{i=1}^n$), we impose it mainly for simplifying our technical proofs.

are observed by the econometrician but idiosyncratic preferences and/or the deviation of the admission tutors' subjective expectation from the true $\mu^P(\cdot,\cdot)$ are not. Part (ii) basically says that the true academic potential of any randomly-picked applicant of a given type (defined by a value of $W=(X,G)$) is equally likely to be over or underestimated. This assumption may be thought of as the median analog of "rational expectations" on the part of admission tutors who might be assigned to the applicant's file. The zero-median restriction is natural here since systematic errors on the part of tutors can be absorbed in γ_g (see Footnote 7).

It is important to note that (i) and (ii) of Assumption 2 are much weaker than requiring ε to be independent of (X,G). One case where full independence will fail is where for some historically under-represented group g, the conditional variance of ε, i.e., $\text{Var}[\varepsilon|X=x,G=g]$ is larger for every x, reflecting larger magnitudes of error in tutors' subjective beliefs regarding those types of individuals with whom the tutor has had less experience. The conditional median restriction is robust to such scale dependence, as is well-known since Manski's (1975) maximum score analysis, and turns out to be sufficient here for identifying γ_g 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) as is effectively assumed via an "index restriction" in Chandra and Staiger (2009, page 7), who analyze fairness of surgical treatment assignment in a healthcare context. Observe also that our zero-median restriction is weaker than requiring the error distributions to be symmetric about zero and thus allows for arbitrary amounts of skewness.

A "descriptive" interpretation of the zero conditional median restriction is as follows. First note that since $\Pr[\varepsilon<0|G=g]=\int \Pr[\varepsilon<0|X=x,G=g]dF_{X|g}(x)$, we have that median $[\varepsilon|X,G]=0$ almost surely implies that median $[\varepsilon|G]=0$ almost surely ($F_{X|g}(x)$ denotes the conditional C.D.F. of X given $G=g$). Now, one may view the right-hand side (RHS) component determining the admission in (3), viz., $\gamma_g+\varepsilon$, as a random admission threshold faced by applicants of type g. The previous argument and (ii) of Assumption 2 then imply that the median of this threshold's distribution for g-type applicants is γ_g. Thus testing the equality of, say, γ_{male} and γ_{female} is equivalent to testing whether the distribution of admission thresholds faced by male applicants has the same median as the distribution of thresholds faced by female applicants.[17]

Given the condition (ii) of Assumption 2, it follows that the marginal admits among type g applicants, i.e., those with values of x satisfying $\mu^P(x,g)=\gamma_g$, will also satisfy

$$\Pr[D=1|X=x,G=g]=\Pr[\varepsilon<0|X=x,G=g]=1/2,$$

which has an intuitive interpretation as follows. According to our model, those applicants whose $\mu^P(X,g)$ is very high relative to γ_g will be admitted with probability close to 1. These are the

[17] Such an interpretation would naturally carry over to assumptions restricting any other conditional quantile, besides the median, to be zero. However, such a restriction will not have the "rational expectations" type structural interpretation possessed by the median and hence we do not consider other quantiles here.

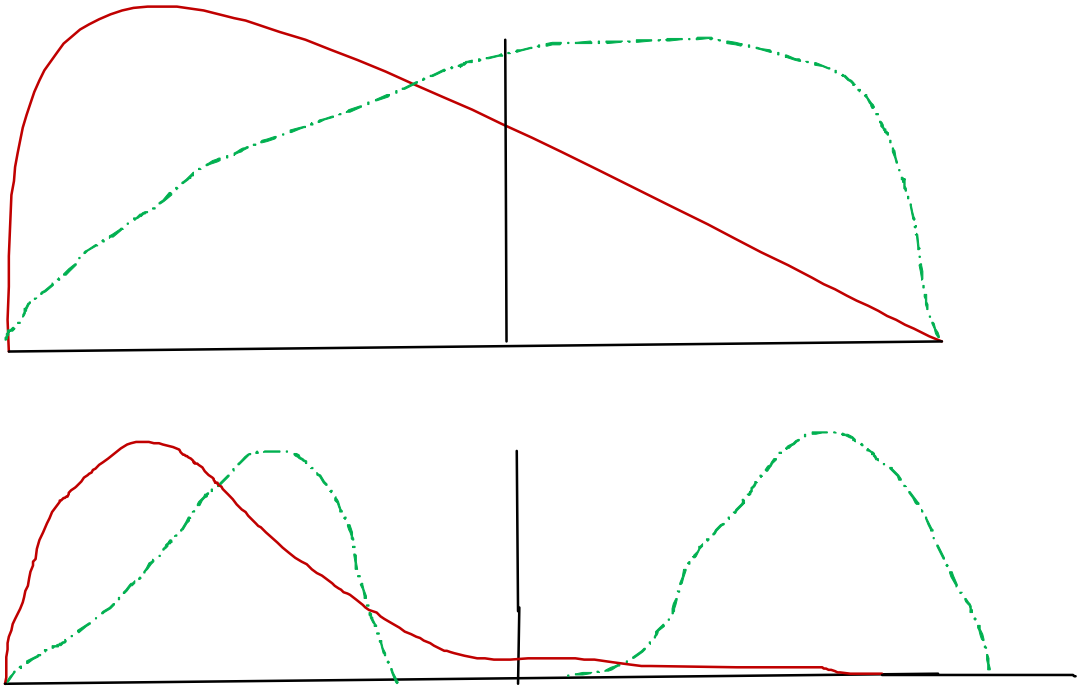
candidates who would get in with certainty if there were no frictions. Conversely, those whose $\mu^P(X,g)$ is very low relative to γ_g will be admitted with probability close to 0. They would not have been admitted in the absence of any frictions. When we have a candidate whose $\mu^P(X,g)$ is exactly at the threshold γ_g, then in the absence of any friction, the university would be indifferent between admitting and not admitting this candidate. In this sense, such candidates are marginal. The stochastic frictions make them equally likely to get in or not and hence the probability of exactly one half.

Finally, part (iii) of Assumption 2 is a regularity condition that aids the proof of identification. It will obviously hold for a wide class of continuously-distributed random variables.

Lastly, we will make a technical assumption which would imply the existence of a common feasible threshold. Toward that end, let Υ_g denote the support of the distribution of $\mu^P(X_i, G_i)$, given $G_i = g$.

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

To interpret this assumption, consider the case where G_i denotes gender and X_i contains one or more continuous variables like pre-admission test scores. Then the assumption says that the (conditional) expected outcome for males and that for females take values in the same set. Therefore, given any value $x \in \Omega_{male}$, where Ω_g denotes the support of the distribution of X_i given $G_i = g$, 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). Fix an arbitrary $\gamma \in I$. Then, under the above assumption, for each g, there exists $x^*(g) \in \text{int}(\Omega_g)$ such that $\mu^P(x^*(g), g) = \gamma$ for every g. So we can define individuals of type g with $X = x^*(g)$ to be the "ideally marginal" admits among type g, i.e., those $(x^*(g), g)$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 efficient, then for every g, $\mu^P(x^*(g), g) = \gamma$; if not, and the marginal admits are denoted by $\tilde{x}(g)$ for group g, then $\mu^P(\tilde{x}(g), g) = \gamma_g$ will differ across g. If the common support assumption did not hold, then it would be possible that admission is academically efficient with a common γ which lies within the support of $\mu^P(X_i, G_i)$ conditional on $G_i = male$ but not of $\mu^P(X_i, G_i)$ given $G_i = female$. In that case, for males, we will have equality at the margin but for females, the marginal admits will have expected outcome exceeding the threshold if γ lies in a "hole" with respect to the support of $\mu^P(X_i, G_i)$ given $G_i = female$. Figure 1 illustrates the point. The common support assumption would hold if a situation as in the top panel of Figure 1 holds, where both curves have positive height at the cutoff-point γ, marked by the vertical line. We in particular note that 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



The red solid curve represents a fictitious conditional density of $\mu^P(X, male)$ and the green dashed curve the density of $\mu^P(X, female)$. 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 $\mu^P(X, female)$. So there is no x in the support of X for females where $\mu^P(X, female)$ can equal the common threshold.

Figure 1:

assumption is that it enables us to interpret the inequality between group-specific thresholds as being symptomatic of academically inefficient admissions.

4 Identification of γ_g

4.1 Identification method

The basic identification idea is to use for each fixed g, the median restriction and the observed $\Pr[D_i = 1|X_i = x, G_i = g]$ to identify the values of X_i defining the marginal admits, i.e., those x for which $\Pr[D_i = 1|X_i = x, G_i = g] = 1/2$ 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 the distribution of G_i in the current year, the distribution of the random variable $\mu^P(X_i, G_i)$ conditional on $G_i = g$ has a strictly positive density (with respect to the Lebesgue measure) on an open interval around γ_g.*

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

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

Proposition 2 *Suppose that Assumptions 1, 2-(ii), 2-(iii) and 4 hold. Then, for each g, the threshold γ_g is point-identified for each g, given $\mu^P(\cdot, \cdot)$.*

Proof. Note that if there exists an $x \in \mathcal{X}_g$ such that $\Pr[D = 1|X = x, G = g] = 1/2$, then we must have

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

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

$$E\left[\mu^P(X, G) - \gamma_g \mid \Pr[D = 1 \mid X, G] = 1/2, G = g\right] = 0. \tag{4}$$

This implies that γ_g can be identified via the equality:

$$\gamma_g = E\left[\mu^P(X, G) \mid \Pr[D = 1 \mid X, G] = 1/2, G = g\right].$$

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 $\Pr[D=1|X=x,G=g] = 1/2$ – 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. ■

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 one half. These are the marginal candidates of type g whose X takes values in the set Π_g. Then, calculate predicted outcome, using data on past years' admits. Finally, average these predicted outcomes across current 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 which illustrates the admission process for a scalar X and for a fixed g. On the horizontal axis we plot values x of X and on the vertical axis we measure $\mu^P(x,g)$ in the top panel and the corresponding probability of offer $p(x,g)$ in the bottom panel. In the top panel of the graph, we plot $\mu^P(x,g)$ against x by the dashed line and mark the admission threshold $\gamma\left(=\gamma_g\right)$ by the horizontal dashed line. In the bottom panel, we plot the corresponding admission probability $p(x,g)$ against x in the absence of errors (dashed line segments) 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 $\mu^P(x,g) < \gamma_g$ and equal to one where $\mu^P(x,g) \geq \gamma_g$. Now consider what happens when there are stochastic perception errors. Such errors will make the perceived expectation at any value x of X 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, are centered at the true $\mu^P(x,g)$. Now, it is probabilistically determined whether a particular applicant with a value x of X is admitted, depending on whether the noisy subjective expectation exceeds γ_g. At a point as xh on the right, we have $\mu^P(xh,g) > \gamma_g$. In this case, the probability $p(xh,g)$ exceeds one half. This probability is computed as the area under the density curve at xh over the region above γ_g in the upper panel, and it is marked by the vertical height of the solid curve in the lower panel. Similarly, at a point as xl on the left, we have $p(xl,g) < 1/2$. Only at the point xm where $\mu^P(xm,g) = \gamma_g$, the density hump at xm is centered around γ_g, which makes the probability of being admitted exactly one half. Notice that this argument does *not* require the density curves to be symmetric or have the same spread. What is required here is that for each x, the area under the density curve over the region above $\mu^P(x,g)$ should be equal to that over the region below $\mu^P(x,g)$, i.e., 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 condition in Assumption 3 in the previous section for $\mu^P(\cdot,\cdot)$.

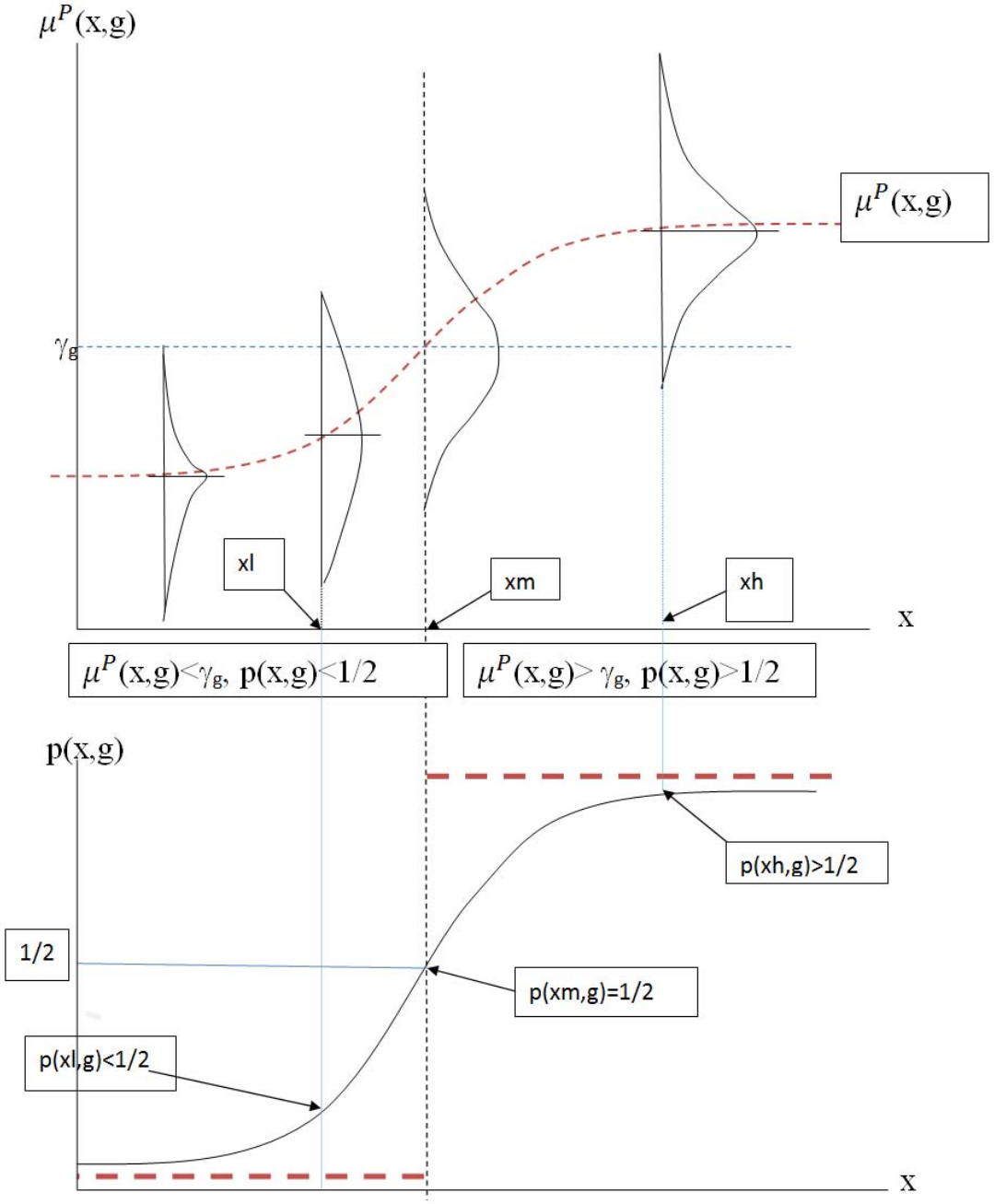


Figure 2: Graphical illustration of identification for fixed G=g

Figure 2:

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 (IB) 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 can simply use IB scores instead of A-level scores as the corresponding variable in X for these students, and can compute predicted outcomes and probabilities of an offer (by corresponding regressions; see Section 5). Thereafter, all applicants are pooled together and the analysis proceeds exactly as before.*

Remark 2 *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 as a numerical score or grade by university officials for easy make comparisons at the end. This score can be used as a component of X in our proposed methodology.*

Remark 3 *Our analysis does not require background information for past years' applicants who were rejected. Universities typically do not store this information and hence it useful to have a method which does not require it.*

4.2 Comparison with other identification strategies

As outlined in Introduction, we are not aware of any existing empirical test of *outcome-based* efficiency or fairness in college admissions.

A previous attempt at identifying treatment thresholds – and consequently the marginal treatment-recipients – in the healthcare context is Chandra and Staiger (2009, CS). CS attempt to identify difference in expected outcome thresholds for surgery by assuming an index restriction on the unobservable's distribution. This approach fails when the unobservable's distribution has general covariate-dependent variance, 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 healthcare context, Bhattacharya (2012) suggests an alternative approach to testing outcome-oriented treatment assignment via a partial identification analysis using a combination of observational data and prior experimental findings from randomized controlled trials. Such experimental results are typically difficult to come by in the college admission context.

