Variation in impacts of letters of recommendation on college admissions decisions: Approximate balancing weights for treatment effect heterogeneity in observational studies*

Eli Ben-Michael, Avi Feller, and Jesse Rothstein

UC Berkeley

July 1, 2020

Abstract

Assessing treatment effect variation in observational studies is challenging because differences in estimated impacts across subgroups reflect both differences in impacts and differences in covariate balance. Our motivating application is a UC Berkeley pilot program for letters of recommendation in undergraduate admissions: we are interested in estimating the differential impacts for under-represented applicants and applicants with differing a priori probability of admission. We develop balancing weights that directly optimize for “local balance” within subgroups while maintaining global covariate balance between treated and control populations. We then show that this approach has a dual representation as a form of inverse propensity score weighting with a hierarchical propensity score model. In the UC Berkeley pilot study, our proposed approach yields excellent local and global balance, unlike more traditional weighting methods, which fail to balance covariates within subgroups. We find that the impact of letters of recommendation increases with the predicted probability of admission, with mixed evidence of differences for under-represented minority applicants.

*email: ebenmichael@berkeley.edu. This work was partially supported by the William T. Grant Foundation. We thank Greg Dubrow, Amy Jarich, Olufeme Ogundole, and James Pustejovsky for helpful conversations and thoughtful comments.
1 Introduction and motivation

Assessing treatment effect variation in observational studies is challenging, even for pre-specified subgroups: variation in estimated impacts reflect both actual treatment effect variation and differences in covariate balance across groups. Our motivating application is a UC Berkeley pilot program for letters of recommendation in undergraduate admissions during the 2016-2017 admissions cycle. Using much of the same applicant information as the UC Berkeley admissions department — as well as information on the process of soliciting letters of recommendation — we design an observational study on the impact of submitting a letter of recommendation on subsequent admission. Our goal is to understand how impacts in this study vary for under-represented applicants and for applicants with differing \textit{a priori} probabilities of admission.

Traditional Inverse Propensity Score Weighting (IPW) is a standard approach to estimating varying effects in this setting: first estimate a propensity score model, typically via logistic regression, and then construct weights based on the estimated model. Researchers can then assess subgroup variation by including interaction terms in the propensity score model (see Green and Stuart, 2014). As with all traditional IPW methods, however, the estimated weights are only guaranteed to have good covariate balancing properties asymptotically, and can behave poorly in more realistic settings, including when the propensity score model is mis-specified. And even if traditional IPW achieves good global balance, the fully interacted propensity score model is typically high dimensional and is unlikely to achieve good balance within subgroups (see Lee et al., 2019; Dong et al., 2020; Yang et al., 2020).

In this paper, we develop a balancing weights approach for estimating subgroup treatment effects by directly optimizing for covariate balance. We present a convex optimization problem that finds weights that directly target the level of local imbalance within each subgroup — while still ensuring global covariate balance between treated and control populations. This approach therefore adapts approximate balancing weights for overall effects (Zubizarreta, 2015; Athey et al., 2018; Hirshberg and Wager, 2019; Ben-Michael et al., 2020) to estimate subgroup treatment effects. We show that controlling local imbalance in turn controls the estimation error of subgroup-specific effects, allowing us to better isolate treatment effect variation.

Next, we demonstrate that this proposal has a dual representation as inverse propensity weighting with a hierarchical propensity score model, building on recent connections between balancing weights and propensity score estimation (Zhao and Percival, 2016; Tan, 2017; Ben-Michael et al., 2019a; Chattopadhyay et al., 2020). In particular, finding weights that minimize both global and local imbalance corresponds to estimating a propensity score model in which the subgroup-specific parameters are partially pooled toward a global propensity score model. Any remaining imbalance after weighting may lead to bias. To adjust for this, we also combine the weighting approach with an outcome model, analogous to bias correction for matching (Rubin, 1973; Athey et al., 2018).

We apply our approach to estimate the impacts of letters of recommendation during the 2016...
UC Berkeley undergraduate admissions cycle. We focus on variation in the effect on admissions rates based on under-represented minority (URM) status and on the a priori predicted probability of admission, estimated using data from the prior year’s admissions cycle. First, we show that the proposed weights indeed yield excellent local and global balance, while traditional propensity score weighting methods yield poor local balance. We then find evidence that the impact of letters increases with the predicted probability of admission. Applicants who are very unlikely to be admitted see little benefit from letters of recommendation while applicants on the cusp of acceptance see a larger impact.

The evidence on the differential effects by URM status is more mixed. Overall, the point estimates for URM and non-URM applicants are close to each other. However, these estimates are noisy and mask important variation by a priori probability of admission. For applicants with the highest baseline admission probabilities, we estimate larger impacts for non-URM than URM applicants, though these estimates are sensitive to augmentation with an outcome model. For all other applicants, we estimate the reverse: larger impacts for URM than non-URM applicants. Since URM status is correlated with the predicted probability of admission, this leads to a Simpson’s Paradox-type pattern for subgroup effects, with a slightly larger point estimate for non-URM applicants pooled across groups (Bickel et al., 1975; VanderWeele and Knol, 2011).

These results hinge on estimating higher-order interaction terms with the treatment. This suggests caution but also highlights the advantages of a design-based approach (Rubin, 2008). Since we separate the design and analysis phases, we can carefully assess covariate balance and overlap in the subgroups of interest — and can tailor the weights to target these quantities directly. This is a challenge for many recent approaches that use automatic machine learning methods to regularize the complexity of estimated heterogeneous treatment effects (Carvalho et al., 2019). Nonetheless, we view our proposed approach as a complement to — not a substitute for — these approaches and explore an augmented estimator as part of our analysis.

The importance of higher-order interactions also suggests that, as in all observational studies, our results are sensitive to violating the strong assumption of ignorable treatment assignment. Thus, we argue our analysis is a reasonable first look at this question, best understood alongside other approaches that rest on different assumptions (such as those in Rothstein, 2017). In Appendix A.2, we explore one alternative approach that instead leverages unique features of the UC Berkeley pilot study, which included an additional review without the letters of recommendation from a sample of 10,000 applicants. The results from this approach are broadly similar to the estimates from the observational study, differing mainly in regions with relatively poor overlap.

The paper proceeds as follows. In the next section we introduce the letter of recommendation pilot program at UC Berkeley. Section 2 introduces the problem setup and notation, and discusses related work. Section 3 proposes and analyzes the approximate balancing weights approach. Section 4 presents a simulation study. Section 5 presents empirical results on the effect of letters of
recommendation. Section 6 concludes with a discussion about possible extensions. The appendix includes additional theoretical discussion and analysis.

1.1 Motivating example: A pilot program for letters of recommendation in college admissions

Through a pilot study during the 2016-17 admissions cycle at the University of California, Berkeley, many applicants for freshman admission were invited to submit letters of recommendation (LORs) as part of their applications. Rothstein (2017) provides additional detail about the pilot study. On its face, the use of LORs in admissions seems consistent with the spirit of what the University of California calls “holistic review”: looking beyond reductive summaries (e.g. SAT scores) and examining the whole applicant, taking account of any contextual factors and obstacles overcome (Hout, 2005). LORs have the potential to offer insight into aspects of the applicant not captured by the available quantitative information or by the essays that applicants submit (Kuncel et al., 2014; Hout, 2005). However, there is legitimate concern that applicants from under-resourced high schools may have trouble identifying potential writers, and that when they do identify writers, the letters that are provided may be less informative or prejudicial against the applicant, due, e.g., to poor writing or grammar, or to lower status of the letter writer; see Schmader et al. (2007) as an example.