In the context of law enforcement and medical treatment, some alternative approaches have been proposed for testing 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 (c.f., Persico, 2009, for a review). The usual approach in this literature is not to detect the marginals directly, as in the present paper, but to utilize some

specific institutional feature of the empirical context under study, which would equate the outcome of the marginal with that of the average in a known subset of the population and thereby eliminate the so-called "infra-marginality" problem. However, none of these existing, context-specific approaches is applicable in our setting. For instance, in the context of policing, Knowles, Persico and Todd (2005)[18] use the assumption that criminals rationally alter their potential outcomes in response to the crackdown regime, e.g., by altering the amount of contraband they carry. Such immediate responses are not feasible in the admissions context where applicants' academic outcomes depend on long-term human capital accumulation. In the medical setting, Anwar and Fang (2011) assume that physicians optimally choose a continuous variable related to diagnostic tests before discharging patients. A test of fair discharge is then based on comparing the average readmission rates of discharged patients of different race who had undergone the diagnostic test at the physician-optimized level of intensity. In the admission set-up, there is usually no such continuous choice variable available to admission tutors.

In an ongoing project on a methodologically related theme, Jiang, Nelson and Vytlacil (2011, JNV) 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 an analog of the $\mu^P(\cdot)$ function in the deterministic model $D = \mathbf{1}\{\mu^P(W) > 0\}$ but when they only observe the distribution of $W|D = 1$. In contrast, we observe W for all applicants, the relevant μ^P function is identified directly from past outcome data, the determination of D involves additional heterogeneity and the goal is to identify the threshold γ's which potentially varies 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.

5 Estimation and Inference

We now consider the calculation of γ_g from admissions data collected for several cohorts of applicants. We may view the current cohort as a random sample from a model describing the superpopulation of all potential applicants. Therefore, the values of γ_g calculated based on the present cohort will suffer from sampling uncertainty and a test of equality of γ_g's across g requires distribution theory, which we derive in this section.

Motivated by the restriction of (4), 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 sample size large enough, one can consider fully nonparametric estimation of $\mu^P(x, g)$,

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

$p(x,g)$ 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 state and prove formal theorems describing the distribution theory for this semiparametric case, c.f., Theorem 1 in Subsection A.2. 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 fully nonparametric estimation of $\mu^P(x,g)$ and $p(x,g)$, c.f., Theorem 2 in Subsection A.3.

In the semiparametric approach, we estimate $\mu^P(\cdot,\cdot)$ and $p(\cdot,\cdot)$ parametrically in the first step using past and current cohort data, $\{(A_j^P, A_j^P Y_j^P, X_j^P, G_j^P)\}_{j=1}^N$ and $\{(D_i, X_i, G_i)\}_{i=1}^n$, respectively, where N is the sample size of past cohorts and n is that of the current cohort.[19] Then, in the second step, we use the current cohort data $\{(X_i, G_i)\}_{i=1}^n$ to estimate γ_g by a weighted average of $\hat{\mu}^P(X_i, G_i)$, where the weights are based on a decreasing function of the distance between $\hat{p}(X_i, G_i)$ and $1/2$,

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

Here $K_h(z) := K(z/h)/h$; $K(\cdot)$ is a kernel function ($\mathbb{R} \to [0,\infty)$); h is a smoothing parameter (bandwidth); $\hat{p}(x,g)$ and $\hat{\mu}^P(x,g)$ are first-step estimators of $p(x,g)$ and $\mu^P(x,g)$, respectively. This $\hat{\gamma}_g$ is a weighted average of predicted outcomes of (X_i, g)-types whose predicted probability of getting an offer, $\hat{p}(X_i, g)$, is close to a half, where closeness is determined by the kernel K and the bandwidth h.

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

We will let the sample size of past cohorts and that of the present cohort to be of the equal order of magnitude. For notational simplicity in deriving our asymptotic theory, we assume that $n = N$ (while this assumption can be easily generalized, say, $n = O(N)$).

For the fully parametric case, due to the smoothness of the estimator of γ_g in the regression parameters, the estimator possess the $\sqrt{n}$-consistency and asymptotic-normality properties, which allows us to use a bootstrap method to obtain standard errors. The semiparametric case is somewhat different from standard two-step estimators where the first step is parametric and the second step involves some form of averaging of the first step estimation errors, leading eventually to$\sqrt{nh}$-consistent estimates. Here, due to kernel smoothing at the second step, even if the first step is

[19] Given $A_j^P = 1$, we can observe the outcome Y_j^P, and if $A_j^P = 0$, we say that the zero outcome is observed. Therefore, we may regard $\left(A_j^P, A_j^P Y_j^P, X_j^P, G_j^P\right)$ is observed for every j.

parametric, one cannot estimate γ_g at the parametric rate. Moreover, because both the conditioning variable $p(X_i, G_i)$ and the dependent variable $\mu^P(X_i, G_i)$ are estimated here, it is not trivial to derive the distribution theory, which is more complicated than standard nonparametric regression analysis. We now outline this distribution theory.

Remark 4 *It is important to note that in the numerator of (5) we have to use $\hat{\mu}^P(X_i, G_i)$ rather than current year outcomes Y_i even if the latter are available at the time of analysis. The reason is that the admission processes and acceptance patterns in the current year might differ from those in the past years so that the distribution of $Y \mid X, G, A = 1$ in the current year could be different from that of $Y^P \mid X^P, G^P, A^P = 1$. It is the latter distribution and not the former which is available to admission tutors at the time of making the admission decision. Therefore, testing efficiency or fairness of admissions in the current year requires the use of $\hat{\mu}^P(X_i, G_i)$ which is based on the latter distribution.*

Distribution of the semiparametric estimator: For the first stage, one may use any parametric model satisfying some mild conditions (c.f., Assumptions 8 and 9, below) e.g., a probit or logit model for $p(x, g)$; and a linear (regression) model for $\mu^P(x, g)$. Define $\tilde{\gamma}_g$ as the infeasible estimator that would result if the true values $\mu^P(X_i, G_i)$ and $p(X_i, G_i)$ were used instead of their estimates:

$$\tilde{\gamma}_g = \frac{\sum_{i=1}^{n} K_h\left(p(X_i, G_i) - 1/2\right) \mu^P(X_i, G_i) \mathbf{1}\{G_i = g\}}{\sum_{l=1}^{n} K_h\left(p(X_l, G_l) - 1/2\right) \mathbf{1}\{G_l = g\}}, \tag{6}$$

for each g. We show in the Appendix (see Theorem 1) that our semiparametric estimator $\hat{\gamma}_{sp}(g)$ has the same asymptotic distribution as $\tilde{\gamma}_g$ (under the assumption that parametric forms of estimators of $p(x, g)$ and $\mu^P(x, g)$ are correctly specified). Since $\tilde{\gamma}_g$ is a nonparametric regression estimator of the dependent variable $\mu^P(X_i, G_i)$ evaluated at $p(X_i, G_i) = 1/2$, we can derive the following asymptotic result under several standard conditions:

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

where $\mathbf{B}(g)$ and $\mathbf{V}(g)$ denote bias and variance components, respectively. 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 sufficient technical conditions are formally stated as Lemma 1 in the Appendix (see also a remark on Assumption 7).

Remark 5 *Note that the convergence rate of $\hat{\gamma}_g$ does not depend on the dimension of X_i (or (X_i, G_i)) since the asymptotic distribution of $\hat{\gamma}_g$ and the infeasible $\tilde{\gamma}_g$ are identical (shown in Theorem 1 in the Appendix). As seen in (6), our estimation problem is essentially of one dimension, i.e., $p(x, g), \mu^P(x, g) \in \mathbb{R}^1$. It is worth noting that the distribution theory derived here differs*

from standard kernel regression theory since both the outcome $\mu^P(x,g)$ and the conditioning variable $p(x,g)$ are unobservable. However, the $\sqrt{nh}$ rate of $\hat{\gamma}_g$ is generic, in that it is obtained even when even when $p(x,g)$ and $\mu^P(x,g)$ are nonparametrically estimated in the first step, as shown in Theorem 2 in the Appendix A.3. This occurs because first-step (nonparametric) estimation errors average out at a fast enough rate to zero in the second step. However, a technical complication arises here due to the estimator's form in which the generated regressor $\hat{p}(X_i,G_i)$ is inside of the kernel function, as we can see in (5). In particular, showing that $K_h(p(X_i,G_i)-1/2)$ is well-approximated by $K_h(\hat{p}(X_i,G_i)-1/2)$ requires careful arguments and particular bandwidth choices, since the convergence of $K_h(\hat{p}(X_i,G_i)-1/2)$ to $K_h(p(X_i,G_i)-1/2)$ only occurs more slowly than that of $\hat{p}(x,g)$ to $p(x,g)$.

Remark 6 *One can find several two-step nonparametric estimators in the literature, for example, Mammen, Rothe and Schienle (2011), Rilstone (1996) and Sperlich (2009). However, these authors analyze the setting where only regressors are generated, while in our setting both dependent and regressor variables are (nonparametrically) generated. The latter case seems not to have been well-investigated in previous studies. The aforementioned papers' results imply that final estimators are unaffected by the first step estimation errors (even though the convergence rate in the first step is slower than that in the second step) under suitable bandwidth choices. We show that this conclusion continues to hold when both the dependent and regressor variables are generated, which seems to be a new result. Additionally, as in Assumption 2, we allow for statistical dependence among observations. This sort of non-I.I.D. setting seems not to have been considered in previous studies on two-step nonparametric estimators.*

Choosing bandwidths: Note that our parameter of interest, γ_g is exactly the conditional mean $E[\mu^P(X_i,G_i)|p(X_i,G_i)=1/2,G_i=g]$. Therefore, we recommend a standard method based on the cross-validation (CV), which uses a global goodness-of-fit criterion for the conditional mean $E[\mu^P(X_i,G_i)|p(X_i,G_i)=1/2,G_i=g]$. In the present context, the CV is achieved by minimizing the leave-one-out criterion

$$R(h)=\sum_{j=1}^{n}\mathbf{1}\{G_j=g\}\times\left[\hat{\mu}^P(X_j,g)-m_{-j}(\hat{p}(X_j,G_j),g;h)\right]^2,$$

where

$$m_{-j}(a,g;h)=\frac{\sum_{1\leq i\leq n;\ i\neq j}K_h(\hat{p}(X_i,G_i)-a)\,\hat{\mu}^P(X_i,G_i)\,\mathbf{1}\{G_i=g\}}{\sum_{1\leq l\leq n;\ l\neq j}K_h(\hat{p}(X_l,G_i)-a)\,\mathbf{1}\{G_l=g\}}$$

is an estimator of $E\left[\mu^P(X_i,G_i)\ |\ p(X_i,G_i)=a,G_i=g\right]$, 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 $h=\hat{h}_{CV}/(\log n)$ 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.[20]

6 Application to Oxford admissions

Background: Our application is based on admissions data for two recent cohorts of applicants to an undergraduate degree programme in a highly popular subject at Oxford University. Like in many other European and Asian countries, students enter British universities to study a specific subject from the start, rather than the US model of following a broad general curriculum in the beginning, followed by specialization in later years. Consequently, admissions are conducted primarily by faculty members (i.e., admission tutors) in the specific discipline to which the candidate has applied. An applicant competes with all other applicants to this specific discipline and no switches are permitted across disciplines in later years. The admission process is in general – and at Oxford in particular – strictly academic where extra-curricular achievements, such as leadership qualities, suitability as team-members, engagement with the community etc., are given no weight. In that sense, undergraduate admissions at Oxford are more comparable with Ph.D. admissions in US universities. Furthermore, almost all UK applicants sit two common school-leaving examinations, viz., the GCSE and the A-levels before entering university. Each of these examinations requires the student to take written tests in specific subjects – e.g., math, history, English etc. – rather than an overall SAT-type aptitude test. The examinations are centrally conducted and hence scores of individual students on these examinations are directly comparable, unlike high-school GPA in the US where candidates undergo school-specific assessments which may not be directly comparable across schools. Consequently, much less weight is placed in the admission process on school-reference letters which tend to be somewhat generic and within-school ranks which are typically unavailable to admission tutors.

Choice of Sample: For our empirical analysis, 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. This form contains the applicant's unique

[20] 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.

identifier number, gender, school type, prior academic performance record, personal statement and a letter of reference from the school. The aptitude-test and essay-assessment scores are separately recorded. All of this information is then entered into a spread-sheet held at a central database which all admission tutors can access.

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 efficiency 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 our application, we use anonymized data for three cohorts of applicants from their records held 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, our personal experience as admission tutors and anecdotal experiences of 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 3. 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 3) to gauge its perceived importance in the admissions process. The A-level score appears 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 (including the essay) and the interview score for our analysis. We also use dummies for whether the applicant studied two specific subjects, at A-level, which are recommended by Oxford. A more detailed description of these covariates is provided in Table 0, below.

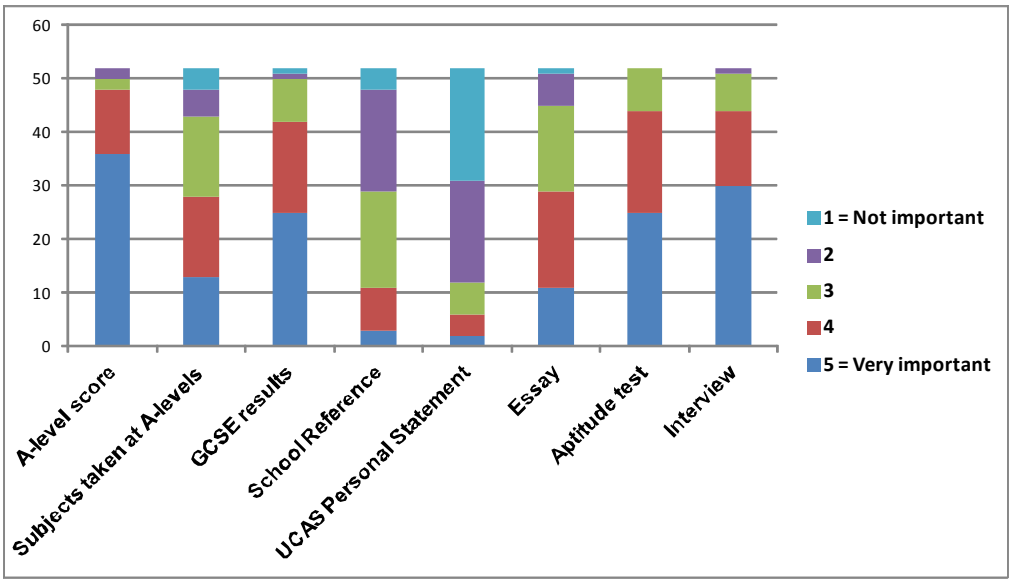


Figure 3:

Group identities G: We consider academic efficiency 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 admission 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.[21] 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 4 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

[21] Source: Guardian newspaper report at:
http://www.guardian.co.uk/education/2009/aug/19/oxford-university-men-places-women

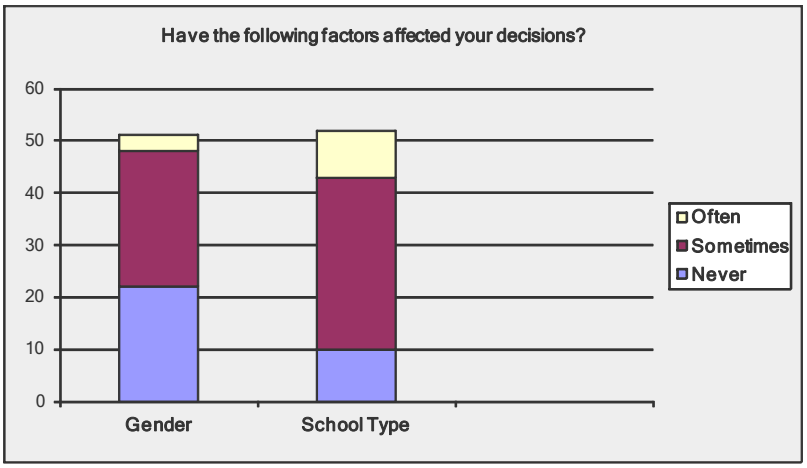


Figure 4:

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. The key advantage of using the preliminary year score as the relevant outcome measure is that every admit sits the same preliminary exam in any given year; so there is no confounding from the difference in score distributions across different optional subjects, as often happens in the final examinations at the end of the 3-year course. In fact, Arcidiacono, Aucejo and Spenner (2011) have documented, for Duke University data, large differences in patterns of major choice between candidates who are the likely beneficiaries of affirmative action policies during admissions compared to the major choice patterns of other enrolled students.