Given this ambiguity, we analyze the results of the 2016 admission cycle to determine whether and how the inclusion of LORs affected admissions readers’ evaluations of URM UC Berkeley freshman applicants relative to non-URM applicants. Throughout, we restrict our analysis to non-athlete California residents who applied to either the College of Letters and Science or the College of Engineering. This leaves 40,541 applicants, 11,143 of whom submitted LORs. For the purposes of this study, we define a URM applicant as one who is a low-income student, a student in a low-performing high school, a first-generation college student, or from an underrepresented racial or ethnic group. We focus our analysis on the impacts for applicants who both were invited to and subsequently did submit LORs.1

1.1.1 Selection into treatment

UC Berkeley uses a two-reader evaluation system. Each reader scores applicants on a three-point scale, as “No,” “Possible,” or “Yes.” Application decisions are based on the combination of these two scores and the major to which a student has applied. In the most selective majors (e.g., mechanical engineering), an applicant must receive two “Yes” scores to be admitted, while in others a single “Yes” is sufficient. In the LOR pilot, applicants were invited to submit letters based

---

1 We could use the methods discussed here to explore a range of different quantities. For this target, the net effect of LORs on admission includes differential rates of submission of a letter given invitation. While non-URM applicants submitted letters at a higher rate than URM applicants, the majority of the discrepancy arises from applicants who were unlikely to be admitted a priori (Rothstein, 2017).
Figure 1: Absolute difference in means, standardized by the pooled standard deviation, between applicants submitting and not submitting letters of recommendation for several key covariates. By design, applicants submitting letters of recommendation disproportionately have a “Possible” score from the first reader (70% of treated applicants vs. 4% of untreated applicants).

As in any observational study of causal effects, selection into treatment is central. Decisions to submit letters were a two-step process. Any applicant who received a “Possible” score from the first reader was invited. In addition, due to concerns that first read scores would not be available in time to be useful, an index of student- and school-level characteristics was generated, and applicants with high levels of the index were invited as well.\(^2\) Of the 40,451 total applicants, 14,596 were invited to submit a letter. Approximately 76% of those invited to submit letters eventually submitted them, and no applicant submitted a letter who was not invited to.

For this analysis, we assume that submission of LORs is effectively random conditional on the first reader score and on both student- and school-level covariates. In particular, the interaction between the covariates and the first reader score plays an important role in the overall selection mechanism, as applicants who received a score of “No” or “Yes” from the first reader could still have been asked to submit an LOR based on their individual and school information. Figure 1 shows covariate imbalance for several key covariates — measured as the absolute difference in means divided by the pooled standard deviation — for applicants who submitted LORs versus those who did not.\(^3\) We see that there are large imbalances in observable applicant characteristics, most

\(^2\) The index was generated from a logistic regression fit to data from the prior year’s admissions cycle, predicting whether an applicant received a “Possible” score (versus either a “No” or a “Yes”). Applicants with predicted probabilities from this model greater than 50% were invited to submit LORs. Because we observe all of the explanatory variables used in the index, this selection depends only on observable covariates. A small share of applicants with low predicted probabilities received first reads after January 12, 2017, the last date that LOR invitations were sent, and were not invited even if they received “Possible” scores.

\(^3\) The full set of student-level variables we include in our analysis are: weighted and unweighted GPA, GPA
notably average school income, GPA, the number of honors and AP classes taken, and SAT score. There were also large imbalances in first reader scores (not shown in Figure 1): 70% of applicants that submitted LORs had “Possible” scores, compared to only 4% of those who did not.

1.1.2 Heterogeneity across \( a \ priori \) probability of admission

To better understand who was invited to submit LORs and any differential impacts between URM and non-URM applicants, we construct a univariate summary of applicant- and school-level characteristics. We use logistic regression to estimate the probability of admission given observable characteristics using the prior year (2015) admissions data.\(^4\) We then use this model to predict \( a \ priori \) admissions probabilities for the applicants of interest in 2016; we refer to this as the Admissibility Index (AI). The overall AUC in predicting 2016 admissions is 0.88 and the mean square error is 10% (see Appendix Table B.2). However, the predictive accuracy decreases for higher AI applicants, slightly under-estimating the probability of admissions for middle-tier applicants and over-estimating for the highest admissibility applicants (see Appendix Figure B.1). Additionally, we see that predictive performance is better for URM applicants than non-URM applicants,

---

\(^4\)This is a different model than the logistic regression used by the admissions office, which predicted a reviewer score of “Possible” rather than admission.
Table 1: Number of applicants and proportion treated by subgroup.

<table>
<thead>
<tr>
<th>AI Range</th>
<th>URM</th>
<th>Number of Applicants</th>
<th>Number Submitting LOR</th>
<th>Proportion Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 5%</td>
<td>URM</td>
<td>11,832</td>
<td>2,157</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>6,529</td>
<td>607</td>
<td>9%</td>
</tr>
<tr>
<td>5% - 10%</td>
<td>URM</td>
<td>3,106</td>
<td>1,099</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>2,099</td>
<td>536</td>
<td>25%</td>
</tr>
<tr>
<td>10% - 20%</td>
<td>URM</td>
<td>2,876</td>
<td>1,212</td>
<td>42%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>2,495</td>
<td>828</td>
<td>33%</td>
</tr>
<tr>
<td>&gt; 20%</td>
<td>URM</td>
<td>4,645</td>
<td>2,345</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>6,959</td>
<td>2,359</td>
<td>34%</td>
</tr>
</tbody>
</table>

particularly for applicants to the College of Engineering (see Appendix Figure B.2).

Figure 2 shows the AI distribution for the 2016 applicant cohort, broken out by URM status and LOR submission. There are several features of this distribution that have important implications for our analysis. First, although the probability of admission is quite low overall, applicants across nearly the full support of probabilities submitted LORs. This is primarily because applicants who received “Possible” scores from the first readers come from a wide range of admissibility levels. This will also allow us to estimate heterogeneous effects across the full distribution, with more precision for applicants with lower AIs. Second, because the admissions model disproportionately predicted that URM students had high chances of receiving “Possible” scores, many more URM applicants were invited to submit letters than non-URM applicants, and so our estimates for URM applicants will be more precise than those for non-URM applicants.

From Figure 2 we know that the distribution of AI varies between URM and non-URM applicants, and so apparent differences in estimated effects between the two groups may be due to compositional differences. Therefore, in the subsequent sections we will focus on estimating effects within subgroups defined by both URM status and admissibility. To do this, we define subgroups by creating four (non-equally-sized) strata of the AI: < 5%, 5% – 10%, 10% – 20% and > 20%. Interacting with URM status, this leads to eight non-overlapping subgroups; we will marginalize over these to estimate the other subgroup effects above. Table 1 shows the total number of applicants in each of the eight groups, along with the proportion submitting letters of recommendation.

2 Treatment effect variation in observational studies

2.1 Setup and estimands

We now describe the letter of recommendation study as an observational study where for each applicant $i = 1, \ldots, n$, we observe applicant and school level-covariates $X_i \in \mathcal{X}$; a group indicator
$G_i \in \{1, \ldots, K\}$ denoting e.g., URM status or coarsened AI; a binary indicator for submitting a letter of recommendation $W_i \in \{0, 1\}$; and whether the applicant is admitted, which we denote as $Y_i \in \{0, 1\}$. We assume that for each applicant, $(X_i, G_i, W_i, Y_i)$ are sampled i.i.d. from some distribution $P(\cdot)$. Additionally, let $n_{1g}$ and $n_{0g}$ be the number of treated and control units in subgroup $G_i = g$, respectively. Following the potential outcomes framework (Neyman, 1923; Holland, 1986), we assume SUTVA (Rubin, 1980) and posit two potential outcomes $Y_i(0)$ and $Y_i(1)$ for each applicant $i$, corresponding to $i$’s outcome if that applicant submits a letter of recommendation or not, respectively; the observed outcome is $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$.$^5$ In this study we are interested in estimating two types of effects. First, we wish to estimate the overall Average Treatment Effect on the Treated (ATT), the treatment effect for applicants who submit a letter,

$$\tau = E[Y(1) - Y(0) \mid W = 1],$$

where we denote $\mu_1 = E[Y(1) \mid W = 1]$ and $\mu_0 = E[Y(0) \mid W = 1]$. Second, for each subgroup $G_i = g$, we would like to estimate the Conditional ATT (CATT),

$$\tau_g = E[Y(1) - Y(0) \mid G = g, W = 1],$$

where similarly we denote $\mu_{1g} = E[Y(1) \mid G = g, W = 1]$ and $\mu_{0g} = E[Y(0) \mid G = g, W = 1]$.