Summary statistics and success rates: We provide summary statistics for the entire 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 1 percentage point (20% of the overall standard deviation) higher in the first year exams. They perform worse on average in their GCSE and A-levels. These differences are statistically significant at the 5% level. 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 Tables 1A and 1B by showing that even after controlling for covariates, gender and school-type do not affect the *average* success rate among applicants. The value of McFadden's pseudo-R^2 for the probit model corresponding to Table 2A is about 50% and the corresponding

R^2 for a linear probability model (not reported here) is about 45% – which are about 10 times higher than the goodness-of-fit measures typically reported by applied researchers working with cross-sectional data. This suggests that the commonly observed covariates explain a very large fraction of admission outcomes. On a more minor note, Tables 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 used in Tables 2A and 2B are different, these two 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 (reported in Table 2B) 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, we use (5) to 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 1/2. 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 horizontal axis, marked "bw" represents the scale multiplying $n^{-1/5} \times sd$, where n is the relevant sample size and sd is the estimated standard deviation of the estimated regressor (the $\hat{p}(\cdot,\cdot)$). The scale bw was varied to ensure that the resulting bandwidths ($\hat{h} = bw \times n^{-1/5} \times sd/(\log n)$) lie between 0.01 and 0.99.

The numerical minimizer bw^* of this criterion over bw is used to compute the optimal bandwidth $\hat{h}^* = bw^* \times n^{-1/5} \times sd/\log(n)$ in each case.

In Table 3A, we show the difference in estimated admission thresholds for a range of bandwidths (which define "closeness to 1/2") and the Epanechnikov kernel $K(u) = (3/4)\left(1-u^2\right) \times \mathbf{1}\{|u| \leq 1\}$. The middle bandwidth (shaded row) is the optimal one ($\hat{h}^*$), described above. The other rows correspond to bandwidths that are 0.5 times the optimal one and 2 times the optimal one, respectively. The last row corresponds to a fully parametric analysis where the parametrically estimated $\hat{\mu}^P(X,g)$ is regressed on the parametrically estimated $\hat{p}(X,g)$ and its square and the predicted value at $\hat{p}(x,g) = 1/2$ is taken to be the estimate of $\hat{\gamma}$. The second row in Table 3A, for instance, may be read as follows. The entry in the first column specifies the scale by which the optimal

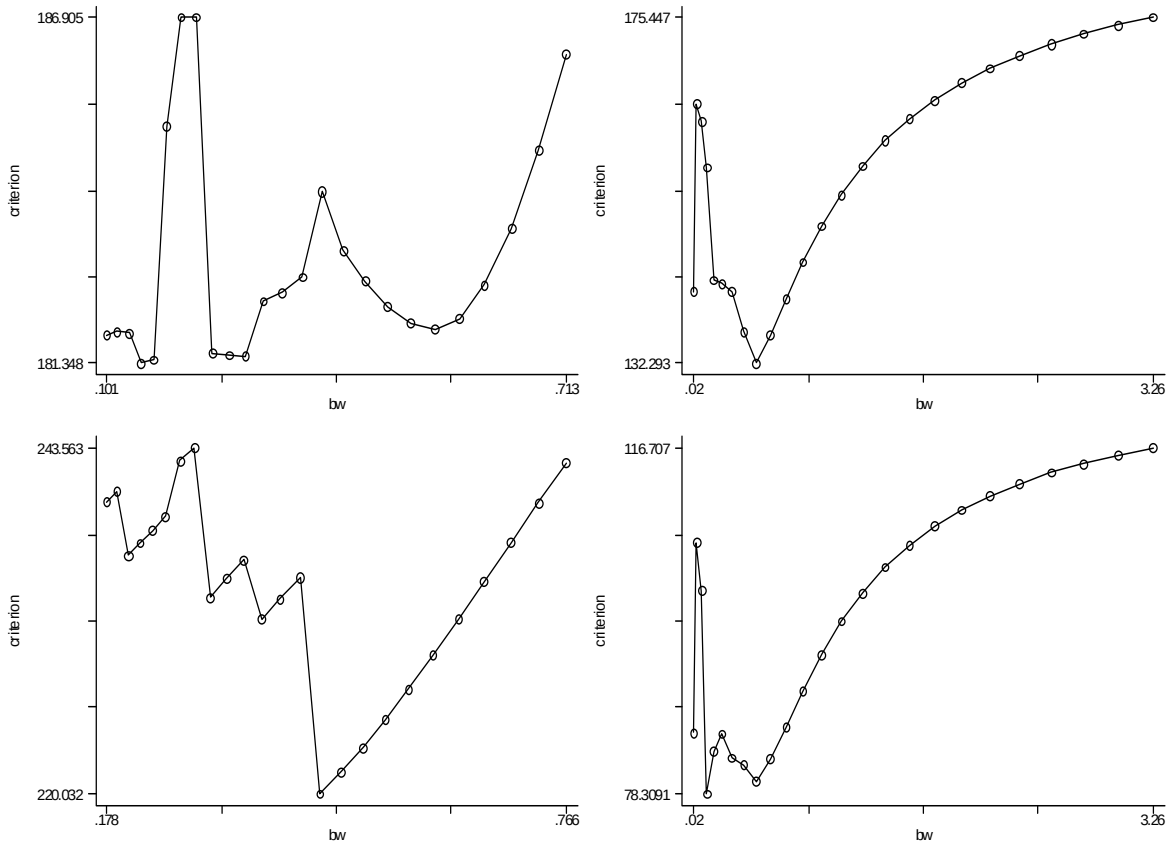


Figure 5:

bandwidth is multiplied (in this case 1), and the second column reports the male threshold computed by the corresponding scaled bandwidth. We see that the marginal male admits are expected to score 59.36 percent in their first year examination. The third column shows that the marginal female admits can be expected to score 55.67 percent, implying a difference of 3.7 percent (reported in the fourth column). This difference has a 1-sided p-value of 0.004 under the null of equal thresholds, reported in the fifth column. The 3.7 percentage point difference amounts to about $100 \times 3.7/6 = 61\%$ of one standard deviation of the overall first-year score distribution and thus represents a relatively large magnitude difference.

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 is much smaller than the 3.7 points difference among the marginal candidates.

Outcome variants: In Table 4A, we consider slightly different forms of the outcome, viz., (i) the chances of scoring at least 60 and (ii) securing at least 55. These correspond roughly to the 50th and 20th percentiles of the overall score distribution, respectively. In particular, the 55+ criterion corresponds to an admission process designed to maximize the probability of securing at least the minimum benchmark of a second class. 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 schools 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 is 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.s of the estimated $\mu^P(X, male)$ and $\mu^P(X, female)$ (the left panel) and those of the estimated $\mu^P(X, indep\_school)$ and $\mu^P(X, state\_school)$ (the right panel). It is clear that the male distribution first-order stochastic dominates the female distribution. This means that even if admissions are centrally conducted and are deterministic conditional on μ^P (i.e., there is no unobserved heterogeneity across admission tutors), *any* common acceptance rate across gender will result in a higher μ^P for the marginal accepted male than the marginal accepted female. This can be seen in Figure 6, by looking along any fixed cutoff on the vertical axis. Any such horizontal cut-off line[22] 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 tutors 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 action, viz., the overweighting of outcomes for historically disadvantaged groups. A second possibility

[22] For instance, if the top 30% of applicants are accepted among both males and among females, then we should be looking along the horizontal line at 1-0.3=0.7 on the vertical axis.

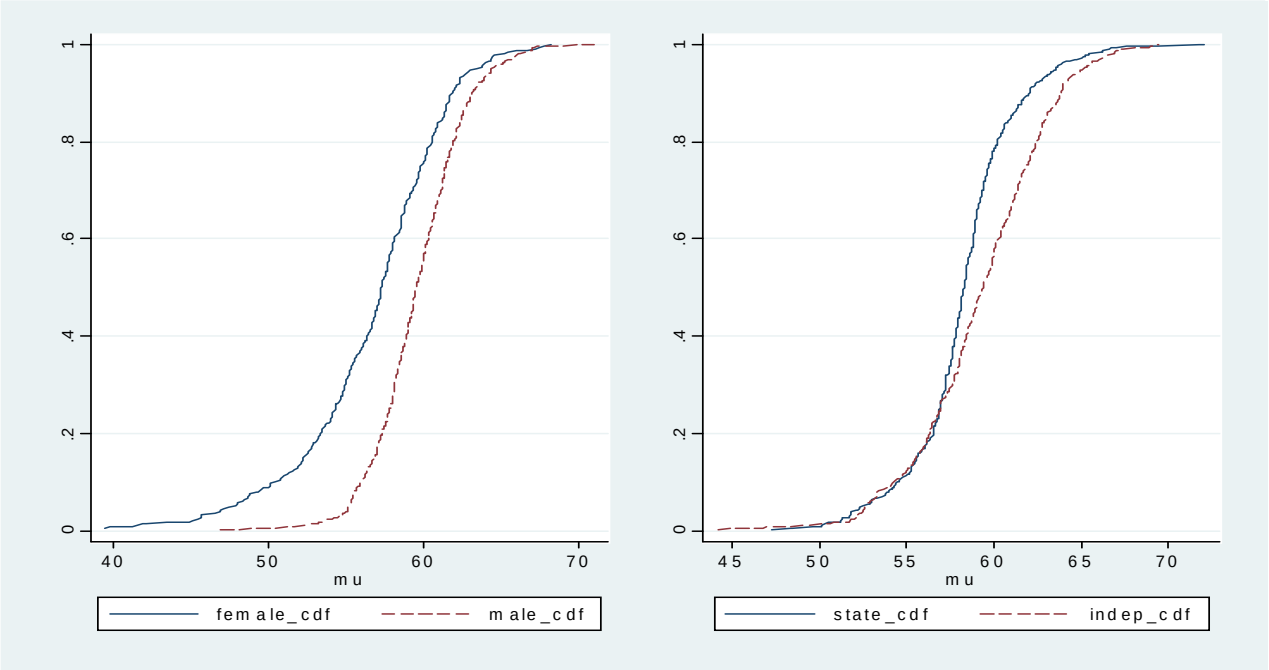


Figure 6:

is that, in face of political and/or media pressure, admission tutors try to equate an application success rate for, say, males with one for females, 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 tutor perception errors are identically distributed. A third possibility is that female applicants are set a lower admission threshold in order to encourage more female candidates to apply in future. Note from Table 1A that the number of female applications is nearly half the number of male ones. Regardless of what the underlying determinants of the tutors' 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.[23]

[23] 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. Indeed, this is the traditional approach which is taken by all of the papers cited above in that they all focus the analysis on a single outcome. It would be interesting to repeat our empirical analysis with performance data in the final examinations; 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 efficient in the sense of equalizing the treatment threshold for potential candidates across demographic groups. 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 interview performance. We assume that a researcher can access this background information by acquiring the application form and the performance scores and combine these with data on academic outcomes of applicants who were admitted to the university in past years. Such admission procedures and data situations are extremely common across universities in the world, making our methodology fairly generally applicable. Furthermore, academic researchers can normally obtain such information, possibly under confidentiality agreements, from their own institutions.

Once the data are obtained, one can use the analytical framework developed in this paper to analyze fairness of admissions. In this framework, the admission process is formulated as a stochastic, threshold-crossing model where academically fair (i.e., economically efficient) admissions correspond to the use of identical thresholds across demographic groups. Under suitable substantive and regularity conditions, we establish how these admission thresholds can be identified from admissions data for current applicants and performance data of students admitted in the past. We then propose methods of statistical inference which can be used to test equality of admission thresholds across demographic groups. Our methods are based on predicted probability of acceptance and predicted performance in university rather than directly on covariates. As such, these methods can be applied to situations where applicants come from diverse backgrounds and report scores from different aptitude tests (e.g., the A-levels versus the International Baccalaureate) since the necessary predicted values can be calculated based on candidate-specific covariates. Furthermore, we do not require any information for past applicants who were not accepted. This feature is convenient since universities normally do not store such data.

We apply our methods to admissions data for a large undergraduate programme of study at Oxford University and focus on first-year examination performance as the outcome of interest. These exams consist of common papers which are answered by all students and are blindly marked, i.e., the marking tutors do not know anything about the students' backgrounds. We find that the admission threshold faced by applicants who are male or from independent schools are higher than those for female or state-school applicants with the gender gap nearly 60% of a standard deviation of the overall exam performance and the private-state school gap nearly 28%. This contrasts sharply with average admission rates, which are 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.[24] 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 separate role and we see qualitatively similar results to those obtained when using the continuous outcome.[25] Nonetheless, for use in other applications involving continuously distributed outcomes, this may be a direction worth further exploration. 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. Third, it may be useful to perform an 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. Fourth, in our analysis of fair admissions, we have taken the applicant pool as given. Indeed, one dimension of enhancing social mobility is to encourage more students from under-represented socio-demographic groups to apply to elite universities (see the interpretation of our gender-results at the end of the previous section). It would be useful for future research to further investigate this issue. Finally, in ongoing work, we are (i) exploring the related but reverse question of how individual characteristics should be weighed in admission decisions and (ii) investigating how median independence and/or symmetry conditions can be used to detect inefficient treatment allocation in medical-type settings where trial data are frequently available but treatment assignment may be significantly affected by covariates unobserved by the data analyst.

[24] 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), Graham (2011) and Carrell, Sacerdote and West (2011).

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

Table 0: Variable labels

Variable-Label	Explanation
gcsescore	Overall score in GCSE, 0-4
alevelscore	Average A-level scores 80-120
took subject 1	Whether studied 1st recommended subject at A-level
took 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.

Table 1A. Summary Statistics by Gender

Variable	Obs	Mean	Obs	Mean	Difference	p-value
	Female		**Male**			
gcsescore	365	3.83	620	3.75	0.08	0
took subject 1	365	0.69	620	0.68	0.01	0.54
took subject 2	365	0.48	620	0.52	-0.04	0.27
alevelscore	365	119.73	620	119.44	0.29	0.01
aptitude test	365	62.53	620	65.24	-2.71	0
essay	365	63.23	620	64.49	-1.26	0
interview	365	64.68	620	65.29	-0.61	0.04
prelim_tot	119	60.98	206	61.89	-0.92	0.04
offer	365	0.363	620	0.357	0.01	0.41
accept	365	0.34	620	0.34	0.00	0.5

Note: The data pertain to two cohorts of applicants, broken up by gender. The variable names are explained in table 0. Column 6 records the p-value corresponding to a test of equal means across gender against a one-sided alternative. Gender differences in unconditional offer rates (highlighted) are seen to be statistically indistinguishable from zero at 5%.

Table 1B. Summary stats by School-Type

Variable	Obs	Mean	Obs	Mean	Difference	p-value
	State		**Independent**			
gcsescore	548	3.70	437	3.87	-0.17	0
took subject 1	548	0.64	437	0.73	-0.09	0.02
took subject 2	548	0.53	437	0.49	0.04	0.004
alevelscore	548	119.60	437	119.73	-0.13	0.02
aptitude test	548	63.82	437	64.94	-1.12	0.0015
essay	548	64.06	437	64.07	-0.01	0.5
interview	548	65.02	437	65.17	-0.15	0.65
prelim_tot	180	61.15	145	62.10	-0.95	0.03
offer	548	0.361	437	0.357	0.00	0.5
accept	548	0.33	437	0.35	-0.01	0.46

Note: The data pertain to two cohorts of applicants, broken up by type of high-school attended prior to applying. The variable names are explained in table 0. Column 6 records the p-value corresponding to a test of equal means across school-type against a one-sided alternative. Differences in unconditional offer rates across school-types (highlighted) are seen to be statistically indistinguishable from zero at 5%.