Estimating $\mu_{1g}$ is relatively straightforward: we can simply use the average outcome for treated units in group $g$, $\hat{\mu}_{1g} \equiv \frac{1}{n_{1g}} \sum_{G_i = g} W_i Y_i$. However, estimating $\mu_{0g}$ is more difficult due to confounding; we focus much of our discussion on imputing this counterfactual mean for the group of applicants who submitted letters of recommendation. To do this, we rely on two key assumptions that together form the usual strong ignorability assumption (Rosenbaum and Rubin, 1983).

Assumption 1 (Ignorability). The potential outcomes are independent of treatment given the covariates and subgroup:

$$Y(1), Y(0) \perp \perp W \mid X, G.$$  \hspace{1cm} (2)

Assumption 2 (One Sided Overlap). The propensity score $e(x, g) \equiv P(W = 1 \mid X = x, G = g)$ is less than 1:

$$e(X, G) < 1.$$  \hspace{1cm} (3)

In our context, Assumption 1 says that conditioned on the first reader score and applicant- and school-level covariates, submission of an LOR is independent of the potential admissions outcomes. Due to the selection mechanism we describe in Section 1.1.1, we believe that this is a reasonable starting point for estimating these impacts; see Rothstein (2017) and Appendix A.2 for alternatives. Assumption 2 corresponds to assuming that no applicant would have been guaranteed to submit a

---

$^5$There is a possibility of interference induced by the number of admitted applicants being capped. With 6874 admitted students, we consider the potential interference to be negligible.
letter of recommendation. Although some applicants were guaranteed to be invited to submit an LOR, we believe that this is a reasonable assumption for actually submitting a letter. In Section 5.1 we assess overlap empirically.

With this setup, let $m_0(x, g) = E[Y(0) \mid X = x, G = g]$ be the **prognostic score**, the expected control outcome conditioned on covariates $X$ and group membership $G$. Under Assumptions 1 and 2, we have the standard identification result:

$$
\mu_0 = E[m_0(X, G) \mid W = 1] = E \left[ \frac{e(X, G)}{1 - e(X, G)} Y \mid W = 0 \right].
$$

(4)

Therefore we can obtain a plug-in estimate for $\mu_0$ with an estimate of the prognostic score, $m_0(\cdot, \cdot)$, an estimate of propensity score, $e(\cdot, \cdot)$, or an estimate of the treatment odds themselves, $\frac{e(\cdot, \cdot)}{1 - e(\cdot, \cdot)}$.

We next review existing methods for such estimation, turning to our proposed weighting approach in the following section.

### 2.2 Related work: methods to estimate subgroup treatment effects

There is an extensive literature on estimating varying treatment effects in in observational studies; see Anoke et al. (2019) and Carvalho et al. (2019) for recent discussions. This is an active area of research, and we narrow our discussion here to methods that assess heterogeneity across pre-defined, discrete subgroups. We also focus on design-based approaches that do not use outcome information in constructing the estimators (Rubin, 2008). See Hill (2011); Künzel et al. (2019); Carvalho et al. (2019); Nie and Wager (2019) for discussions of approaches that instead focus on outcome modeling.

**Methods based on estimated propensity scores.** A canonical approach in this setting is Inverse Propensity Weighting (IPW) estimators for $\mu_0$ (see Green and Stuart, 2014). Traditionally, this proceeds in two steps: first estimate the propensity score $\hat{e}(x, g)$, e.g. via logistic regression; second, construct a weighted average of control unit outcomes,

$$
\hat{\mu}_0 = \sum_{W_i=0, G_i=g} \frac{\hat{e}(X_i, G_i) Y_i}{1 - \hat{e}(X_i, G_i)}
$$

(5)

where these are “odds of treatment” weights to target the ATT. A natural approach to estimating $\hat{e}(X_i, G_i)$, recognizing that $G_i$ is discrete, is to estimate a logistic model for treatment separately for each group or, equivalently, with full interactions between $G_i$ and (possibly transformed) covariates $\phi(X_i) \in \mathbb{R}^p$:

$$
\text{logit}(e(x, g)) = \alpha_g + \beta_g \cdot \phi(x).
$$

(6)

Estimating Equation (6) without any regularization can be problematic, due to complete separation, particularly when some groups are small; examples include $L^1$ penalization (Lee et al., 2019),
hierarchical Bayesian modelling (Li et al., 2013), and generalized boosted models (McCaffrey et al., 2004). Other possible modifications include Dong et al. (2020), who propose a stochastic search algorithm to estimate a similar model when the number of subgroups is large, and Li et al. (2017) and Yang et al. (2020), who propose overlap weights, which upweight regions of greater overlap. We explore overlap weights further in Section 5.

Under suitable assumptions and conditions, methods utilizing the estimated propensity score will converge to the true ATT asymptotically. However, these methods can often fail to achieve good covariate balance in the sample of interest; as we show in Section 5.1, these methods fail to balance covariates in the UC Berkeley LOR study. The key issue is that traditional IPW methods focus on estimating the propensity score itself (i.e., the conditional probability of treatment) rather than finding weights that achieve good in-sample covariate balance.

Balancing weights. Unlike traditional IPW, balancing weights estimators instead find weights that directly target in-sample balance. These methods have a long history in calibrated survey weighting (see, e.g. Deming and Stephan, 1940; Deville et al., 1993), and have recently been extensively studied in the observational study context (e.g. Hainmueller, 2011; Zubizarreta, 2015; Athey et al., 2018; Hazlett, 2018; Hirshberg et al., 2019). They have also been shown to estimate the propensity score with a loss function designed to achieve good balance (Zhao and Percival, 2016; Wang and Zubizarreta, 2019; Chattopadhyay et al., 2020). See Ben-Michael et al. (2020) for a recent review.

While balancing weights achieve better balance than the traditional IPW methods above, we must take special care to use them appropriately when estimating subgroup treatment effects. As we will show in Section 5.1, designing balancing weights estimators without explicitly incorporating the subgroup structure also fails to balance covariates within subgroups in the LOR study. We turn to designing such weights in the next section.

3 Approximate balancing weights for treatment effect variation

Now we describe a specialization of balancing weights that minimizes the bias for subgroup treatment effect estimates by incorporating the subgroup structure into the balance measure and optimizing for the “local balance” within each subgroup. First we will show that the error for a subgroup treatment effect estimate is bounded by the level of local imbalance within the subgroup. Then we describe a convex optimization problem to minimize the level of imbalance within each subgroup while ensuring global balance in the full sample. Next, we connect the procedure to IPW with a hierarchical propensity score model, using the procedure’s Lagrangian dual formulation. We conclude by describing how to augment the weighting estimate with an outcome model.
3.1 Local balance and differential confounding

For weights $\hat{\gamma} \in \mathbb{R}^n$, a linear weighting estimator for $\mu_{0g}$ is a weighted average of the control outcomes in the subgroup:

$$\hat{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) Y_i.$$  \hspace{1cm} (7)

In general, we are interested both in estimating the main effect through the overall ATT as well as interaction effects through the subgroup ATTs. We also want these estimates to be self-consistent in the sense that the overall mean is a weighted average of the subgroup-specific means: $\hat{\mu}_0 = \sum_{g=1}^K \frac{n_{1g}}{n} \hat{\mu}_{0g}$. We therefore want to find a single set of weights for both the overall and subgroup effects, rather than estimating separate weights for each.