Table 2A. Probit of receiving offer

Regressor	Coef.	Std. Err.	z	p-value
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

N=985, Pseudo-R-squared=0.5

Note: The data pertain to two cohorts of applicants. The variable names are explained in table 0. The table presents the coefficients in a probit regression of getting an offer. The last column reports a 2-sided p-value corresponding to a test of zero effect.

Table 2B. Regression of first-year score

	Coefficient	Std. Err.	t	p-value
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

N=325, R-squared=0.16

Note: The data pertain to two cohorts of applicants. The variable names are explained in table 0. The table presents the coefficients in a linear regression (with heteroskedastic errors) of performance in first-year examinations at Oxford on pre-admission characteristics. The last column reports a 2-sided p-value corresponding to a test of zero effect.

Table 3A. Thresholds by Gender

Outcome mean=61.54, std dev=5.2

Method	Male-thld	Fem-thld	Male-Fem	p-value
Scale=0.5	59.16	55.5	3.66	0.0004
Scale=1.00	59.36	55.67	3.69	0.0004
Scale=2	59.91	56.15	3.76	0.0001
Parametric	60.51	56.86	3.65	0.0004

Note: This table presents the estimated admission thresholds for expected performance by gender. These thresholds are calculated via equation (5) in the text where mu-hat and p-hat are estimated via linear regression and probit respectively and the threshold is obtained via a nonparametric regression of the estimated muhat on the estimated phat evaluated at phat equals one-half. Each of the first three rows corresponds to a different choice of bandwidth. The middle, highlighted bandwidth is the one which minimizes the cross-validation criteria and the first and third rows correspond respectively to one-half and twice the middle bandwidth. The last row reports results from a fully parametric analysis where the threshold is obtained via a linear regression of the estimated muhat on the estimated phat and its square evaluated at phat equals a half. The last column reports a 2-sided p-value corresponding to a test of zero effect.

Table 3B. Thresholds by School-type

Outcome mean=61.54, std dev=5.2

Method	Indep thld	State thld	Ind-State	p-value
Scale=0.5	60.21	58.61	1.6	0.08
Scale=1.00	58.55	56.84	1.71	0.05
Scale=2.00	59.44	57.78	1.66	0.04
Parametric	60.34	58.7	1.64	0.05

Note: This table presents the estimated admission thresholds for expected performance by school-type. These thresholds are calculated via equation (5) in the text where mu-hat and p-hat are estimated via linear regression and probit respectively and the threshold is obtained via a nonparametric regression of the estimated muhat on the estimated phat evaluated at phat equals one-half. Each of the first three rows corresponds to a different choice of bandwidth. The middle, highlighted bandwidth is the one which minimizes the cross-validation criteria and the first and third rows correspond respectively to one-half and twice the middle bandwidth. The last row reports results from a fully parametric analysis where the threshold is obtained via a linear regression of the estimated muhat on the estimated phat and its square evaluated at phat equals a half. The last column reports a 2-sided p-value corresponding to a test of zero effect.

Table 4A. Other outcomes by Gender

Outcome	Male-thld	Fem-thld	Male-Fem	p-value
60+ (mean 0.52)	0.5	0.2	0.3	0.02
55+ (mean 0.78)	0.78	0.55	0.23	0.06
Avg (mean 61.54)	59.36	55.67	3.69	0.0004

Note: This table presents the estimated admission thresholds for expected performance by gender. Three different measures of performance are considered, viz., securing at least a high second class mark (60+), at least a second class mark (55+) and the actual score out of 100 (avg.). The mean of each performance measure across the entire sample is reported in parantheses. The thresholds are calculated via equation (5) in the text where mu-hat and p-hat are estimated via linear regression and probit respectively and the threshold is obtained via a nonparametric regression of the estimated muhat on the estimated phat evaluated at phat equals one-half. The optimal bandwidth is used. The last column reports a 2-sided p-value corresponding to a test of zero effect.

Table 4B. Other outcomes by School-type

Outcome	Indep-thld	State-thld	Indep-State	p-value
60+ (mean 0.52)	0.58	0.35	0.23	0.24
55+ (mean 0.78)	0.62	0.71	-0.09	0.65
Avg (mean 61.54)	58.55	56.84	1.71	0.03

Note: This table presents the estimated admission thresholds for expected performance by school-type. Three different measures of performance are considered, viz., securing at least a high second class mark (60+), at least a second class mark (55+) and the actual score out of 100 (avg.). The mean of each performance measure across the entire sample is reported in parantheses. The thresholds are calculated via equation (5) in the text where mu-hat and p-hat are estimated via linear regression and probit respectively and the threshold is obtained via a nonparametric regression of the estimated muhat on the estimated phat evaluated at phat equals one-half. The optimal bandwidth is used. The last column reports a 2-sided p-value corresponding to a test of zero effect.

References

[1] Ahn, H. & J.L. Powell (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism, Journal of Econometrics, 58, 3-29.

[2] Antonovics, K.L. & B.G. Knight, (2009) A New Look at Racial Profiling: Evidence from the Boston Police Department, Review of Economics and Statistics, 91, 163-177.

[3] Anwar, S & H. Fang (2011) Testing for the role of prejudice in emergency departments using bounceback rates, NBER Working Paper 16888.

[4] Arcidiacono, P. (2005) Affirmative Action in Higher Education: How do Admission and Financial Aid Rules Affect Future Earnings?, Econometrica, 73-5, 1477-1524.

[5] Arcidiacono, P., E. Aucejo, H. Fang, & K. Spenner (2011) Does Affirmative Action Lead to Mismatch? A New Test and Evidence, Quantitative Economics, 2-3, 303-333.

[6] Arcidiacono, P, E. Aucejo & K. Spenner (2011) What Happens After Enrollment? An Analysis of the Time Path of Racial Differences in GPA and Major Choice?, working paper, Duke University.

[7] Becker, G. (1957) The economics of discrimination, University of Chicago Press.

[8] Bertrand, M. & S. 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.

[9] Bertrand, M., R. Hanna & 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.

[10] 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.

[11] Bhattacharya, D. & P. Dupas (2010) Inferring Efficient Treatment Assignment under Budget Constraints, forthcoming, Journal of Econometrics.

[12] Bhattacharya, D. (2011) Evaluating Treatment Protocols Using Data Combination, Mimeo, University of Oxford.

[13] Bosq, D (1998) Nonparametric Statistics for Stochastic Processes, 2nd Ed., Springer-Verlag.

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

[15] Bradley. R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions, Probability Surveys 2, 107-144.

[16] Brock, W.A., J. Cooley, S. Durlauf & S. Navarro (2011) On the Observational Implications of Taste-Based Discrimination in Racial Profiling, forthcoming, Journal of Econometrics.

[17] Card, D. & A.B. Krueger (2005) Would The Elimination Of Affirmative Action Affect Highly Qualified Minority Applicants? Evidence From California And Texas, Industrial and Labor Relations Review, 58-3, 416-434.

[18] Carneiro, P., J.J. Heckman & E.J. Vytlacil (2011) Evaluating marginal policy changes and the average effect treatment for individuals at the margin, NBER Working Paper 15211.

[19] Carrell, S., B.I. Sacerdote & J.E. West (2011) From Natural Variation to Optimal Policy? The Lucas Critique Meets Peer Effects, NBER Working Paper 16865.

[20] Chandra, A. & D. Staiger (2009) Identifying provider prejudice in medical care, Mimeo, Harvard University and Dartmouth College.

[21] Davidson, J. (1994) Stochastic Limit Theory, Oxford University Press.

[22] Fang, H. & A. Moro (2008) Theories of Statistical Discrimination and Affirmative Action: A Survey, NBER Working Paper 15860.

[23] Fryer Jr., R.G. & G.C. Loury (2005) Affirmative Action and Its Mythology, Journal of Economic Perspectives, 19-3, 147-162.

[24] Fryer, R.G., G.C. Loury & T. Yuret (1996) Color-Blind Affirmative Action, NBER Working Paper 10103.

[25] Gospodinov, N. & M. Hirukawa (2012) Nonparametric Estimation of Scalar Diffusion Models of Interest Rates Using Asymmetric Kernels, forthcoming in Journal of Empirical Finance.

[26] Graddy, K. & 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, 435-451.

[27] Graham, B.S. (2011) Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers, Handbook of Social Economics 1B: 965 - 1052 (J. Benhabib, A. Bisin, & M. Jackson, Eds.), Amsterdam: North-Holland.

[28] Grogger, J. & G. Ridgeway (2006) Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness, Journal of the American Statistical Association, 101, 878-887.

[29] Hansen, B.E. (2008) Uniform convergence rates for kernel estimation with dependent data, Econometric Theory, 24, 726-748.

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

[31] Holzer, H.J. & D. Neumark (2000) What Does Affirmative Action Do?, Industrial and Labor Relations Review, 53-2, 240-271.

[32] Hoxby, C.M. (2009) The Changing Selectivity of American Colleges, Journal of Economic Perspectives, American Economic Association, 23-4, 95-118.

[33] Jiang, W., R. Nelson & E. Vytlacil (2011): Nonparametric Identification and Estimation of a Binary Choice Model of Loan Approval Using Only Approved Loans, Working Paper, Yale University.

[34] Kanaya, S. (2012) Uniform convergence rates of kernel-based nonparametric estimators for diffusion processes: A damping function approach, Working Paper, University of Oxford.

[35] Kane, T. J. & W.T. William (1998) Racial and Ethnic Preference in College Admissions, in Christopher Jencks and Meredith Phillips (eds.), The Black-White Test Score Gap, Washington: Brookings Institution.

[36] Keith, S., R.M. Bell, A.G. Swanson & A.P. Williams (1985) Effects of Affirmative Action in Medical Schools – A Study of the Class of 1975, The New England Journal of Medicine, 313, 1519-1525.

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

[38] Kobrin, J.L., B.F. Patterson, E.J. Shaw, K.D. Mattern & S.M. Barbuti (2008) Validity of the SAT for Predicting First-year College Grade Point Average, College Board, New York.

[39] Kristensen, D. (2009) Uniform Convergence Rates of Kernel Estimators with Heterogeneous, Dependent Data, Econometric Theory 25, 1433-1445.

[40] Kuncel, N. R., S.A. Hezlett & D.S. Ones (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.

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

[42] Mammen, E., C. Rothe & M. Schienle (2011) Nonparametric Regression with Nonparametrically Generated Covariates, forthcoming in Annals of Statistics.

[43] Manski, C. (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice, Journal of Econometrics, 3-3, 205-228.

[44] Manski, C. (1988) Identification of Binary Response Models, Journal of the American Statistical Association, 83. 729-738.

[45] Manski, C. (2004) Statistical Treatment Rules for Heterogeneous Populations, Econometrica, 72-4, 1221-1246.

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

[47] Ogg , T., A. Zimdars & A. 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, 35-5.

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

[49] Parks, G. (2011) Academic Performance of International Baccalaureate Students at Cambridge by School, available online at:
http://www.admin.cam.ac.uk/offices/admissions/research/docs/ib_performance.pdf

[50] Persico, N (2009) Racial Profiling? Detecting Bias Using Statistical Evidence. Annual Review of Economics, 1, 229-254.

[51] Rilstone, P. (1996) Nonparametric Estimation of Models with Generated Regressors, International Economic Review 37, 299-313.

[52] Rothstein, J. (2004) College Performance Predictions and the SAT, Journal of Econometrics, 121, 297-317.

[53] Sackett, P., N. Kuncel, J. Arneson, G. Cooper & S. Waters (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

[54] 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

[55] Sperlich, S. (2009) A Note on Nonparametric Estimation with Predicted Variables, Econometrics Journal 12, 382-395.

[56] Zimdars, A. (2010) Fairness and undergraduate admission: a qualitative exploration of admissions choices at the University of Oxford, Oxford Review of Education. 36-3, 307-323.

[57] Zimdars, A., A. Sullivan & A. Heath (2009) Elite Higher Education Admissions in the Arts and Sciences: Is Cultural Capital the Key?, Sociology, 4, 648-66.

A Technical Appendix

The appendix contains three subsections: subsection A.1 presents the proof of (1) in Proposition 1; subsection A.2 formally states and derives the asymptotic distribution of the semiparametric estimator of γ_g, on which our application is based and, finally, subsection A.3 states and derives the distribution theory for the fully nonparametric estimator of γ_g.

A.1 Proof of Proposition 1

Consider any feasible rule $p(\cdot)$ satisfying the budget constraint. Since $p^{opt}(\cdot)$ satisfies the budget constraint with equality (recall the definition of γ and q) and $p(\cdot)$ is feasible, we must have

$$\int_{w\in\mathcal{W}} \alpha(w)\, p^{opt}(w)\, dF_W(w) = c \geq \int_{w\in\mathcal{W}} \alpha(w)\, p(w)\, dF_W(w), \tag{7}$$

implying that

$$\int_{w\in\mathcal{W}} \alpha(w) \left[p^{opt}(w) - p(w)\right] dF_W(w) \geq 0. \tag{8}$$

Let $\mathbb{W}(p) := \int_{w\in\mathcal{W}} p(w)\,\alpha(w)\,\beta(w)\, dF_W(w)$. Now, the productivity resulting from $p(\cdot)$ differs from that from $p^{opt}(\cdot)$ by

$$\begin{aligned}
& \mathbb{W}\left(p^{opt}\right) - \mathbb{W}(p) \\
= & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) + \gamma \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w)\, dF_W(w) \\
\geq & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= & \int_{\beta(w)>\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
& + \int_{\beta(w)<\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= & \int_{\beta(w)>\gamma} \left[1 - p(w)\right] \left[\beta(w) - \gamma\right] \alpha(w)\, dF_W(w) + \int_{\beta(w)<\gamma} p(w) \left[\gamma - \beta(w)\right] \alpha(w)\, dF_W(w) \geq 0,
\end{aligned} \tag{9}$$

where the first inequality holds by (8) and that $\gamma > 0$. Therefore, we have $\mathbb{W}\left(p^{opt}\right) \geq \mathbb{W}(p)$ for any feasible $p(\cdot)$, and the solution $p^{opt}(\cdot)$ given in (1) is optimal.

To show the uniqueness, consider any feasible rule $p(\cdot)$ which differs from $p^{opt}(\cdot)$ on some set whose measure is not zero, i.e., $\int_{w\in\mathbf{S}(p)} dF_W(w) > 0$ for $\mathbf{S}(p) := \{w \in \mathcal{W} \mid p^{opt}(w) \neq p(w)\}$. Now, assume that $\mathbb{W}\left(p^{opt}\right) = \mathbb{W}(p)$ for this $p(\cdot)$. In this case, since the last equality on the RHS of (9) holds with equality, $p(\cdot)$ must take the following form:

$$p(w) = \begin{cases} 1 & \text{if } \beta(w) > \gamma; \\ 0 & \text{if } \beta(w) < \gamma, \end{cases}$$

for almost every w (with respect to F_W). This implies that $p(w) = p^{opt}(w)$ for almost every w except when $\beta(w) = \gamma$. Since the measure of $\mathbf{S}(p)$ is not zero, we must have $p^{opt}(w) \neq p(w)$ for

$\beta(w) = \gamma$, and $\mathbf{S}(p) = \{w \in \mathcal{W} \mid \beta(w) = \gamma\}$, which, together with the budget constraint, implies that $q > p(w)$ when $\beta(w) = \gamma$. However, this in turn implies that we have a strict inequality in the third line on the RHS of (9), which contradicts our assumption. Therefore, we now have shown that $\mathbb{W}\left(p^{opt}\right) > \mathbb{W}(p)$ for any feasible $p(\cdot)$ with $\int_{w \in \mathbf{S}(p)} dF_W(w) > 0$, leading to the desired uniqueness property of $p^{opt}(\cdot)$ in the stated sense. ■

A.2 Asymptotic results for the semiparametric case

In this and next subsections, we often write $W_i = (X_i, G_i)$ (as defined in Section 3) for notational simplicity. 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 S^d is finite). Note that $\mathcal{W} = S^c \times S^d$ in the notation of previous sections. We let the last one or more components of the vector W_i^d be G_i, denote by S^G the support of G_i (e.g., if we are interested only in the gender difference, $S^G = \{female, male\}$).

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 $\mu^P(x, g) = \mu^P(w)$ or $\mu\left(w^c, w^d\right)$. 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, $||E|| := \max_{1 \leq i \leq I;\ 1 \leq j \leq J} |E_{i.j}|$. And, we often write $z_0 = 1/2$ below in proofs.

As stated previously, our analyses are based on the estimator of the form in (5). However, to consider the semi and non parametric cases separately, we below re-define our estimators. Now, we consider the following semiparametric estimator (while the nonparametric one is presented in the next subsection):

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

where $\bar{p}(w; \theta_p)$ $(= \bar{p}(x, g; \theta_p))$ is a (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; $\bar{\mu}^P(w; \theta_\mu)$ $(= \bar{\mu}^P(x, g; \theta_\mu))$ is a (semi) parametric estimator of $\mu^P(w)$; and θ_μ, $\hat{\theta}_\mu$ and θ_μ^0 are defined analogously. We may use various (semi) parametric models, e.g., a probit or logit model for $p(w)$ and a linear (regression) model for $\mu^P(w)$, whose requirements presented in Assumptions 8 and 9 are quite mild.

Asymptotic behavior of the semiparametric estimator: To investigate the asymptotic properties of (10), we consider the following estimator:

$$\tilde{\gamma}_{sp}(g) := \frac{(1/n) \sum_{i=1}^n K_h\left(\bar{p}\left(W_i; \theta_p^0\right) - 1/2\right) \bar{\mu}^P\left(W_i; \theta_\mu^0\right) \mathbf{1}\{G_i = g\}}{(1/n) \sum_{l=1}^n K_h\left(\bar{p}\left(W_l; \theta_p^0\right) - 1/2\right) \mathbf{1}\{G_l = g\}}. \tag{11}$$

This is not an feasible estimator, requiring (pseudo) true objects $\bar{p}\left(w;\theta_p^0\right)$ and $\bar{\mu}^P\left(w;\theta_\mu^0\right)$. However, we below show that the feasible and infeasible estimators, $\hat{\gamma}_{sp}(g)$ and $\tilde{\gamma}_{sp}(g)$ share the same asymptotic distribution.

To derive the asymptotic distribution of the infeasible estimator $\tilde{\gamma}_{sp}(g)$, we work with the following conditions:

Assumption 5 *Let*

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

For each $g \in S^G$, $m(\cdot,g)$ 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 ($\nu(z,g)\,dz = \Pr[p(W_i) \in dz, G_i = g]$); and for each $g \in S^G$, $\nu(\cdot,g)$ is twice continuously differentiable on $[0,1]$.

Assumption 6 *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 $\int_{\mathbb{R}} u^2 |K(u)|\,du \leq \bar{K}$.*

Assumption 7 *There exist some θ_p^0 and θ_μ^0 such that $\bar{p}(w;\theta_p) = p(w)$ and $\bar{\mu}^P\left(w;\theta_\mu^0\right) = \mu^P(w)$.*

Assumptions 5-6 are standard technical requirements for kernel-based estimation. Note that under Assumptions 1, 2-(ii), 2-(iii) and 4, there exists some constant $\epsilon > 0$ such that

$$\inf_{(z,g)\in[1/2-\epsilon,1/2+\epsilon]\times S^G} \nu(z,g) > 0. \tag{12}$$

Note also that given the correct specification condition in Assumption 7, $\tilde{\gamma}_{sp}(g)$ is identical to the infeasible estimator $\tilde{\gamma}_g$ (defined in (6), Section 5).

Lemma 1 *Suppose that Assumptions 1, 2 and 2-7 hold. Then, it holds that as $n \to \infty$ and $h \to 0$ with $nh \to \infty$ and $nh^5 = O(1)$,*

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

for each $g \in S^G$, where

$$\mathbf{B}(g) := \int_{\mathbb{R}} u^2 K(u)\,du \left[(\partial/\partial z)\, m(z,g)\,(\partial/\partial z)\,\nu(z,g)/\nu(z,g) + (1/2)\left(\partial^2/\partial z^2\right) m(z,g)\right]\Big|_{z=1/2};$$

$$\mathbf{V}(g) := \int_{\mathbb{R}} K^2(u)\,du \mathrm{Var}\left[\mu(X_i,g)\,|p(X_i,g) = z\right]/\nu(z,g)\Big|_{z=1/2}.$$

Given the stated conditions, 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 have supposed correct parametric specifications here, but even when the parametric models are misspecified, the lemma's result in still holds with

slight modification. In such a case, objects in Assumption 5 and the lemma, γ_g, $m(z,g)$ $\nu(z,g)$, $\mathbf{B}(g)$ and $\mathbf{V}(g)$, should be interpreted in terms of pseudo true objects, say, the parameter γ_g should be interpreted as $E[\bar{\mu}^P\left(W_i;\theta_\mu^0\right)|\bar{p}\left(W_i;\theta_p^0\right)=1/2,\, G_i=g]$, rather than the "true" one considered in Sections 4 and 5. Note that the same remark applies to the result in Theorem 1 below.

We now analyze our semiparametric estimator $\hat{\gamma}_{sp}(g)$ under the following conditions:

Assumption 8 *(i) The estimator $\hat{\theta}_p$ is consistent for the (pseudo) true parameter θ_p^0 with*

$$|\hat{\theta}_p-\theta_p^0|=O_{a.s.}(1/\sqrt{n}). \tag{13}$$

(ii) There exists some compact set Θ_p such that θ_p^0 is in the interior of Θ_p; for each $w\in S^c\times S^d$, $\bar{p}(w;\cdot)$ is twice continuously differentiable on Θ_p;

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

Assumption 9 *The estimator $\hat{\theta}_\mu$ is consistent for the (pseudo) true parameter θ_μ^0 with*

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

The condition on $\hat{\theta}_\mu$ in Assumption 9 is fairly weak. We do not presuppose any data generating mechanism on the past cohort data $\{(A_j^P,A_j^PY_j^P,X_j^P,G_j^P)\}$, except for the $\sqrt{n}$-consistency of the function, which should be satisfied by many (semi-)parametric models and estimators. The conditions on $\hat{\theta}_p$ in Assumption 8 are slightly stronger, but are also satisfied in many cases. In particular, for various estimators, (i) of Assumption 2 and some boundedness condition on relevant functions are often sufficient for the strong $\sqrt{n}$-consistency in (13) (see, e.g., the strong law of large numbers as found in Ch. 20 of Davidson, 1994). Assumptions 8 and 9 are respectively satisfied by probit and linear regression models (employed in Section 6).

To show the asymptotic equivalence of $\hat{\gamma}_{sp}(g)$ and $\tilde{\gamma}_{sp}(g)$, we also impose the following condition on $K(\cdot)$.

Assumption 10 *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 *Suppose that Assumption 10 and the same conditions as in Lemma 1 hold. Then, it holds that as $n\to\infty$ and $h\to 0$ with $nh^3\to\infty$,*

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

Therefore, additionally if $nh^5 = O(1)$ (as $n \to \infty$ and $h \to 0$), the asymptotic bias and distribution of $\hat{\gamma}_{sp}(g)$ are the same as those for $\tilde{\gamma}_{sp}(g)$ given in Lemma 1.

Proof. First, consider the convergence of the numerator of (10). We have the following decomposition:

$$\begin{aligned}&(\sqrt{nh}/n)\sum_{i=1}^{n}\left[K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)\bar{\mu}^P\left(W_i;\hat{\theta}_\mu\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\bar{\mu}^P\left(W_i;\theta_\mu^0\right)\right]\\&=\mathcal{A}_n+\mathcal{B}_n+\mathcal{C}_n,\end{aligned}\tag{14}$$

where

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

By Assumption 9, we can easily show that $\mathcal{A}_n = O_P(\sqrt{h})$. To consider the convergence rate of $\mathcal{B}_n$, look at

$$\bar{p}\left(W_i;\hat{\theta}_p\right)-\bar{p}\left(W_i;\theta_p^0\right)=(\partial/\partial\theta_p)\,\bar{p}\left(W_i;\theta_p^0\right)\left[\hat{\theta}_p-\theta_p^0\right]+O_{a.s.}\left(n^{-1}\right)$$

uniformly over i, which follows from the Taylor expansion and Assumption 8. Therefore,

$$\begin{aligned}&K_h\left(\bar{p}\left(W_i;\hat{\theta}_p\right)-z_0\right)-K_h\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)\\=\;&h^{-2}K'((\bar{p}\left(W_i;\theta_p^0\right)-z_0)/h)\left\{(\partial/\partial\theta_p)\,\bar{p}\left(W_i;\theta_p^0\right)\left[\hat{\theta}_p-\theta_p^0\right]+O_{a.s.}\left(n^{-1}\right)\right\}\\&+(1/2)\,h^{-3}K''((\bar{p}\left(W_i;\tilde{\theta}_p\right)-z_0)/h)\times O_{a.s.}\left(n^{-1}\right),\end{aligned}\tag{15}$$

where $\tilde{\theta}_p$ is on the line segment connecting $\hat{\theta}_p$ to θ_p^0 ($\tilde{\theta}_p$ may depend on i) while $O_{a.s.}\left(n^{-1}\right)$s on the RHS are uniform over i. By Assumption 10, there exist some function $\bar{\mathcal{K}}^\star(\cdot)$ and some positive constant $\bar{\varepsilon}\,(>0)$ such that $\sup_{|\varepsilon|\le\bar{\varepsilon}}|K''(u+\varepsilon)| \le \mathcal{K}^\star(u)$ for any $u \in \mathbb{R}$, $\sup_{u\in\mathbb{R}}\mathcal{K}^\star(u) < \infty$ and $\int_{\mathbb{R}}\mathcal{K}^\star(u)\,du < \infty$. Since $\bar{p}\left(W_i;\tilde{\theta}_p\right)/h = \bar{p}\left(W_i;\theta_p^0\right)/h + o_{a.s.}(1)$ uniformly over i, which follows from Assumption 8 and the condition that $\sqrt{n}h \to \infty$, for any n large enough, it almost surely holds that

$$\left|K''\left(\left(\bar{p}\left(W_i;\tilde{\theta}_p\right)-z_0\right)/h\right)\right| \le \mathcal{K}^\star\left(\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)/h\right),\tag{16}$$

uniformly over i.[26] Now, we let

$$\psi_{n,i}:=h^{-2}K'\left(\left(\bar{p}\left(W_i;\theta_p^0\right)-z_0\right)/h\right)\bar{\mu}^P\left(W_i;\theta_\mu^0\right)\mathbf{1}\left\{G_i=g\right\}(\partial/\partial\theta_p)\,\bar{p}\left(W_i;\theta_p^0\right).$$

[26] This is because it holds that for each $\omega \in \Omega^*$ (Ω^* is an event with $\Pr(\Omega^*)=1$) and for any n large enough, $|\bar{p}\left(W_i;\tilde{\theta}_p\right)/h - \bar{p}\left(W_i;\theta_p^0\right)/h| \le \bar{\varepsilon}$.

Then, by using (15) and (16) and noting the uniform boundedness of the function $\bar{\mu}^P\left(\cdot;\theta_\mu^0\right)$, we can consider the following bound:

$$\begin{aligned}|\mathcal{B}_n| \quad \leq \quad & \sqrt{nh}||E\left[\psi_{n,i}\right]|| \times ||\hat{\theta}_p - \theta_p^0|| + \sqrt{nh}||n^{-1}\textstyle\sum_{i=1}^n \left\{\psi_{n,i} - E\left[\psi_{n,i}\right]\right\}|| \times ||\hat{\theta}_p - \theta_p^0|| \\ & + (\sqrt{nh}/nh^3)\textstyle\sum_{i=1}^n \mathcal{K}^\star\left(\left(\bar{p}\left(W_i;\theta_p^0\right) - z_0\right)/h\right) \times O_P\left(n^{-1}\right). \end{aligned} \tag{17}$$

The first term on the RHS of (17) is $O_P(\sqrt{h})$, since $||E\left[\psi_{n,i}\right]|| = O(1)$, which follows from the standard change-of-variable arguments for kernel-based estimators and the uniform boundedness of relevant functions. The second term on the RHS of (17) is $O_P(1/\sqrt{nh})$ since $n^{-1}\sum_{i=1}^n \left\{\psi_{n,i} - E\left[\psi_{n,i}\right]\right\} = O_P(1/\sqrt{nh^2})$, which can be obtained by standard arguments for kernel-based estimation of derivatives (as those in Theorem 6 and its proof of Hansen, 2008). Finally, the last term of the RHS of (17) is $O_P(1/\sqrt{nh^3})$, since we have $E[\mathcal{K}^\star\left(\left(\bar{p}\left(W_i;\theta_p^0\right) - z_0\right)/h\right)] = O(h)$ uniformly over i, which follows from the standard change-of-variable arguments and the kernel-like property of $\mathcal{K}^\star$ stated above. We now have shown that

$$\mathcal{B}_n = O_P(\sqrt{h}) + O_P(1/\sqrt{nh}) + O_P(1/\sqrt{nh^3}) = O_P(\sqrt{h} + 1/\sqrt{nh^3}).$$

We can easily show that $\mathcal{C}_n = O_P(1/\sqrt{nh^3})$ by using (15) and Assumption 9, and omit details. From the expression (14) and arguments above, we can see that the scaled version of the numerator of (10) can be written as

$$(\sqrt{nh}/n)\sum\nolimits_{i=1}^n K_h\left(\bar{p}\left(W_i;\theta_p^0\right) - z_0\right)\bar{\mu}^P\left(W_i;\theta_\mu^0\right) + O_P(\sqrt{h} + 1/\sqrt{nh^3}). \tag{18}$$

By arguments analogous to those for $\mathcal{B}_n$, we can also write the denominator of (10) as

$$(1/n)\sum\nolimits_{l=1}^n K_h\left(\bar{p}\left(W_i;\theta_p^0\right) - z_0\right)\mathbf{1}\left\{G_l = g\right\} + O_P(\sqrt{h}) + O_P(1/\sqrt{nh^3}),$$

which, together with (18), leads to the desired result. ■

A.3 Asymptotic results for the fully nonparametric case

To consider the nonparametric case, we explicitly present the forms of our first-step estimators of $p(w)$ and $\mu^P(w)$. For the estimation of $\mu^P(w)$, recall that we have assumed the availability of the past cohort data $\{(A_j^P, A_j^P Y_j^P, X_j^P, G_j^P)\}_{j=1}^N$ in Section 5, where we let $n = N$ for simplicity. For a variable from the past cohort, we write $W_j^P = (W_j^{P,c}, W_j^{P,d}) = (X_j^P, G_j^P)$ in the same manner as for one from the current cohort (as explained in the previous subsection).

Now, our nonparametric estimator of γ_g is defined as:

$$\hat{\gamma}_{np}(g) := \frac{(1/n)\sum_{i=1}^n K_h\left(\hat{p}_{-i}(W_i) - 1/2\right)\hat{\mu}_{-i}^P(W_i)\mathbf{1}\left\{G_i = g\right\}}{(1/n)\sum_{l=1}^n K_h\left(\hat{p}_{-l}(W_l) - 1/2\right)\mathbf{1}\left\{G_l = g\right\}}, \tag{19}$$

where $\hat{p}_{-i}(w)$ is a so-called leave-one-out nonparametric estimator of $p(w)$ and $\mu^P(w)$ is an estimator of the Nadaraya-Watson type as follows:

$$\hat{p}_{-i}(w) := \frac{\sum_{1\leq j\leq n;\ j\neq i} L_{\xi_p}\left(W_j^c - w^c\right)\mathbf{1}\left\{W_j^d = w^d\right\} D_j}{\sum_{1\leq k\leq n;\ k\neq i} L_{\xi_p}\left(W_k^c - w^c\right)\mathbf{1}\left\{W_k^d = w^d\right\}}; \tag{20}$$

$$\hat{\mu}^P(w) := \frac{\sum_{j=1}^n M_{\xi_\mu}\left(W_j^{P,c} - w^c\right)\mathbf{1}\left\{W_j^{P,d} = w^d\right\} Y_j^P A_j^P}{\sum_{k=1}^n M_{\xi_\mu}\left(W_k^{P,c} - w^c\right)\mathbf{1}\left\{W_k^{P,d} = w^d\right\} A_j^P}; \tag{21}$$

$L_a(z) := L(z/a)/a^{d_1} = L(z_1/a,\ldots,z_{d_1}/a)/a^{d_1}$ 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 ($\mathbb{R}^{d_1}\to\mathbb{R}$) and ξ_μ is another bandwidth.