We will compare the estimate $\hat{\mu}_{0g}$ to $\tilde{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) m_0(X_i, g)$, our best approximation to $\mu_{0g}$ if we knew the true prognostic score. Defining the residual $\varepsilon_i = Y_i - m_0(X_i, G_i)$, the error is

$$\hat{\mu}_{0g} - \tilde{\mu}_{0g} = \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) m_0(X_i, g)}_{\text{bias}_g} - \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} W_i m_0(X_i, g)}_{\text{noise}} + \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i \varepsilon_i}_{\text{bias}_g}. \hspace{1cm} (8)$$

Since the weights $\hat{\gamma}$ are design-based, they will be independent of the outcomes, and the noise term will be mean-zero and have variance proportional to the sum of the squared weights $\frac{1}{n_{1g}} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2$. At the same time, the conditional bias term, $\text{bias}_g$, depends on the imbalance in the true prognostic score $m_0(X_i, G_i)$. The idea is to bound this imbalance by the worst-case imbalance in all functions $m$ in a model class $\mathcal{M}$. While the setup is general, we describe the approach assuming that the prognostic score within each subgroup is a linear function of transformed covariates $\phi(X_i) \in \mathbb{R}^p$ with $L^2$-bounded coefficients; i.e., $\mathcal{M} = \{ m_0(x, g) = \eta_g \cdot \phi(x) \mid \| \eta_g \|_2 \leq C \}$. We can then bound the bias by the level of local imbalance within the subgroup via the Cauchy-Schwarz inequality:

$$|\text{bias}_g| \leq C \left\| \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right\|_2.$$  \hspace{1cm} (9)

This is the subgroup-specific analogue of arguments that relate the error in estimating the global ATT to the level of global balance (e.g., Zubizarreta, 2015; Athey et al., 2018; Hirshberg and Wager, 2019).

We now turn to finding a set of weights that achieve both good global balance and local balance;

\[6\] In the general case with heteroskedastic errors, the variance of the noise term is $\frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i^2 \text{Var}(\varepsilon_i) \leq \max_i \{ \text{Var}(\varepsilon_i) \} \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i^2$.

\[7\] See Wang and Zubizarreta (2019) for the case where the prognostic score can only be approximated by a linear function; see Hazlett (2018) for a kernel representation and Hirshberg et al. (2019) for a general nonparametric treatment.
the former to estimate the ATT and the latter to estimate the subgroup effects. Additionally, as we show in Appendix A.1, achieving good local balance also allows for the overall ATT estimate to be more robust to heterogeneity across subgroups.

3.2 Optimizing for both global and local balance

We now describe a convex optimization procedure to find weights that optimize for local balance while ensuring global balance across the sample. The idea is to stratify across subgroups and find approximate balancing weights within each stratum, while still constraining the overall level of balance. In our setting, we stratify on first reader score, URM status, the coarsened AI measure, and the college that the applicant is applying to; see Section 5. We then find weights $\hat{\gamma}$ that solve the following optimization problem:

$$\min_{\gamma} \sum_{g=1}^{K} \left\| \sum_{G_i=g,W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g,W_i=1} \phi(X_i) \right\|_2^2 + \frac{\lambda_g}{2} \sum_{G_i=G,W_i=0} \gamma_i^2$$

subject to

$$\sum_{W_i=0} \gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i)$$

$$\sum_{G_i=G,W_i=0} \gamma_i = n_{1g}$$

$$\gamma_i \geq 0 \quad \forall i = 1, \ldots, n$$

The optimization problem (10) has several key components. First, following Equation (9) we try to find weights that minimize the local imbalance for each stratum defined by $G$; this is a proxy for the stratum-specific bias. Second, we include an $L^2$ regularization term that penalizes the sum of the squared weights in the stratum; from Equation (8), we see that this is a proxy for the variance of the weighting estimator. For each stratum, the optimization problem includes a hyper-parameter $\lambda_g$ that negotiates the bias-variance tradeoff within that stratum. When $\lambda_g$ is small, the optimization prioritizes minimizing the bias through the local imbalance, and when $\lambda$ is large it prioritizes minimizing the variance through the sum of the squared weights. As a heuristic, we set $\lambda_g = \frac{1}{n_g}$; for larger strata where better balance is possible, this heuristic will prioritize balance — and thus bias — over variance; for smaller strata, by contrast, this will prioritize lower variance.

We also incorporate three constraints into the optimization problem. First, we constrain the weights to exactly balance the covariates globally over the entire population. This ensures that we are minimizing the overall bias as well as the bias within each stratum. This constraint is not crucial: if we could find weights that exactly balance the covariates within each stratum, then this would also yield exact balance globally. Typically, however, the sample sizes are too small to achieve exact balance within each stratum, and so this constraint at least guarantees global
balance. In Appendix A.1 we show that guaranteeing exact global balance can aid estimation of the ATT even when the outcome process can vary by subgroup as above. Second, we include a fine balance constraint (Rosenbaum et al., 2007): within each stratum the weights sum up to the number of treated units in that stratum, \( n_{1g} \). Since each stratum maps to only one subgroup, this also guarantees that the weights sum to the number of treated units in each subgroup. Finally, we restrict the weights to be non-negative, which stops the estimates from extrapolating outside of the support of the control units (King and Zeng, 2006). Together, these induce several stability properties, including that the estimates are sample bounded.

In our setting the strata \( G \) are part of a hierarchy: each stratum is a unique combination of first reader score, URM status, admissibility group, and college. Thus, we could also extend the optimization problem in Equation (10) to balance intermediate levels between global balance and local balance. Incorporating additional balance constraints for each intermediate level, is unwieldy in practice due to the proliferation of hyperparameters. Instead, we expand \( \phi(x) \) to include additional interaction terms between covariates and levels of the hierarchy. In our application, we interact the admissibility index with both URM status and the AI group, which means that we exactly balance AI within each URM-AI group.

Finally, we compute the variance of our estimator conditioned on the design \((X_1, Z_1, W_1), \ldots, (X_n, Z_n, W_n)\) or, equivalently, conditioned on the weights. The conditional variance is

\[
\text{Var}(\hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i = g} (1 - W_i) \hat{\gamma}_i^2 \text{Var}(Y_i).
\]  

Using the \( i \)th residual to estimate \( \text{Var}(Y_i) \) yields the empirical sandwich estimator for the treatment effect

\[
\hat{\text{Var}}(\hat{\mu}_{1g} - \hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i = g} W_i (Y_i - \hat{\mu}_{1g})^2 + \frac{1}{n_{1g}^2} \sum_{G_i = g} (1 - W_i) \hat{\gamma}_i^2 (Y_i - \hat{\mu}_{0g})^2,
\]  

where, as above, \( \hat{\mu}_{1g} \) is the average outcome for applicants in subgroup \( g \) who submit an LOR. This is the fixed-design Huber-White heteroskedastic robust standard error for the weighted average. See Hirshberg et al. (2019) for discussion on asymptotic normality and semi-parametric efficiency for estimators of this form.

### 3.3 Dual relation to partially pooled propensity score estimation

Thus far, we have motivated the approximate balancing weights approach by appealing to the connection between local bias and local balance. We now draw on recent connections between approximate balancing weights and (calibrated) propensity score estimation through the Lagrangian dual problem. The weights that solve optimization problem (10) correspond to estimating the
inverse propensity weights with a (truncated) linear odds function with the stratum $Z$ interacted with the covariates $\phi(X)$,\(^9\)

\[
\frac{P(W = 1 \mid X = x, Z = z)}{1 - P(W = 1 \mid X = x, Z = z)} = [\alpha_g + \beta_g \cdot \phi(x)]_+,
\]

where the coefficients $\beta_g$ are partially pooled towards a global model.

To show this, we first derive the Lagrangian dual. For each stratum $g$, the sum-to-$n_g$ constraint induces a dual variable $\alpha_g \in \mathbb{R}$, and the local balance measure induces a dual variable $\beta_g \in \mathbb{R}^p$. These dual variables are part of the balancing loss function for stratum $z$:

\[
L_g(\alpha_g, \beta_g) \equiv \sum_{W_i=0, G_i=g} [\alpha_g + \beta_g \cdot \phi(X_i)]_+^2 - \sum_{W_i=1, G_i=g} (\alpha_g + \beta_g \cdot \phi(X_i)) ,
\]

where $[x]_+ = \max\{0, x\}$. With this definition we can now state the Lagrangian dual.

**Proposition 1.** With $\lambda_g > 0$, if a feasible solution to (10) exists, the Lagrangian dual is

\[
\min_{\alpha, \beta_1, \ldots, \beta_J, \mu} \sum_{g=1}^{K} L_g(\alpha_g, \beta_g) + \sum_{z=1}^{J} \lambda_g \Vert \beta_g - \mu \beta \Vert_2^2 .
\]

If $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_J$ are the solutions to the dual problem, then the solution to the primal problem (10) is

\[
\hat{\gamma}_i = \left[ \hat{\alpha}Z_i + \hat{\beta}_Z \cdot \phi(X_i) \right]_+ .
\]

The Lagrangian dual formulation sheds additional light on the approximate balancing weights estimator. First, applying results on the connection between approximate balancing weights and propensity score estimation (e.g., Zhao and Percival, 2016; Wang and Zubizarreta, 2019; Hirshberg and Wager, 2019; Chattopadhyay et al., 2020), we see that this approach estimates propensity scores of the form (13). This corresponds to a fully interacted propensity score model where the coefficients on observed covariates vary across strata. Recall that we find approximate balancing weights for each stratum because the number of units per stratum might be relatively small; therefore we should not expect to be able to estimate this fully interacted propensity score well.

The dual problem in Equation (15) also includes a global dual variable $\mu$ induced by the global balance constraint in the primal problem (10). By penalizing the deviations between the stratum-specific variables and the global variables via the $L^2$ norm, $\| \beta_g - \mu \beta \|_2^2$, the dual problem partially pools the stratum-specific parameters towards a global model. Thus, we see that the approximate

\(^9\)The truncation arises from constraining weights to be non-negative, and the linear odds form arises from penalizing the $L^2$ norm of the weights. We can consider other penalties that will lead to different forms. See Ben-Michael et al. (2020) for a review of the different choices.
balancing weights problem in Equation (10) corresponds to a hierarchical propensity score model (see, e.g. Li et al., 2013), as in Section 2.2, fit with a loss function designed to provide covariate balance. Excluding the global constraint removes the global dual variable $\mu_\beta$, and the dual problem shrinks the stratum-specific variables $\beta_g$ towards zero without any pooling. In contrast, ignoring the local balance measure by setting $\lambda_g \to \infty$ constrains the stratum-specific variables $\beta_g$ to all be equal to the global variable $\mu_\beta$, resulting in a fully pooled estimator.

Finally, recall that in the primal problem (10), the hyperparameter $\lambda_g$ controlled the bias-variance tradeoff within stratum $z$ between prioritizing local balance or effective sample size. In the dual problem $\lambda_g$ performs the same role by controlling the level of partial pooling. When $\lambda_g$ is large the dual parameters are heavily pooled towards the global model, and when $\lambda_g$ is small the level of pooling is reduced. By setting $\lambda_g = \frac{1}{n_g}$ as above, larger strata will be pooled less than smaller strata.$^{10}$

### 3.4 Augmentation with an outcome estimator

Finally, we consider augmenting the balancing weights estimator with an outcome model, following similar proposals in a variety of settings (see, e.g. Athey et al., 2018; Hirshberg and Wager, 2019; Ben-Michael et al., 2019b). Analogous to bias correction for matching (Rubin, 1973) or model-assisted estimation in survey sampling (Särndal et al., 2003), the essential idea is to adjust the weighting estimator using an estimate of the bias. Specifically, we can estimate the prognostic score $m_0(x,g)$ with a working model $\hat{m}_0(x,g)$, e.g., with a flexible regression model. An estimate of the bias in group $g$ is then:

$$\hat{\text{bias}}_g = \frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i,g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i,g).$$

(17)

This is the bias due to imbalance in estimated prognostic score in group $g$ after weighting. With this estimate of the bias, we can explicitly bias-correct our weighting estimator, estimating $\mu_{0g}$ as

$$\hat{\mu}_{0g}^{\text{aug}} = \hat{\mu}_{0g} + \hat{\text{bias}}_g$$

$$= \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i Y_i + \left[ \frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i,g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i,g) \right].$$

(18)

Thus, if the balancing weights fail to achieve good covariate balance in a given subgroup, the working outcome model, $\hat{m}_0(X_i,g)$, can further adjust for any differences. See Ben-Michael et al. (2020) for further discussion.

---

$^{10}$It is also possible to have covariate-specific shrinkage by measuring imbalance in the primal problem (10) with a weighted $L^2$ norm, leading to an additional $p$ hyper-parameters. We leave exploring this extension and hyper-parameter selection methods to future work.
4 Simulation study

Before estimating the differential impacts of letters of recommendation, we first present simulations assessing the performance of our proposed approach versus traditional inverse propensity score weights fit via logistic regression. For \( n = 10,000 \) units, we draw \( d = 50 \) covariates \( X_{id} \sim N(0, 1) \) and subgroup indicators \( G_i \in \{1, \ldots, G\} \) as Multinomial\((\frac{1}{G}, \ldots, \frac{1}{G})\), where \( G \in \{10, 50\} \). We then use a separate logistic propensity score model for each group following Equation (6), \(^{11}\)

\[
\logit e(X_i, G_i) = \alpha_{G_i} + (\mu_\beta + U_\beta \odot B_\beta) \cdot X_i,
\]

and also use a separate linear outcome model for each group,

\[
Y_i(0) = \eta_0G_i + (\mu_\eta + U_\eta \odot B_\eta) \cdot X_i + \epsilon_i,
\]

where \( \epsilon_i \sim N(0, 1) \) and \( \odot \) denotes element-wise multiplication. We then draw group-specific treatment effects \( \tau_g \overset{iid}{\sim} N(0, 1) \) and set the treated potential outcome as \( Y_i(1) = Y_i(0) + \tau_{G_i}W_i \). The true ATT in simulation \( j \) is thus \( \tau_j = \frac{1}{n_1} \sum_{i=1}^{n} W_i(Y_i(1) - Y_i(0)) \).

We draw the fixed effects and varying slopes for each group according to a hierarchical model with sparsity. We draw the fixed effects as \( \alpha_0 \overset{iid}{\sim} N(0, 1) \) and \( \eta_0 \overset{iid}{\sim} N(0, 1) \). For the slopes, we first start with a mean slope vector \( \mu_\beta, \mu_\eta \in \left\{-\frac{3}{\sqrt{d}}, \frac{3}{\sqrt{d}}\right\}^K \), where each element is chosen independently with uniform probability. Then we draw isotropic multivariate normal random variables \( U_\beta, U_\eta \overset{iid}{\sim} MVN(0, I_d) \). Finally, we draw a set of \( d \) binary variables \( B_{gj}, B_{gj} \overset{iid}{\sim} \text{Bernoulli} \) with probability \( p = 0.25 \). The slope is then constructed as a set of sparse deviations from the mean vector: \( \mu_\beta + U_\beta \odot B_\beta \) for the propensity score and \( \mu_\eta + U_\eta \odot B_\eta \) for the outcome model.

For \( j = 1, \ldots, m \) with \( m = 500 \) Monte Carlo samples, we estimate the treatment effects for group \( g \), \( \hat{\tau}_{gj} \), and the overall ATT, \( \hat{\tau}_j \), and compute a variety of metrics. Following the metrics studied by Dong et al. (2020), for subgroup treatment effects we compute (a) the mean absolute bias across the \( G \) treatment effects, \( \frac{1}{m} \sum_{j=1}^{m} \frac{1}{G} \sum_{g=1}^{G} |\hat{\tau}_{gj} - \tau_g| \), and (b) the mean root mean square error \( \sqrt{\frac{1}{mG} \sum_{j=1}^{m} \sum_{g=1}^{G} (\hat{\tau}_{gj} - \tau_g)^2} \). For the overall ATT we measure (a) the absolute bias \( \frac{1}{m} \sum_{j=1}^{m} |\hat{\tau}_j - \tau_j| \) and (b) the root mean square error \( \sqrt{\frac{1}{m} \sum_{j=1}^{m} (\hat{\tau}_j - \tau_j)^2} \).