Remark 7 *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 (20) is replaced by $L\left(\Xi_p^{-1}\left(W_j^c - w^c\right)\right)/\det(\Xi_p)$, where $\det(H)$ is the determinant of H (an analogous argument applies to (21)).*

Remark 8 *The suggested estimators (19), (20) and (21) 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 $\tilde{\gamma}_g$ (recall that $\tilde{\gamma}_g = \tilde{\gamma}_{sp}(g)$ under Assumption 71 and the asymptotic property of $\tilde{\gamma}_{sp}(g)$ is given in Lemma 1). For this purpose, we work with the following conditions:

Assumption 11 *There exists the probability function of $W_i(=(W_i^c, W_i^d) = (X_i, G_i))$, i.e., a function $f(w)\,(=f(w^c,w^d) = f(x,g))$ satisfying $f(w^c,w^d)\,dw^c = \Pr\left[W_i^c\in dw^c, W_i^d = w^d\right]$. For each $w^d\in S^d$, the functions $p(\cdot,w^d)$ and $f(\cdot,w^d)$ are compactly supported on S^c. Let κ_p be some positive integer with $\kappa_p\geq 2$. For each $w^d\in S^d$, $p(\cdot,w^d)$ and $f(\cdot,w^d)$ are κ_p-times continuously differentiable on S^c.*

Assumption 12 *There exists the probability function $q(w)$ of $(W_j^P, A_j^P) = (W_j^{P,c}, W_j^{P,d}, A_j^P)$ for $A_j^P = 1$, i.e., a function $q(w)\,(=q(w^c,w^d) = q(x,g))$ satisfying $q(w^c,w^d)\,dw^c = \Pr[W_j^{P,c}\in dw^c, W_j^{P,d} = w^d, A_j^P = 1]$. For each $w^d\in S^d$, the functions $\mu^P(\cdot,w^d)$ and $q(\cdot,w^d)$ are compactly*

supported on S^c. *Let* κ_μ *be some positive integer with* $\kappa_\mu \geq 2$. *For each* $w^d \in S^d$, $\mu^P\left(\cdot, w^d\right)$ *and* $q\left(\cdot, w^d\right)$ *are* κ_μ*-times continuously differentiable on* S^c.

Assumption 13 *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_p*th-order kernel, i.e.,* $\int_{\mathbb{R}^{d_1}} [\bigotimes_{l=1}^{k} u] L(u)\,du = 0$ *for* $k = 1, \ldots, (\kappa_p - 1)$.

Assumption 14 *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_μ*th-order kernel, i.e.,* $\int_{\mathbb{R}^{d_1}} [\bigotimes_{l=1}^{k} u] M(u)\,du = 0$ *for* $k = 1, \ldots, (\kappa_\mu - 1)$.

Assumption 15 *(i) There exists some constant* $C_1 \in (0, \infty)$ *such that*

$$C_1 \leq \inf_{\left(w^c, w^d\right) \in S^c \times S^d} f\left(w^c, w^d\right) \quad \text{and} \quad C_1 \leq \inf_{\left(w^c, w^d\right) \in S_\circ^c \times S^d} q\left(w^c, w^d\right),$$

where f *and* q *are the probability functions defined in Assumptions 11 and 12, respectively. (ii) There exists some set* $S_\circ^c$ *such that if* $D_i = 1$, *then*

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

where any boundary points of $S_\circ^c$ *are in the interior of* S^c.

Assumption 16 *(i)* $\{\varepsilon_i\}$ *is geometrically* α*-mixing (i.e.,* $a_m \leq \tilde{A} \exp\left\{-\tilde{b} m\right\}$ *for some positive constants* $\tilde{A}$ *and* $\tilde{b}$*). (ii)* $\{(A_j^P, A_j^P Y_j^P, W_j^P)\}_{j=1}^n$ *is first-order stationary and geometrically* α*-mixing, and there exists some constant such that* $|Y_j^P| \leq C_2$. *(iii)* (D_i, W_i) *is independent of* $(A_j^P, A_j^P Y_j^P, W_j^P)$ *for any* i *and* j.

Assumptions 11-14 are quite standard for establishing uniform convergence results for $\hat{p}_{-i}(w)$ and $\hat{\mu}^P(w)$ (see Lemmas 2 and 3 below). Assumptions 13-14 require that the kernels, L and M, are of higher order (bias reducing) of orders κ_p and κ_μ, respectively. These, together with the differentiability conditions in Assumptions 11-12, are used to guarantee that the estimation errors due to the first step are negligible in the second step.

We impose Assumption 15 to avoid the so-called boundary-bias problem. Our first-step nonparametric estimators $\hat{p}_{-i}(w)$ and $\hat{\mu}^P(w)$ are of the Nadaraya-Watson type (with symmetric kernel functions), and have slower uniform convergence rates around the boundary points of the support (see, e.g., Bouezmarni and Scaillet, 2005).[27] (ii) of Assumption 15 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

[27] The boundary bias may be avoided by using asymmetric kernels as in Bouezmarni and Scaillet (2005) and Gospodinov and Hirukawa (2012), or by usimg the local polynomial method as in Masry (1996).

are imposed only for simplicity. We may be able to proceed without (i) and/or (ii) of Assumption 15. However, to do so, we will require a trimming device and more intricate conditions on the bandwidths and trimming parameters.

The conditions in Assumption 16 control for the data dependence structure. The geometric mixing conditions in (i) and (ii) allows us to derive sharp convergence rates of the first-step estimators. We can relax these as in Hansen (2008), Kristensen (2009) and Kanaya (2012), allowing for polynomial mixing cases (with relatively strong dependence of sequences). However, we consider only the geometric case for simplicity, where we can work with less complicated restrictions on bandwidth choices.

Given these conditions, we obtain the following result:

Theorem 2 *Suppose that Assumptions 1, 2, 2-6 and 11-16 hold. Let*

$$\begin{aligned}\Delta_n \ :\ &= \sqrt{(\log n)^2/n\xi_\mu^{d_1}} + \sqrt{h} + \sqrt{nh}[\xi_\mu^{\kappa_\mu} + \xi_\mu\sqrt{(\log n)/n\xi_\mu^{d_1}}] \\ &+(1/\sqrt{nh^{5/2}\xi_p^{d_1}}) + \sqrt{nh}\xi_p^{\kappa_p} + \xi_p\sqrt{1/nh^2\xi_p^{d_1}} + (\xi_p^{\kappa_p}/h) + \sqrt{n/h^3}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2 \\ &+\sqrt{n/h}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}][\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}].\end{aligned}$$

It holds that as $n \to \infty$*, and* h, ξ_p *and* $\xi_\mu \to 0$ *with* $[\log(\log n)]^4 (\log n)^2/n\xi_p^{d_1} \to 0$*,* $(\log n)/nh^2\xi_p^{d_1} \to 0$*, and* $[\log(\log n)]^4 (\log n)^2/nh\xi_\mu^{d_1} \to 0$*,*

$$\sqrt{nh}[\hat{\gamma}_{np}(g) - \tilde{\gamma}_g] = O_P(\Delta_n) \quad \textit{for each } g \in S_G,$$

Therefore, additionally if $nh^5 = O(1)$ *and* $\Delta_n \to 0$*, the asymptotic bias and distribution of* $\hat{\gamma}_{np}(g)$ *are the same as those for* $\tilde{\gamma}_g (= \tilde{\gamma}_{sp}(g))$ *given in Lemma 1.*

While the rate conditions of the bandwidths h, ξ_p and ξ_μ for the asymptotic equivalence between $\hat{\gamma}_{np}(g)$ and $\tilde{\gamma}_g$ may look somewhat complicated, they can be easily satisfied when κ_p and κ_μ are large relatively to d_1 (i.e., the orders of the kernel functions are high enough and the relevant functions are sufficiently smooth). As an example, consider $h = O(1/n^{1/5}(\log n))$, which is slightly oversmooth as we did in our empirical application. In this case, if we set $\xi_p^{\kappa_p} = o(1/n^{2/5}\sqrt{\log n})$ and $\xi_\mu^{\kappa_\mu} = o(1/n^{2/5}\sqrt{\log n})$ with

$$(\log n)^3/n^{1/5}\xi_p^{d_1} \to 0; \quad 1/n^{3/5}\xi_\mu^{d_1} \to 0; \quad \text{and} \quad (\log n)^2/n^{1/5}\xi_\mu^{d_1-2} \to 0 \tag{22}$$

all the bandwidth conditions of the theorem are satisfied. As apparent from (22), we need more restrictive conditions on the shrinking rate of ξ_p than on that of ξ_μ (κ_p need to be larger). This is because the estimator $\hat{p}_{-i}(w)$ is in the inside of the kernel function K and it need to have a faster convergence rate than $\hat{\mu}^P(w)$.

To prove the above theorem, we will utilize the following two lemmas, which derive (so-called) uniform Bahadur representations and convergence rates of the first-step nonparametric estimators:

Lemma 2 *Suppose that Assumptions 1, 2, 2-6, 11, 13, 15 and 16-(i) 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*

$$\begin{aligned}\hat{p}_{-i}(w) - p(w) &= (1/n) \sum\nolimits_{1 \le j \le n;\ j \ne i} L_{\xi_p}\left(W_j^c - w^c\right) \mathbf{1}\{W_j^d = w^d\} \left[D_j - p(W_j)\right] / f(w) \\ &\quad + (1/n) \sum\nolimits_{1 \le j \le n;\ j \ne i} L_{\xi_p}\left(W_j^c - w^c\right) \mathbf{1}\{W_j^d = w^d\} \left[p(W_j) - p(w)\right] / f(w) \\ &\quad + O_{a.s.}([\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2); \end{aligned} \tag{23}$$

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

$$\hat{p}_{-i}(w) - p(w) = O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}), \tag{24}$$

uniformly over $i \in \{1, \ldots, n\}$ and $w \in S^c \times S^d$, where $f(w)$ is the probability function defined in Assumption 11.

Lemma 3 *Suppose that Assumptions 12, 14, 15 and 16-(ii) 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*

$$\begin{aligned}\hat{\mu}^P(w) - \mu^P(w) &= (1/n) \sum\nolimits_{j=1}^n M_{\xi_\mu}(W_j^{P,c} - w^c) \mathbf{1}\{W_j^{P,d} = w^d\} A_j^P [Y_j^P - \mu^P(w)] / q(w) \\ &\quad + O_P(\xi_\mu^{\kappa_\mu} + \xi_\mu \sqrt{(\log n)/n\xi_\mu^{d_1}}); \end{aligned} \tag{25}$$

$$\hat{\mu}^P(w) - \mu^P(w) = O_P(\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}), \tag{26}$$

uniformly over $w \in S_\circ^c \times S^d$, where $q(w)$ is the probability function defined in Assumption 12.

The proofs of these lemmas are provided below.[28] Now, we start the proof of Theorem 2.

Proof of Theorem 2. First, we look at the denominator of (19). By applying the mean-value theorem,

$$(1/n) \sum\nolimits_{l=1}^n K_h(\hat{p}_{-l}(W_l) - z_0) \mathbf{1}\{G_l = g\} = \mathbf{I}_n + \mathbf{J}_n,$$

where

$$\begin{aligned}\mathbf{I}_n &: = (1/nh) \sum\nolimits_{l=1}^n K((p(W_l) - z_0)/h) \mathbf{1}\{G_l = g\}; \\ \mathbf{J}_n &: = \left(1/nh^2\right) \sum\nolimits_{l=1}^n K'((\check{p}_{-l}(W_l) - z_0)/h) \mathbf{1}\{G_l = g\} [\hat{p}_{-l}(W_l) - p(W_l)]; \end{aligned}$$

[28] 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 $K_h(p(W_i) - 1/2)$ and $K_h(\hat{p}_{-i}(W_i) - 1/2)$ 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. For these technical points, see the arguments in deriving (27). Note also that except for (24), 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 15. We may be able to derive uniform rates over $S^c \times S^d$. However, under the compact-support condition of K and the exogenous trimming condition ((ii) of Assumption 15), the uniform results over $S_\circ^c$ are sufficient for our purpose.

and $\check{p}_{-l}(W_l)$ is on the line segment connecting $\hat{p}_{-l}(W_l)$ to $p(W_l)$. We below find the probability bounds of $\mathbf{I}_n$ and $\mathbf{J}_n$. Now, by standard arguments for kernel-based estimators (see, e.g., Ch. 3 of Li and Racine, 2007), we can show that $\mathbf{I}_n = \nu(z_0, g) + O_P(h^2 + 1/\sqrt{nh})$. To find the bound of $\mathbf{J}_n$, we use arguments analogous to those for (16) in the proof of Theorem 1. That is, we can find some kernel-like dominant function $\mathcal{K}^*(\cdot)$ for $K'(\cdot)$ (as $\mathcal{K}^\star(\cdot)$ for $K''(\cdot)$) and use this function, to obtain

$$\begin{aligned}\mathbf{J}_n &\leq \left(1/nh^2\right)\sum_{l=1}^n \mathcal{K}^*((p(W_l) - z_0)/h) \times \max_{1\leq l\leq n} \sup_{w\in S^c\times S^d} |\hat{p}_{-l}(w) - p(w)| \\ &= O_P(1/h) \times O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}). \end{aligned} \tag{27}$$

Now, we have shown that

$$\begin{aligned} & (1/n)\sum_{l=1}^n K_h(\hat{p}_{-l}(W_l) - z_0)\mathbf{1}\{G_l = g\} \\ = \;& \nu(z_0, g) + O_P(h^2 + 1/\sqrt{nh} + (1/h)[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]), \end{aligned} \tag{28}$$

where the reminder term is $o_P(1)$ under the stated conditions on the bandwidths.

Next, we look at the numerator of (19):

$$(\sqrt{nh}/n)\sum_{i=1}^n K_h(\hat{p}(W_i) - z_0)\hat{\mu}^P(W_i) = \mathbf{A}_n + \mathbf{B}_n + \mathbf{C}_n, \tag{29}$$

where

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

We can show the following results:

$$\mathbf{A}_n = O_P(\sqrt{(\log n)^2/n\xi_\mu^{d_1}} + \sqrt{h} + \sqrt{nh}[\xi_\mu^{\kappa_\mu} + \xi_\mu\sqrt{(\log n)/n\xi_\mu^{d_1}}]); \tag{30}$$

$$\begin{aligned}\mathbf{B}_n &= O_P(\sqrt{h} + (1/\sqrt{nh^{5/2}\xi_p^{d_1}}) + \sqrt{nh}\xi_p^{\kappa_p}) \\ &\quad + O_P(\xi_p\sqrt{1/nh^2\xi_p^{d_1}} + (\xi_p^{\kappa_p}/h) + \sqrt{n/h^3}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2); \end{aligned} \tag{31}$$

$$\mathbf{C}_n = O_P(\sqrt{n/h}[\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}][\xi_\mu^{\kappa_\mu} + \sqrt{(\log n)/n\xi_\mu^{d_1}}]), \tag{32}$$

whose proofs are provided below. Now, by the results (28)-32 and the boundedness condition of $v(z, g)$ (stated in (12)), we can obtain the conclusion of the theorem.

The convergence rates of the term $\mathbf{A}_n$ in (30). Recall that if $W_i^c \notin S_\circ^c$, then $D_i = 0$ and $p(W_i) = 0$ (by (ii) of Assumption 15). In this case, we have

$$K_h(p(W_i) - z_0) = h^{-1}K(-z_0/h) = 0 \;\text{ for } h(>0) \text{ small enough}, \tag{33}$$

since the support of K is bounded and $-z_0/h(=-1/2h)$ is large enough, and thus, for h small enough, we can restrict our attention to the case $W_i^c \in S_\circ^c$. Therefore, by using (25) in Lemma 3, we can obtain the following expression:

$$\begin{aligned}\mathbf{A}_n &= \mathbf{A}_{n,1} + \mathbf{A}_{n,2} \\ &\quad + (\sqrt{nh}/n) \sum_{i=1}^n K_h\left(p\left(W_i\right) - z_0\right) \mathbf{1}\left\{G_i = g\right\} \times O_P(\xi_\mu^{\kappa_\mu} + \xi_\mu \sqrt{\left(\log n\right)/n\xi_\mu^{d_1}}), \end{aligned} \tag{34}$$

where

$$\begin{aligned}\mathbf{A}_{n,1} &:= (\sqrt{nh}/n) \sum_{j=1}^n A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] n^{-1} \sum_{i=1}^n \left[a_{n,i}(W_j^P) - \bar{a}_n(W_j^P)\right]; \\ \mathbf{A}_{n,2} &:= (\sqrt{nh}/n) \sum_{j=1}^n A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] \bar{a}_n(W_j^P); \\ a_{n,i}\left(w\right) &:= K_h\left(p\left(W_i\right) - z_0\right) M_{\xi_\mu}\left(w^c - W_i^c\right) \mathbf{1}\{w^d = W_i^d\}/q\left(W_i\right); \quad \bar{a}_n\left(w\right) := E\left[a_{n,i}\left(w\right)\right]. \end{aligned}$$

We below derive the convergence rates of the three terms on the RHS of (34).