We compute treatment effects for five weighting estimators:

- **Partially pooled balancing weights**: approximate balancing weights that solve (10), using \( G \) as the stratifying variable and prioritizing local balance by setting \( \lambda_g = \frac{1}{n_1g} \).

- **Fully pooled balancing weights**: approximate balancing weights that solve (10), but ignore local balance by setting \( \lambda \) to be very large and fully pooling towards the global model. This is

\[^{11}\]The logistic specification differs from the truncated linear odds in Equation 13. If the transformed covariates \( \phi(X_i) \) include a flexible basis expansion, the particular form of the link function will be less important.
Figure 3: Performance of approximate balancing weights and traditional IPW with logistic regression for estimating subgroup treatment effects.

equivalent to stable balancing weights (Zubizarreta, 2015) with an exact balance constraint.

- **No pooled balancing weights**: approximate balancing weights that solve (10), but without the exact global balance constraint.

- **Full interaction IPW**: traditional IPW with a fully interacted model that estimates a separate propensity score within each stratum as in Equation (6).

- **Fixed effects IPW**: full interaction IPW with stratum-specific coefficients constrained to be equal to a global parameter $\beta_g = \beta$ for all $g$.

We fit each logistic regression via maximum likelihood with an $L^1$ penalty to induce sparsity; for the fully interacted specification we also include a set of global parameters $\mu_\beta$ so that the slope for group $g$ is $\mu_\beta + \Delta_g$, with an $L^1$ penalty for each component. For both logistic regression specifications, we estimate the models with glmnet (Friedman et al., 2010) using an $L^1$ penalty on the parameters with hyperparameter chosen through 5-fold cross validation.\(^\text{12}\)

Figure 3 shows the results for the overall ATT and for subgroup effects. We see that with 10 subgroups, prioritizing local balance with either the partially pooled or no-pooled approximate balancing approaches yields lower bias and RMSE than ignoring local balance entirely with the fully pooled approach. These approaches also have better performance than either of the traditional logistic regression approaches. In this setting where there are 1,000 units per group, it is possible to achieve good balance in each group and there is no benefit to partially pooling via the exact global balance constraint. However, with 50 subgroups and 200 units per group, it is difficult to balance within each subgroup and there is a benefit to partial pooling. Partially pooling balancing

\(^{12}\)This amounts to partial pooling towards a sparse global model with sparse deviations. We can also consider partially pooling via multilevel modelling.
weights yields much lower bias for the overall ATT than the no-pooled approach, and has lower bias for the subgroup effects as well, although this comes at the cost of higher RMSE for subgroup effects.

5 Differential impacts of letters of recommendation

We now turn to estimating the differential impacts of letters of recommendation on admissions decisions. We focus on the eight subgroups defined in Table 1, based on the interaction between URM status (2 levels) and admissibility index (4 levels). Due to the selection mechanism described in Section 1.1, however, it is useful to create even more fine-grained strata and then aggregate to these eight subgroups. Specifically, we define $G = 41$ fine-grained strata based on URM status, AI grouping, first reader score, and college applied to.\textsuperscript{13} While we are not necessarily interested in treatment effect heterogeneity across all 41 strata, this allows us to exactly match on key covariates and then aggregate to obtain the primary subgroup effects.

Another key component in the analysis is the choice of transformation of the covariates $\phi(\cdot)$. Because we have divided the applicants into many highly informative strata, we choose $\phi(\cdot)$ to include all of the raw covariates. Additionally, because of the importance of the admissibility index, we also include a natural cubic spline for AI with knots at the sample quantiles. Finally, we include the output of the admissions model and a binary indicator for whether the predicted probability of a “Possible” score is greater than 50%. If desired, we could also consider other transformations such as a higher order polynomial transformation, using a series of basis functions for all covariates, or computing inner products via the kernel trick to allow for an infinite dimensional basis (see, e.g. Hazlett, 2018; Wang and Zubizarreta, 2019; Hirshberg and Wager, 2019). We further prioritize local balance in the admissibility index by exactly balancing the AI within each URM $\times$ subgroup. As we discuss above, this ensures local balance in the admissibility index at an intermediate level of the hierarchy between global balance and local balance. Finally, we standardize each component of $\phi(X)$ to have mean zero and variance one.

5.1 Diagnostics: local balance checks and assessing overlap

Before estimating effects, we first assess the level of local balance within each subgroup, following the discussion in Section 3.1. We consider the five estimators described in Section 4. We also use the estimated fully interacted propensity score model to create subgroup overlap weights as in Yang et al. (2020).

\textsuperscript{13}Of the 48 possible strata, we drop 7 strata where no applicants submitted a letter of recommendation. These are non-URM applicants in both colleges in the two lowest AI strata but where the first reader assigned a “Yes” or “No”. This accounts for \(~2\%\) of applicants. The remaining 41 strata have a wide range of sizes with a few very large strata. Min: 15, p25: 195, median: 987, p75: 1038, max: 8000
Figure 4 shows the distribution of the imbalance in each of the 51 (standardized) components of \( \phi(X) \), for the three balancing weights approaches as well as the fully interacted IPW estimator. The fully interacted IPW approach has very poor balance overall, due in part to the difficulty of estimating the high-dimensional propensity score model. As expected, both the fully- and partially-pooled balancing weights achieve perfect balance overall; however, only the partially pooled balancing weights achieve excellent local balance. The partially- and no-pooled approaches have similar global and local balance overall, but the partially-pooled approach sacrifices a small amount of local balance for an improvement in global balance. Appendix Figure B.3 shows these same metrics for the fixed effects IPW and overlap weights, which uses the same propensity score estimates as in the fully interacted IPW approach. Both yield poor local balance.

Appendix Figure B.4 shows imbalance in the one-dimensional summary admissibility index. Our proposed approach, which directly balances this summary index within URM and AI subgroups, again achieves excellent balance overall and within each group. This is not true for other approaches, especially fully interacted IPW weights, which fail to achieve reasonable balance in the admissibility index for most subgroups, with worse imbalance relative to the unweighted comparisons for some subgroups. Here we see the effect of partial pooling. The no-pooled approach—only targeting balance within the fine-grained strata, ignoring global balance as well as balance in our primary subgroups—fails to achieve good balance in some subgroups, notably for high admissibility URM applicants, while the partially pooled approach achieves exact balance by design.

Finally, we assess overlap within each subgroup. A key benefit of weighting approaches is that any overlap issues manifest in the distribution of our weights \( \hat{\gamma} \). Figure 5 plots the distribution
of the weights over the comparison applicants by URM status and AI group, normalized by the number of treated applicants in the subgroup. The vast majority of control units receive zero weight and are excluded from the figure. Of the 28,556 applicants who did not submit an LOR, only 5,702 (20%) receive positive weight. This is indicative of a lack of “left-sided” overlap: very many applicants who did not submit a letter of recommendation had nearly zero odds of doing so in the pilot program. This is problematic for estimating the overall average treatment effect, but is less of a concern when we focus on estimating the average treatment effect on the treated.

For each AI subgroup we also see that the distribution of weights is skewed more positively for the non-URM applicants. In particular, for the lower AI, non-URM subgroups we see a non-trivial number of comparison applicants that “count for” over 2% of the re-weighted sample, with a handful of outliers that are equivalent to over 5%. While large weights do not necessarily affect the validity of the estimator — though they suggest caution in terms of “right-sided” overlap — large weights decrease the effective sample size, reducing the precision of our final estimates. Appendix Figure B.5 shows the effective sample size, \( n_{1g}/\sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2 \), for each subgroup \( g \). We see that the URM subgroups have larger effective sample sizes than the non-URM subgroups, with particularly stark differences for the lower AI subgroups. Furthermore, for all non-URM subgroups with AI ≤ 20%, the effective sample size is ≤ 100. From this, we should expect to have far greater precision in the estimates for URM applicants than non-URM applicants.

5.2 Treatment effect estimates

After assessing local balance and overlap, we can now turn to estimating the differential impacts of letters of recommendation. Figure 6 shows (1) the percent of applicants who submitted an LOR who were accepted, \( \hat{\mu}_{1g} \) (2) the imputed counterfactual mean, \( \hat{\mu}_{0g} \) and (3) the ATT, \( \hat{\mu}_{1g} - \hat{\mu}_{0g} \). The standard errors are computed via the sandwich estimator in Equation (12). Overall, we estimate an
Figure 6: Estimated treated and control means and treatment effect of letters of recommendation on admission ± two standard errors, overall and by URM status and Admissibility Index.

We observe an increase in admission rates of 5 percentage points (pp). While we estimate a larger effect for non-URM applicants (6 pp) than URM applicants (4 pp), there is insufficient evidence to distinguish between the two effects. Second, we see a roughly positive trend between treatment effects and the AI, potentially with a peak for the 10%-20% group. This is driven by the very small estimated effect for applicants with AI < 5% and who are thus very unlikely to be accepted a priori. Substantively, this corresponds to letters of recommendation having a very low impact for applicants unlikely to be accepted, but a larger impact for applicants that are perhaps on the cusp on acceptance.

Appendix Figure B.6 shows an estimate of the log risk ratio, \( \log \frac{E[Y(1)|G=g]}{E[Y(0)|G=g]} \), for the subgroups. From the estimated risk ratios, we see that this pattern, while noisy, is consistent with impacts that are roughly constant on the multiplicative scale, perhaps with a dip for both the low and high admissibility applicants.

Figure 7 further stratifies the subgroups, showing the effects jointly by URM status and AI. While the point estimate for the overall increase in admission rates is slightly larger for non-URM applicants than for URM applicants, this is mainly a composition effect. For applicants very unlikely to be admitted (AI < 5%) the point estimates are nearly identical for URM and non-URM applicants, although the URM subgroup is estimated much more precisely. For the next two levels of the admissibility index (AI between 5% and 20%), URM applicants have a higher estimated impact, with imprecise estimates for non-URM applicants. For the highest admissibility groups (AI > 20%), non-URM applicants have larger positive effects, though again these estimates are
Figure 7: Estimated treated and control means and treatment effect of letters of recommendation on admission ± two standard errors, further broken down by URM status interacted with the Admissibility Index.

noisy. Since URM applicants have lower AI on average, the overall estimate is also lower for URM applicants. Furthermore, the peak in the effect for middle-tier applicants is more pronounced for URM applicants than non-URM applicants. From Figure 7a we see that this is primarily because high admissibility URM applicants with a letter of recommendation are admitted at very high rates; the imputed baseline after re-weighting is similarly large.

We also consider augmenting the weighting estimator with an estimate of the prognostic score, \( \hat{m}(x, g) \). In Appendix Figure B.7 we show estimates after augmenting with ridge regression; we compute standard errors via Equation (12), replacing \( Y_i - \hat{\mu}_0 g \) with the empirical residual \( Y_i - \hat{m}(X_i, g) \). Because the partially pooled balancing weights achieve excellent local balance for \( \phi(X) \), augmenting with a model that is also linear in \( \phi(X) \) results in minimal adjustment. We therefore augment with a nonlinear outcome model, random forests. Tree-based estimators are a natural choice for a nonlinear outcome model, creating “data-dependent strata” similar in structure to the strata we define for \( G \). For groups where the weights \( \hat{\gamma} \) have good balance across the estimates \( \hat{m}(x, g) \), there will be little adjustment due to the outcome model. Conversely, if the raw and bias-corrected estimate disagree for a subgroup, then the weights have poor local balance across important substantive data-defined strata. For these subgroups we should be more cautious of our estimates.

Figure 8 shows the augmented effect estimates relative to the un-augmented estimates; the
Figure 8: Estimated effect of letters of recommendation on admission rates with and without augmentation.

(a) Overall and by URM status and AI.
(b) By URM status interacted with AI.

In the Appendix we consider alternative estimates. First, Appendix Figure B.8 shows the estimated effects on admission rates using all five weighting procedures we consider above. Despite failing to achieve good local balance, the IPW approaches and fully pooled balancing weights approach yield effect estimates that are similar to our proposed approach. The overlap weighting approach, however, leads to substantively different conclusions, perhaps due to the change in the estimand. These differences appear to be driven by that estimator’s negative estimated effect of LORs for high admissibility, non-URM applicants, suggesting that there are other substantively important sources of heterogeneity beyond URM status and admissibility.

Second, we consider effects on an intermediate outcome: whether the second reader — who has access to the LOR — gives a “Yes” score. Because these are design-based weights, we use the same set of weights to estimate effects on both second reader scores and admissions decisions. With this outcome we can also make use of a within-study design to estimate treatment effects, leveraging scores from additional third readers who did not have access to the letters of recommendation;
we describe this design in Appendix A.2. Appendix Figures B.9 and B.10 show the results for both approaches. Overall for second reader scores we see a similar structure of heterogeneity as for admission rates, although there does not appear to be an appreciable decline in the treatment effect for the highest admissibility non-URM applicants. The two distinct approaches yield similar patterns of estimates overall, with the largest discrepancy for applicants with a predicted probability of admission between 5% and 10%, particularly for non-URM applicants. However, this group has a very low effective sample size, and so the weighting estimates are very imprecise.

Taken together, these results paint a relatively clear picture of differential impact of letters of recommendation across applicants’ a priori probability of admission. Treatment effects are low for applicants who are unlikely to be accepted and high for applicants on the margin for whom letters provide useful context, with some evidence of a dip for the highest admissibility applicants. Our estimates of differential impacts between URM and non-URM students are more muddled, due to large sampling errors, and do not support strong conclusions. Point estimates indicate that LORs benefit URM applicants more than they do non-URM applicants at all but the highest academic indexes. Because non-URM applicants are overrepresented in the high-AI category, the point estimate for the average treatment effect is larger for non-URMs; however, there is insufficient precision to distinguish between the two groups.

6 Discussion

Estimating heterogeneous treatment effects and assessing treatment effect variation in observational studies is a challenge, even for pre-specified subgroups. Focusing on weighting estimators that estimate subgroup treatment effects by re-weighting control outcomes, we show that the estimation error depends on the level of local imbalance between the treated and control groups after weighting. We then present a convex optimization problem that finds approximate balancing weights that directly target the level of local imbalance within each subgroup, while ensuring exact global balance to also estimate the overall effect. Using this method to estimate heterogeneous effects in the UC Berkeley letters of recommendation pilot study, we find evidence that letters of recommendation lead to better admissions outcomes for stronger applicants, with mixed evidence of differences between URM and non-URM applicants.

There are several directions for future work. First, we directly estimate the effect of submitting an LOR among those who submit. However, we could instead frame the question in terms of non-compliance and use the invitation to submit an LOR as an instrument for submission. Using the approximate balancing weights procedure described above we could adjust for unequal invitation probabilities, and estimate the effect on compliers via weighted two-stage least squares. Second, we could consider deviations from the ignorability assumption via a sensitivity analysis. One potential path is to extend the balancing weights sensitivity procedure from Soriano et al. (2020) to the setting with distinct subgroups. Third, we could adapt our approach to explore treatment effect
variation in other types of observational studies, for instance in settings that mimic the structure of multisite trials.
A Appendix

A.1 Global and local balance in estimating the ATT

Even when estimating the overall treatment effect, \( \tau \), the imbalance within each subgroup continues to play a key role in the bias. To see this, note that the estimated overall counterfactual mean for the treated units is \( \hat{\mu}_0 = \frac{1}{n_1} \sum_{g=1}^{K} n_{1g} \hat{\mu}_{0g} \). Again we compare to our best estimate if we knew the prognostic score, \( \tilde{\mu}_0 = \frac{1}{n_1} \sum_{g=1}^{K} n_{1g} \tilde{\mu}_{0g} \), and see that the local imbalance plays a part. The error is

\begin{align*}
\hat{\mu}_0 - \tilde{\mu}_0 &= \bar{\eta} \cdot \left( \frac{1}{n_1} \sum_{i=1}^{n} n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^{n} W_i \phi(X_i) \right) \\
&\quad + \frac{1}{n_1} \sum_{g=1}^{K} n_{1g} (\eta_g - \bar{\eta}) \cdot \left( \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n} \hat{\gamma}_i (1 - W_i) \epsilon_i,
\end{align*}