To find the rate of $\mathbf{A}_{n,1}$, let

$$\begin{aligned}\psi_n\left(w\right) &:= (1/n) \sum_{i=1}^n \left[a_{n,i}\left(W_j^P\right) - \bar{a}_n\left(W_j^P\right)\right]; \\ \bar{\psi}_n\left(w\right) &:= \psi_n\left(w\right) \times \mathbf{1}\{\psi_n\left(w\right) \leq \sqrt{\left(\log n\right)^2/nh\xi_\mu^{d_1}}\}. \end{aligned}$$

Then, we can write

$$\begin{aligned}\mathbf{A}_{n,1} &= (\sqrt{nh}/n) \sum_{j=1}^n A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] \psi_n(W_j^P) \\ &+ (\sqrt{nh}/n) \sum_{j=1}^n A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] \bar{\psi}_n(W_j^P) \\ &+ (\sqrt{nh}/n) \sum_{j=1}^n A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] \psi_n(W_j^P) \mathbf{1}\left\{\psi_n(W_j^P) > \sqrt{\left(\log n\right)^2/nh\xi_\mu^{d_1}}\right\} \end{aligned} \tag{35}$$

Note that $\psi_n\left(w\right)$ is the sum of geometrically α-mixing and zero-mean variables with the kernel weight of K and M. Therefore, given that $\left[\log(\log n)\right]^4 \left(\log n\right)^2/nh\xi_\mu^{d_1} \to 0$, we can show that $\sup_{w \in S^c \times S^d} |\psi_n\left(w\right)| = O_{a.s.}(\sqrt{\left(\log n\right)/nh\xi_\mu^{d_1}})$ (by arguments as in the proof of Theorem 3 in Hansen, 2008; see also our discussions in deriving (60) in the proof of Lemma 2). Therefore, for n large enough and for any $w \in S^c \times S^d$, it almost surely holds that

$$\mathbf{1}\{\psi_n\left(w\right) > \sqrt{\left(\log n\right)^2/nh\xi_\mu^{d_1}}\} = 0. \tag{36}$$

This means that for n large enough, the second term on the RHS of (35) is zero almost surely, implying that the convergence rate of $\mathbf{A}_{n,1}$ is determined by the first term on the RHS of (35). Now, let

$$\delta_{n,j} := A_j^P \left[Y_j^P - \mu^P(W_j^P)\right] \bar{\psi}_n(W_j^P),$$

and note that by the definition of $\bar{\psi}_n\left(w\right)$, as well as by the boundedness of $\mu^P\left(\cdot\right)$ and Y_j^P (Assumptions 12 and 16-(ii)), there exists some constant C such that $\delta_{n,j} \leq C\sqrt{\left(\log n\right)^2/nh\xi_\mu^{d_1}}$. Then, look

at

$$
\begin{aligned}
&E[\left|\textstyle\sum_{j=1}^{n}\delta_{n,j}\right|^2]\ \ \left(=E[\left|\textstyle\sum_{j=1}^{n}A_j^P[Y_j^P-\mu^P\left(W_j^P\right)]\bar{\psi}_n\left(W_j^P\right)\right|^2]\right)\\
&=\textstyle\sum_{j=1}^{n}E[\delta_{n,j}^2]+2\sum\sum_{1\le j<l\le n}E[\delta_{n,j}\delta_{n,l}]\\
&\le nE[\delta_{n,1}^2]+2\sum\sum_{1\le j<l\le n}4a_{l-j}\times C^2\left(\log n\right)^2/nh\xi_\mu^{d_1}\\
&\le n[C^2\left(\log n\right)^2/nh\xi_\mu^{d_1}]+8n\textstyle\sum_{l=1}^{n}\tilde{A}\exp\{-\tilde{b}l\}\times C^2\left(\log n\right)^2/nh\xi_\mu^{d_1}=O((\log n)^2/h\xi_\mu^{d_1}),
\end{aligned}
\tag{37}
$$

where the first inequality uses the independence between $(A_j^P, A_j^P Y_j^P, W_j^P)$ and (D_i, W_i) (Assumption 16-(iii)) and the Billingsley inequality (see, e.g., Corollary 1.1 in Bosq, 1998), and the last inequality holds by noting the geometric mixing condition (Assumption 16) and the fact that $\sum_{l=1}^{n}\exp\{-\tilde{b}l\}=O\left(1\right)$ for any $\tilde{b}>0$. From (35)-(37), we see that $\mathbf{A}_{n,1}=O_P(\sqrt{\left(\log n\right)^2/n\xi_\mu^{d_1}})$.

Next, to consider the rate of $\mathbf{A}_{n,2}$, we look at

$$
\bar{a}_n\left(w\right)=E\left[a_{n,i}\left(w\right)\right]=\left[r\left(z,w\right)/q\left(w\right)\right]\left[1+o\left(1\right)\right]\ \ \text{uniformly over } w\in S^c\times S^d, \tag{38}
$$

where $r\left(z,w\right)$ is the probability function of $p\left(W_i^c,W_i^d\right)$ and W_i, i.e.,

$$
r(z,w^c,w^d)dzdw^c=\Pr[p(W_i^c,w^d)\in dz, W_i^c\in w^c, W_i^d=w^d].
$$

This result can be easily shown by using the differentiability and boundedness of $r\left(z,w\right)$ and $q\left(w\right)$, as well as the stated conditions on the kernel functions.[29] Noting that $\bar{a}_n\left(w\right)$ is uniformly bounded, we can show that $E[\left|\sum_{j=1}^{n}A_j^P\left[Y_j^P-\mu^P\left(W_j^P\right)\right]\bar{a}_n\left(W_j^P\right)\right|]^2=O\left(n\right)$ by arguments analogous to those for (37). Therefore, we have $\mathbf{A}_{n,2}=O(\sqrt{h})$.

Finally, we can easily show that the last term on the RHS of (34) is $O_P(\sqrt{nh}[\xi_\mu^{\kappa_\mu}+\xi_\mu\sqrt{\left(\log n\right)/n\xi_\mu^{d_1}}])$. From these arguments, we now have shown (30) as desired.

The convergence rate of $\mathbf{B}_n$ in (31). By the same arguments as for (33), we only need to consider the case where $W_i^c\in S_\circ^c$. Applying the Taylor expansion to $K_h\left(\hat{p}\left(W_i\right)-z_0\right)-K_h\left(p\left(W_i\right)-z_0\right)$, we can write $\mathbf{B}_n=\mathbf{B}_{n,1}+\mathbf{B}_{n,2}$, where

$$
\begin{aligned}
\mathbf{B}_{n,1}&=(\sqrt{nh}/nh^2)\textstyle\sum_{i=1}^{n}K'((p\left(W_i\right)-z_0)/h)[\hat{p}_{-i}\left(W_i\right)-p\left(W_i\right)]\mu^P\left(W_i\right);\\
\mathbf{B}_{n,2}&=(\sqrt{nh}/2nh^3)\textstyle\sum_{i=1}^{n}K''((\check{p}_{-i}\left(W_i\right)-z_0)/h)[\hat{p}_{-i}\left(W_i\right)-p\left(W_i\right)]^2\mu^P\left(W_i\right);
\end{aligned}
$$

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 can show that

$$
\mathbf{B}_{n,2}=O_P(\sqrt{n/h^3}[\xi_p^{\kappa_p}+\sqrt{\left(\log n\right)/n\xi_p^{d_1}}]^2) \tag{39}
$$

by arguments as those for (27) (with the boundedness of $\mu^P\left(w\right)$ and (24) of Lemma 2).

[29] Note that the existence of $r\left(\cdot,\cdot,w^d\right)$ and its differentiability follow from the existence of the density $f\left(\cdot,w^d\right)$ of W_i^c and the differentiability of $f\left(\cdot,w^d\right)$ and $p\left(\cdot,w^d\right)$ in Assumption 11)

To find the rate of $\mathbf{B}_{n,1}$, we use the expression (23) in Lemma 2. By letting

$$\begin{aligned}
b_{n,i}(w) &:= \left(1/h^2\right) K'((p(W_i)-z_0)/h)\mu^P(W_i)\, L_{\xi_p}(w^c - W_i^c)\mathbf{1}\{w^d = W_i^d\}/f(W_i);\\
\bar{b}_n(w) &:= E[b_{n,i}(w)];\\
\eta_{n,i}(w) &:= \left(1/h^2\right) K'((p(W_i)-z_0)/h)\mu^P(W_i)\, L_{\xi_p}(w^c - W_i^c)\mathbf{1}\{w^d = W_i^d\}\left[p(w)-p(W_i)\right]/f(W_i);\\
\bar{\eta}_n(w) &:= E[\eta_{n,i}(w)],
\end{aligned}$$

we can write

$$\begin{aligned}
\mathbf{B}_{n,1} &= (\sqrt{nh}/n^2)(n-1)\textstyle\sum_{j=1}^n \bar{b}_n(W_j)\left[D_j - p(W_j)\right]\\
&+ (\sqrt{nh}/n^2)\textstyle\sum_{i=1}^n \sum_{1\le j\le n;\ j\neq i}[b_{n,i}(W_j) - \bar{b}_n(W_j)]\left[D_j - p(W_j)\right]\\
&+ (\sqrt{nh}/n^2)(n-1)\textstyle\sum_{j=1}^n \bar{\eta}_n(W_j) + (\sqrt{nh}/n^2)\sum_{i=1}^n \sum_{1\le j\le n;\ j\neq i}[\eta_{n,i}(W_j) - \bar{\eta}_n(W_j)]\\
&+ (\sqrt{nh}/nh^2)\textstyle\sum_{i=1}^n \left|K'\left(\left(p(W_i)-z_0\right)/h\right)\mu^P(W_i)\right| \times O_{a.s.}([\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}]^2), \qquad (40)
\end{aligned}$$

where we below investigate the convergence rates of the five terms on the RHS.

To consider the rate of the first term, note that

$$\bar{b}_n(w) = -\left[\mu(w)/f(w)\right] r_1(z_0, w)\left[1 + o(1)\right] \text{ uniformly over } w \in S_\circ^c \times S^d,$$

where $r_1\left(z, w^c, w^d\right) := (\partial/\partial z)\, r\left(z, w^c, w^d\right)$ and $r\left(z, w^c, w^d\right)$ is the probability function of $p\left(W_i^c, W_i^d\right)$ and W_i (used in (38)). This follows from standard integration-by-parts and change-of-variable techniques as in the proof 6 of Hansen (2008). Therefore, by using the boundedness of $[D_j - p(W_j)]\,\bar{b}_n(W_j)$ and the Billingsley inequality (as in deriving (37)), we can show that $E[\left|\sum_{j=1}^n [D_j - p(W_j)]\,\bar{b}_n(W_j)\right|^2] = O(n)$ and thus, the first term on the RHS of (40) is $O_P(\sqrt{h})$.

To investigate the rate of the second term on the RHS of (40), we let

$$\pi_n(i,j) := [b_{n,i}(W_j) - \bar{b}_n(W_j)]\left[D_j - p(W_j)\right]. \qquad (41)$$

Then, we consider the following moment:

$$\begin{aligned}
&E[\left|\textstyle\sum_{i=1}^n \sum_{1\le j\le n;\ j\neq i}[b_{n,i}(W_j) - \bar{b}_n(W_j)]\left[D_j - p(W_j)\right]\right|^2]\\
&= \sum_{1\le i,j\le n;\ i\neq j}\sum \left\{E[|\pi_n(i,j)|^2] + E[\pi_n(i,j)\,\pi_n(j,i)]\right\} + \sum_{1\le i,j,k\le n;\ i\neq j\neq k;}\sum\sum E\left[\pi_n(i,j)\,\pi_n(i,k)\right] \qquad (42)
\end{aligned}$$

where the equality holds since $\{W_i\}$ is I.I.D., $E[b_{n,i}(W_j) - \bar{b}_n(W_j) \,|\, \{W_l :\ 1 \le l \le n;\ l \neq i\}] = 0$ for any $i \neq j$ and $E[D_j - p(W_j)\,|W_j] = 0$ for any j (i.e., it holds that $E[\pi_n(i,j)\,\pi_n(k,i)] = E[\pi_n(i,j)\,\pi_n(j,k)] = E[\pi_n(i,j)\,\pi_n(k,j)] = 0$ for $i \neq j \neq k$ and $E[\pi_n(i,j)\,\pi_n(k,l)] = 0$ for $i \neq j \neq k \neq l$). Then, we can show that

$$\begin{aligned}
E[|\pi_n(i,j)|^2] &= O(1/h^3\xi_p^{d_1}),\\
E[\pi_n(i,j)\,\pi_n(j,i)] &\le \sqrt{E[|\pi_n(i,j)|^2]E[|\pi_n(j,i)|^2]} = O(1/h^3\xi_p^{d_1}),
\end{aligned}$$

uniformly over i and j ($i \neq j$), by standard change-of-variable arguments, and also that

$$\sum_{1\leq i,j,k\leq n;}\sum_{i\neq j\neq k;}\sum E\left[\pi_n\left(i,j\right)\pi_n\left(i,k\right)\right] = O(n^2/h^{7/2}\xi_p^{d_1}), \tag{43}$$

where we below provide the proof of (43). Now, these results imply that the RHS of (42) is $O(n^2/h^{7/2}\xi_p^{d_1})$ and therefore, the second term on the RHS of (40) is $O_P(1/\sqrt{nh^{5/2}\xi_p^{d_1}})$.

The third term on the RHS of (40) can be shown to be $O_P(\sqrt{nh}\xi_p^{\kappa_p})$, since

$$\bar{\eta}_n\left(w\right) = O(\xi_p^{\kappa_p}) \text{ uniformly over } w \in S_\circ^c \times S^d,$$

which follows from by the standard change-of-variable and Talylor-expansion arguments.

As for the fourth term on the RHS of (40), we look at

$$\begin{aligned}
& E[\left|\sum_{i=1}^n \sum_{1\leq j\leq n;\ j\neq i}[\eta_{n,i}(W_j) - \bar{\eta}_n(W_j)]\right|^2] \\
= \ & n\left(n-1\right)\left\{E[\left|\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)\right|^2] + E[\left[\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)\right]\left[\eta_{n,2}\left(W_1\right) - \bar{\eta}_n\left(W_1\right)\right]]\right\} \\
& + n\left(n-1\right)\left(n-2\right)E[[\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)][\eta_{n,1}\left(W_3\right) - \bar{\eta}_n\left(W_3\right)]] \\
= \ & O(n^2/h^3\xi_p^{d_1-2}) + O(n^3\xi_p^{2\kappa_p}/h^3),
\end{aligned} \tag{44}$$

where the first equality follows from the I.I.D. condition of $\{W_i\}$ and the fact that $E[\eta_{n,i}\left(W_j\right) - \bar{\eta}_n\left(W_j\right) \mid \{W_l:\ 1\leq l\leq n;\ l\neq i\}] = 0$; and the last equality uses the following results

$$\begin{aligned}
E[\left|\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)\right|^2] &= O(1/h^3\xi_p^{d_1-2}); \\
E[\left[\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)\right]\left[\eta_{n,2}\left(W_1\right) - \bar{\eta}_n\left(W_1\right)\right]] &= O(1/\xi_p^{d_1-2}); \text{ and} \\
E[\left[\eta_{n,1}\left(W_2\right) - \bar{\eta}_n\left(W_2\right)\right]\left[\eta_{n,1}\left(W_3\right) - \bar{\eta}_n\left(W_3\right)\right]] &= O(\xi_p^{2\kappa_p}/h^3),
\end{aligned}$$

which can be shown by the Taylor-expansion, change-of-variable and integration-by-parts techniques. Given (44), we can see that the fourth term on the RHS of (40) is $O_P(\xi_p\sqrt{1/nh^2\xi_p^{d_1}} + \xi_p^{\kappa_p}/h)$.

The last term on the RHS of (40) can be easily shown to be $O_P(\sqrt{nh}[\xi_p^{\kappa_p} + \sqrt{\left(\log n\right)/n\xi_p^{d_1}}]^2)$. Therefore, by putting these arguments together, we have

$$\begin{aligned}
\mathbf{B}_{n,1} &= O_P(\sqrt{h}) + O_P(1/\sqrt{nh^{5/2}\xi_p^{d_1}}) + O_P(\sqrt{nh}\xi_p^{\kappa_p}) \\
&\quad + O_P(\xi_p\sqrt{1/nh^2\xi_p^{d_1}} + \xi_p^{\kappa_p}/h) + O_P(\sqrt{n/h}[\xi_p^{\kappa_p} + \sqrt{\left(\log n\right)/n\xi_p^{d_1}}]^2)
\end{aligned} \tag{45}$$

Now, by (39) and (45), we obtain the desired result (31). It remains to show (43).

Proof of (43). We consider two moment bounds of $E\left[\pi_n\left(i,j\right)\pi_n\left(i,k\right)\right]$. First, by recalling the definition of π_n in (41) and by using standard change-of-variable arguments, we can easily show that

$$\left|E\left[\pi_n\left(i,j\right)\pi_n\left(i,k\right)\right]\right| = O\left(1/h^3\right) \tag{46}$$

uniformly over i, j and k with $i \neq j \neq k$. Second, by the Davydov inequality (see, e.g., Corollary 1.1 of Bosq, 1998), we have

$$\begin{aligned}
& |E\left[\pi_n(i,j)\,\pi_n(i,k)\right]| \\
= \;& \left|E\left[\; E\left[\left\{b_{n,i}(W_j)-\bar{b}_n(W_j)\right\}\left\{b_{n,i}(W_k)-\bar{b}_n(W_k)\right\} \mid W_j, W_k\right]\left\{D_j-p(W_j)\right\}\times\left\{D_k-p(W_k)\right\}\;\right]\right| \\
\leq \;& 8(2a_{|j-k|})^{1/4}\left\{E\left[\;\left|E\left[\left\{b_{n,i}(W_j)-\bar{b}_n(W_j)\right\}\left\{b_{n,i}(W_k)-\bar{b}_n(W_k)\right\}|W_j,W_k\right]\left\{D_j-p(W_j)\right\}\right|^2\;\right]\right\}^{1/2} \\
& \times\left\{E[|D_k-p(W_k)|^4]\right\}^{1/4} \\
= \;& \exp\{-\tilde{b}\,|j-k|\}\times O(\sqrt{1/h^7\xi_p^{2d_1}}), \qquad (47)
\end{aligned}$$

uniformly over $i \neq j \neq k$ with $i \neq j \neq k$ and $|j-k| \geq T_n+1$, where the last equality uses the geometric mixing condition of $\{D_i, W_i\}$ (Assumption 16), the Jensen inequality, the boundedness of $|D_j - p(W_j)|(\leq 2$ for any $j)$, and the following result:

$$E[\left\{b_{n,i}(W_j)-\bar{b}_n(W_j)\right\}^2\left\{b_{n,i}(W_k)-\bar{b}_n(W_k)\right\}^2] = O(1/h^7\xi_p^{2d_1}).$$

Now, let $\{T_n\}$ be a sequence of integers tending to ∞ (as $n \to \infty$). Then,

$$\begin{aligned}
& \sum_{1\leq i,j,k\leq n;\; i\neq j\neq k}\sum\sum E\left[\pi_n(i,j)\,\pi_n(i,k)\right] \\
\leq \;& \left\{\sum_{1\leq i,j,k\leq n;\; i\neq j\neq k;\; |j-k|<T_n+1}\sum\sum + \sum_{1\leq i,j,k\leq n;\; i\neq j\neq k;\; |j-k|\geq T_n+1}\sum\sum\right\}\left|E\left[\pi_n(i,j)\,\pi_n(i,k)\right]\right| \\
= \;& n^2 T_n \times O\left(1/h^3\right) + \sum_{1\leq i,j,k\leq n;\; i\neq j\neq k;\; |j-k|\geq T_n+1}\sum\sum \exp\{-\tilde{b}\,|j-k|\}\times O(\sqrt{1/h^7\xi_p^{2d_1}}) \\
= \;& O\left(n^2T_n/h^3\right) + n^2\textstyle\sum_{l=T_n+1}^{\infty}\exp\{-\tilde{b}l\}\times\sqrt{1/h^7\xi_p^{2d_1}} = O(n^2/h^{7/2}\xi_p^{d_1}),
\end{aligned}$$

where the first equality follows from (46) and (47); and the last equality holds by letting $T_n := 1/h^{1/2}\xi_p^{d_1}$ (this T_n is some polynomial order of n under the stated bandwidth conditions and thus $\sum_{l=T_n+1}^{\infty}\exp\{-\tilde{b}l\} < \infty$). We now have proved (43) as desired.

The convergence rate of $\mathbf{C}_n$ in (32). Let $\varepsilon\,(>0)$ any (small) constant. Then, by (24) 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 $\max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(W_i)| \leq \varepsilon$, implying that 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 (33), means that if $W_i \notin S_\circ^c$ and n is large enough (with h small enough), $[K_h(\hat{p}_{-i}(W_i) - z_0) - K_h(p(W_i) - z_0)]\left[\hat{\mu}^P(W_i) - \mu^P(W_i)\right] = 0$. Thus, for deriving the upper bound of $\mathbf{C}_n$, it is sufficient to consider only the case where $W_i \in S_\circ^c$ and therefore, for h small enough, it almost surely holds that

$$\begin{aligned}
|\mathbf{C}_n| \leq\; & (\sqrt{n/h})\times(1/nh)\textstyle\sum_{i=1}^{n}\mathcal{K}^*((p(W_i)-z_0)/h) \\
& \times\max_{i\in\{1,\ldots,n\}}\sup_{w\in S^c\times S^d}|\hat{p}_{-i}(w)-p(w)|\times\sup_{w\in S_\circ^c\times S^d}|\hat{\mu}^P(w)-\mu^P(w)|,
\end{aligned}$$

where $\mathcal{K}^*$ is the function used for deriving the bound of $\mathbf{J}_n$ in (27). By (24) and (26), we now obtain the desired result (32). ■

It remains to prove two auxiliary lemmas:

Proof of Lemma 2. Let

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

Then, for each i, we can write

$$\hat{p}_{-i}(w) - p(w) = [\hat{H}_{-i}(w) + \hat{\Gamma}_{-i}(w)] \times \Big[\frac{1}{f(w)} + \frac{f(w) - \hat{f}_{-i}(w)}{\hat{f}_{-i}(w) f(w)}\Big]. \tag{48}$$

For the components on the RHS of (48), we can show the following convergence results:

$$\max\nolimits_{i \in \{1,\ldots,n\}} \sup\nolimits_{w \in S_\circ^c \times S^d} |\hat{f}_{-i}(w) - f(w)| = O_{a.s.}(\xi_p^{\kappa_p} + \sqrt{(\log n)/n\xi_p^{d_1}}); \tag{49}$$
$$\max\nolimits_{i \in \{1,\ldots,n\}} \sup\nolimits_{w \in S^c \times S^d} |\hat{f}_{-i}(w) - f(w)| = O_{a.s.}(\xi_p + \sqrt{(\log n)/n\xi_p^{d_1}}); \tag{50}$$
$$\max\nolimits_{i \in \{1,\ldots,n\}} \sup\nolimits_{w \in S^c \times S^d} |\hat{\Gamma}_{-i}(w)| = O_{a.s.}(\xi_p^{\kappa_p} + \xi_p\sqrt{(\log n)/n\xi_p^{d_1}}); \tag{51}$$
$$\max\nolimits_{i \in \{1,\ldots,n\}} \sup\nolimits_{w \in S^c \times S^d} |\hat{H}_{-i}(w)| = O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}}), \tag{52}$$

whose proofs are provided below. Now, fix any $\omega \in \Omega^*$, where Ω^* is an event with $\Pr(\Omega^*) = 1$. Then, (50) implies that as $n \to \infty$, $\max_{i \in \{1,\ldots,n\}} \sup_{w \in S^c \times S^d} |\hat{f}_{-i}(w) - f(w)| < C_1/2$ (C_1 is given in Assumption 15), and therefore,

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

implying that

$$1/\hat{f}_{-i}(w) = O_{a.s.}(1), \text{ uniformly over } i \in \{1, \ldots, n\} \text{ and } w \in S^c \times S^d. \tag{53}$$

Now, by (48) and (51)-(53), we have the following expression:

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

Then, (49) and (50) imply the first and second results (23) and (24), respectively. It remains to show show (49), (50), (51) and (52).

Proofs of (49) and (50). Letting $\hat{f}(w) := n^{-1} \sum_{1 \leq k \leq n} L_{\xi_p}(W_k^c - w^c) \mathbf{1}\{W_k^d = w^d\}$, we have the following decomposition:

$$\begin{aligned}
\hat{f}_{-i}(w) - f(w) &= \hat{f}(w) - f(w) - (n\xi_p^{d_1})^{-1} L\left((W_i^c - w^c)/\xi_p\right) \mathbf{1}\{W_i^d = w^d\} \\
&= \{\hat{f}(w) - E[\hat{f}(w)]\} + \{E[\hat{f}(w)] - f(w)\} + O(1/n\xi_p^{d_1}),
\end{aligned} \tag{54}$$

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 analogous arguments as in the proof of Theorem 3 of Hansen (2008), we can show that the first term on the RHS of (54) is $O_{a.s.}(\sqrt{(\log n)/n\xi_p^{d_1}})$ uniformly over $w \in S^c \times S^d$.[30]

As for the second term, noting that $S_\circ^c$ is strictly in the interior of S^c, we can also show that

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

which follows from standard arguments for biases of kernel-based estimators, say change-of-variable and Taylor-approximation arguments with Assumption 13 (see, e.g., proofs of Theorems 6 and 8 in Hansen, 2008). This implies the desired result (49). Next, if we let the domain of w^c as the whole set S^c (instead of $S_\circ^c$), we have

$$\begin{aligned}
&\sup_{(w^c,w^d)\in S^c\times S^d} |E[\hat{f}(w)] - f(w)| \\
&= \sup_{(w^c,w^d)\in S^c\times S^d} \left| \textstyle\sum_{u^d\in S^d} (1/\xi_p^{d_1}) \int_{u^c\in S^c} L((u^c - w^c)/\xi_p) \mathbf{1}\{u^d = w^d\} f(u^c, u^d) du^c - f(w) \right| \\
&= \sup_{(w^c,w^d)\in S^c\times S^d} \left| \int_{v^c\in T^c(w^c,\xi_p)} L(v^c) \left[f(w^c + \xi_p v^c, w^d) - f(w^c, w^d) \right] dv^c \right| \\
&= \sup_{(w^c,w^d)\in S^c\times S^d} \left| \int_{v^c\in T^c(w^c,\xi_p)} L(v^c) \langle \xi_p v^c, (\partial/\partial w^c) f(\tilde{w}^c, w^d) \rangle dv^c \right| \\
&\leq \xi_p \int_{v^c\in S^L} |L(v^c)| \times ||v^c|| dv^c \times \sup_{(w^c,w^d)\in S^c\times S^d} ||(\partial/\partial w^c) f(w^c, w^d)|| = O(\xi_p),
\end{aligned} \tag{55}$$

where $T^c(w^c, \xi_p) := \{v |\ w^c + \xi_p v \in S^c\}$; $\langle a, b\rangle$ 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 $(u^c - w^c)/\xi_p = v^c$; and the third equality uses the mean-value theorem; and the inequality uses the fact that $T^c(w^c, \xi_p) \supset S^L$ (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 (55) implies the desired result (50).

Proof of (51). We write

$$\hat{\Gamma}(w) := n^{-1} \textstyle\sum_{1\leq j\leq n} L_{\xi_p}(W_j^c - w^c) \mathbf{1}\{W_j^d = w^d\} [D_j - p(w)]. \tag{56}$$

By the same arguments as for (54), it holds that

$$\hat{\Gamma}_{-i}(w) = E[\hat{\Gamma}(w)] + \{\hat{\Gamma}(w) - E[\hat{\Gamma}(w)]\} + O(1/n\xi_p^{d_1}), \tag{57}$$

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

First, 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

[30] We note that we only suppose the first-order stationarity of the sequence, while Hansen (2008) considers the strict stationarity case. However, the key to Hansen's results is the mixing condition and the strict stationarity condition is not an essential one. In fact, Kristensen (2009) work without any stationarity condition and derives results analogous to those in Hansen (2008).

boundary points of S_+^c are in the interior of S^c. By Assumption (15), such S_+^c exists. Let $N_+^c := \left\{u \in S^c |\ u \notin S_+^c\right\}$. Then, we look at the following bound:

$$\sup_{\left(w^c,w^d\right)\in S^c\times S^d} |E[\hat{\Gamma}\left(w\right)]| \leq \sup_{\left(w^c,w^d\right)\in S_+^c\times S^d} |E[\hat{\Gamma}\left(w\right)]| + \sup_{\left(w^c,w^d\right)\in N_+^c\times S^d} |E[\hat{\Gamma}\left(w\right)]|. \quad (58)$$

The first term on the RHS of (58) is $O(\xi_p^{\kappa_p})$. This can be shown by standard arguments for 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 (58), we see

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

where the second equality holds since $p(w^c,w^d) = 0$ for $(w^c,w^d) \in N_+^c \times S^d$, and the last equality holds for ξ_p small enough, since $p(w^c + \xi_p v^c, w^d) = 0$ 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

$$\sup_{\left(w^c,w^d\right)\in S^c\times S^d} |E[\hat{\Gamma}\left(w\right)]| = O(\xi_p^{\kappa_p}). \quad (59)$$

Second, we derive the uniform bound of $\{\hat{\Gamma}\left(w\right) - E[\hat{\Gamma}\left(w\right)]\}$. For this purpose, let $\{Z_{n,i}\}$ be an array:

$$\begin{aligned}
Z_{n,i} := &L\left((W_j^c - w^c)/\xi_p\right) \mathbf{1}\{W_j^d = w^d\} \left[p\left(W_j\right) - p\left(w\right)\right] \\
&- E[L\left((W_j^c - w^c)/\xi_p\right) \mathbf{1}\{W_j^d = w^d\} \left[p\left(W_j\right) - p\left(w\right)\right]].
\end{aligned}$$

Note that $(n\xi_p^{d_1})^{-1} \sum_{i=1}^n Z_{n,i} = \sum_{i=1}^n (\hat{\Gamma}\left(w\right) - E[\hat{\Gamma}\left(w\right)])$. By arguments similar to those for $E[\hat{\Gamma}\left(w\right)]$, we can show that for any $m \leq n$, $E[(\sum_{i=1}^m Z_{n,i})^2] = \sum_{i=1}^m E[Z_{n,i}^2] = O(m\xi_p^{d_1+2})$ uniformly over $\left(w^c, w^d\right) \in S^c \times S^d$. Given this variance bound and using techniques based on Bernstein-type inequality (see, e.g., the proof of Theorem 3 in Hansen, 2008), we can prove that

$$\sup_{\left(w^c,w^d\right)\in S^c\times S^d} |\hat{\Gamma}\left(w\right) - E[\hat{\Gamma}\left(w\right)]| = O_{a.s.}(\xi_p \sqrt{\left(\log n\right)/n\xi_p^{d_1}}), \quad (60)$$

which, together with (57) and (59) imply the desired result (51).

Proof of (52). This result follows from arguments analogous to above, and we omit details (we use arguments as in the proof of Theorem 3, Hansen, 2008). Now, the proof is completed. ■

Proof of Lemma 3. The proof proceeds quite analogously to that of Lemma (2), and we omit details for brevity. By considering the decomposition of $\hat{\mu}^P\left(w\right) - \mu^P\left(w\right)$ as in (48) and derive results corresponding to (49)-(52), we can obtain the desired expressions (we note that the uniform rates are established only over $w^c \in S_\circ^c$ (the interior set) and in terms of convergence in probability, where we use arguments analogous to those for Theorems 2, 6 and 8, Hansen, 2008). ■