(21)

where \( \bar{\eta} \equiv \frac{1}{K} \sum_{g=1}^{K} \eta_g \) is the average of the model parameters across all subgroups. Again using Cauchy-Schwarz we see that the overall bias is controlled by the local imbalance within each subgroup as well as the global balance across subgroups:

\begin{align*}
|\text{bias}| &\leq \|\bar{\eta}\|_2 \left| \frac{1}{n_1} \sum_{i=1}^{n} n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^{n} W_i \phi(X_i) \right|_2 \\
&\quad + \sum_{g=1}^{G} \frac{n_{1g}}{n_1} \|\eta_g - \bar{\eta}\|_2 \left| \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right|_2. 
\end{align*}

(22)

Equation (22) shows that the relative importance of local and global balance for estimating the overall ATT is controlled by the level of similarity in the outcome process across groups. In the extreme case where the outcome process does not vary across groups — i.e., \( \eta_g = \bar{\eta} \) for all \( g \) — then controlling the global balance is sufficient to control the bias. In the other extreme where the outcome model varies significantly across subgroups — e.g., \( \|\eta_g - \bar{\eta}\|_2 \) is large for all \( g \) — we will primarily seek to control the local imbalance within each subgroup in order to control the bias for the ATT. In general, we will want to achieve good local balance within each subgroup and good global balance across subgroups.

A.2 Within-subject comparison

We compare the weighting estimates for the effect of submitting an LOR on the second reader scores to estimates exploiting an additional feature of the pilot study. After the admissions process concluded, 10,000 applicants who submitted letters were randomly sampled and the admissions office recruited several readers to conduct additional evaluations of the applicants (Rothstein, 2017).
During this supplemental review cycle, the readers were not given access to the letters of recommendation, but otherwise the evaluations were designed to be as similar as possible to the second reads that were part of the regular admissions cycle; in particular, readers had access to the first readers’ scores.

With these third reads we can estimate the treatment effect by taking the average difference between the second read (with the letters) and the third read (without the letters). One major issue with this design is that readers might have applied different standards during the supplemental review cycle. Regardless, if the third readers applied a different standard consistently across URM and admissibility status, we can distinguish between treatment effects within these subgroups. We show the results in Figures B.9 and B.10.

A.3 Proofs

Proof of Proposition 1. First, we will augment the primal optimization problem in Equation (10) with auxiliary covariates $E_1, \ldots, E_j$ so that $E_g = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i)$. Then the optimization problem becomes:

$$
\min_{\gamma} \sum_{z=1}^{J} \frac{1}{2\lambda_g} \|E_j\|_2^2 + \frac{\lambda_g}{2} \sum_{z_i=W_i=0} \gamma_i^2 + \mathcal{I}(\gamma_i \geq 0)
$$

subject to

$$
\gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i)
$$

$$
E_j = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i), \ z = 1, \ldots, J
$$

$$
\sum_{G_i=g, W_i=0} \gamma_i = n_{1g},
$$

where $\mathcal{I}(x \geq 0) = \{0 \ x \geq 0 \ \infty \ x < 0$ is the indicator function. The first constraint induces a Lagrange multiplier $\mu_\beta$, the next $J$ constraints induce Lagrange multipliers $\delta_1, \ldots, \delta_j$, and the sum-to-one constraints induce Lagrange multipliers $\alpha_1, \ldots, \alpha_J$. Then the Lagrangian is

$$
\mathcal{L}(\gamma, E, \mu_\beta, \delta, \alpha) = \sum_{z=1}^{J} \left[ \frac{1}{2\lambda_g} \|E_j\|_2^2 - E_j \cdot \delta_j + \sum_{G_i=g, W_i=0} \frac{1}{2} \gamma_i^2 + \mathcal{I}(\gamma_i \geq 0) - \gamma_i(\alpha + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \right]
$$

$$
+ \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i))
$$

(24)
The dual objective is:

\[
q(\mu, \delta, \alpha) = \sum_{z=1}^{J} \left[ \min_{E_j} \left\{ \frac{1}{2\lambda g} \|E_j\|^2_2 - E_j \cdot \delta_j \right\} + \sum_{G_i=g, W_i=0} \min_{\gamma_i \geq 0} \left\{ \frac{1}{2} \gamma_i^2 - \gamma_i(\alpha + (\mu \beta + \delta_j) \cdot \phi(X_i)) \right\} \right]
+ \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} \left( 1 + (\mu \beta + \delta_j) \cdot \phi(X_i) \right)
\]

Note that the inner minimization terms are the negative convex conjugates of \( \frac{1}{2} \|x\|^2_2 \) and \( \frac{1}{2} x^2 + I(X \geq 0) \), respectively. Solving these inner optimization problems yields that

\[
q(\mu, \delta, \alpha) = - \sum_{z=1}^{J} \left[ \frac{\lambda g}{2} \|\delta_j\|^2_2 + \sum_{G_i=g, W_i=0} [\alpha_j + (\mu \beta + \delta_j) \cdot \phi(X_i)]^2_+ \right] + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu \beta + \delta_j) \cdot \phi(X_i))
\]

Now since there exists a feasible solution to the primal problem (10), from Slater’s condition we see that the solution to the primal problem is equivalent to the solution to \( \max_{\mu, \delta, \alpha} q(\mu, \delta, \alpha) \). Defining \( \beta_j \equiv \mu \beta + \delta_j \) gives the dual problem (15). Finally, note that the solution to the minimization over the weights in Equation (25) is \( \gamma_i = [\alpha_j + \beta_j \cdot \phi(X_i)]_+ \), which shows how to map from the dual solution to the primal solution.

B Additional figures and tables
Figure B.1: (a) Mean square error (Brier score) and (b) admission rates for the Admissibility Index predicting the 2016-2017 cycle admissions results, computed in 2% groups.

<table>
<thead>
<tr>
<th>College</th>
<th>URM</th>
<th>AUC</th>
<th>Brier Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letters and Science</td>
<td>URM</td>
<td>89%</td>
<td>9%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>88%</td>
<td>11%</td>
</tr>
<tr>
<td>Engineering</td>
<td>URM</td>
<td>92%</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>Not URM</td>
<td>89%</td>
<td>11%</td>
</tr>
</tbody>
</table>

Table B.2: AUC and Brier score for the Admissibility Index predicting the 2016-2017 cycle admissions results.

Figure B.2: ROC curve for Admissibility Index predicting the 2016-2017 cycle admissions results.
Figure B.3: Distribution of covariate balance measured by the mean standardized difference for different weighting methods.

(a) Overall and by URM status and AI.  
(b) By URM status interacted with AI.

Figure B.4: Imbalance in the admissibility index after weighting relative to before weighting, overall and within each subgroup. For several subgroups, the fully pooled balancing weights procedure results in increased imbalance in the admissibility index, denoted by an arrow.
Figure B.5: Effective sample size for each subgroup, with weights solving the approximate balancing weights problem (10).

Figure B.6: Estimated log risk ratio of admission with and without letters of recommendation ± two standard errors computed via the delta method, overall and by URM status and AI.
Figure B.7: Estimated effect of letters of recommendation on admission rates with and without augmentation via ridge regression.

(a) Overall and by URM status and AI.  
(b) By URM status interacted with AI.

Figure B.8: Estimated effect of letters of recommendation on admission rates for comparable weighting estimators.

(a) Overall and by URM status and AI.  
(b) By URM status interacted with AI.
(a) Partially pooled balancing weights  
(b) Within-subject design

Figure B.9: Effects on second reader scores overall, by URM status, and by AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.

(a) Partially pooled balancing weights  
(b) Within-subject design

Figure B.10: Effects on second reader scores by URM status interacted with AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.
C Additional simulation results

Figure C.1: Performance of approximate balancing weights for estimating subgroup treatment effects as $\lambda$ varies.
References


