Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing[*]

Susan Athey[†] Guido W. Imbens[‡] Stefan Wager[§]

Current version June 2016

Abstract

There are many settings where researchers are interested in estimating average treatment effects and are willing to rely on the unconfoundedness assumption, which requires that the treatment assignment be as good as random conditional on pre-treatment variables. The unconfoundedness assumption is often more plausible if a large number of pre-treatment variables are included in the analysis, but this can worsen the finite sample properties of standard approaches to treatment effect estimation. There are some recent proposals on how to extend classical methods to the high dimensional setting; however, to our knowledge, all existing method rely on consistent estimability of the propensity score, i.e., the probability of receiving treatment given pre-treatment variables. In this paper, we propose a new method for estimating average treatment effects in high dimensional linear settings that attains dimension-free rates of convergence for estimating average treatment effects under substantially weaker assumptions than existing methods: Instead of requiring the propensity score to be estimable, we only require overlap, i.e., that the propensity score be uniformly bounded away from 0 and 1. Procedurally, out method combines balancing weights with a regularized regression adjustment.

Keywords: Causal Inference, Potential Outcomes, Propensity Score, Sparse Estimation

1 Introduction

In many observational studies, researchers are interested in estimating average causal effects. A common approach is to assume that conditional on observed features of the units, assignment to the treatment is as good as random, or unconfounded; see, e.g., Rosenbaum and Rubin [1983] and Imbens and Rubin [2015] for general discussions. There is a large literature on adjusting for differences in observed features between the treatment and control groups under unconfoundedness; some popular methods include regression, matching, propensity score weighting and subclassification, as well as doubly-robust combinations thereof [e.g., Abadie and Imbens, 2006, Heckman et al., 1998, Hirano et al., 2003, Robins and Rotnitzky, 1995, Rosenbaum, 2002].

In practice, in order to make the assumption of unconfoundedness more plausible, researchers may need to account for a substantial number of features or observed confounders. For example, in an observational study of the effect of flu vaccines on hospitalization, we may be concerned that only controlling for differences in the age and sex distribution between controls and treated may not be sufficient to eliminate biases. In contrast, controlling for detailed medical histories and personal characteristics may make unconfoundedness more plausible. But the formal asymptotic theory in the earlier literature only considers the case where the sample size increases while the number of features remains fixed, and so

[*] We are grateful for detailed comments from Bryan Graham, Chris Hansen and Nishanth Mundru, and for discussions with seminar participants at Stanford University and the Cowles Econometrics Conference.

[†] Professor of Economics, Graduate School of Business, Stanford University, and NBER, athey@stanford.edu.

[‡] Professor of Economics, Graduate School of Business, Stanford University, and NBER, imbens@stanford.edu.

[§] Department of Statistics, Stanford University, swager@stanford.edu.

approximations based on those results may not yield valid inferences in settings where the number of features is large, possibly even larger than the sample size.

There has been considerable recent interest in adapting methods from the earlier literature to high-dimensional settings. Belloni et al. [2014, 2016] show that biases can arise when applying standard high-dimensional methods such as the lasso to estimating the outcome relationship. The reason for this bias is that the lasso focuses solely on accurate prediction of outcomes, at the expense of adjusting for covariates that affect treatment assignment, that is, covariates that enter in the propensity score. Belloni et al. [2014] propose an augmented variable selection scheme to avoid this effect, while Farrell [2015] discusses how a doubly robust approach to average treatment effect estimation in high dimensions can also be used to compensate for the bias of the lasso. Despite the breadth of research on the topic, a common requirement of all these methods is that they all rely on consistent estimability of the propensity score, i.e., the conditional probability of receiving treatment given the features. For example, several of the above methods assume that the propensity scores can be consistently estimated using a sparse logistic model.

In this paper, we show that efficient inference of average treatment effects in high dimensions is possible under substantially weaker assumptions. Rather than assuming that the propensity score is estimable, we only require overlap, i.e., that the propensity score be uniformly bounded away from 0 and 1 for all values in the support of the pretreatment variables. In particular, our results do not rely on a sparse propensity model—or even a well-specified logistic propensity model. Given this overlap assumption, as well as standard linearity and sparsity assumptions on the model relating outcomes to pretreatment variables, we show that our estimator has at most the same asymptotic error as the best semi-parametric average treatment effect estimators in low dimensions [Bickel et al., 1998].

Our approach builds on the classical literature on weighted estimation of treatment effects, going back to the work of Rosenbaum and Rubin [1983] who showed that controlling for the propensity score is sufficient to remove all biases associated with observed covariates. Recent studies have sought to extend the applicability of this result by using machine learning techniques to estimate the propensity score, in combination with conventional methods for estimating average treatment effects given the estimated propensity score: McCaffrey et al. [2004] recommend estimating the propensity score using boosting and then use inverse propensity weighting, while Westreich et al. [2010] consider support vector machines, neural networks, and classification trees. In related approaches, Chan et al. [2015], Graham et al. [2012, 2016], Hainmueller [2012], Imai and Ratkovic [2014] and Zubizarreta [2015] propose weighting methods where the weights are not equal to the inverse of the propensity score but are chosen explicitly to optimize balance between the covariate distributions in the treatment and control groups.

None of these methods, however, achieve systematically good performance in high dimensions. The reason plain propensity-based methods fall short in high dimensions is closely related to the reason why pure lasso regression adjustments are not efficient: In high dimensions, it is not in general possible to exactly balance all the features, and small imbalances can result in substantial biases in the presence of strong effects. Our proposal starts from an attempt to remove these biases by first fitting a standalone pilot model to capture any strong effects, and then applying weighting to the residuals.

We study the following two-stage approximate residual balancing algorithm that tightens the connection between the estimation strategy and the goal of estimating the average treatment effect. First, we fit a regularized linear model for the outcome given the features separately in the two treatment groups. In the current paper we focus on the elastic net [Zou and Hastie, 2005] and the lasso [Chen et al., 1998, Tibshirani, 1996] for this component, and present formal results for the latter. In a second stage, we re-weight the first stage residuals using weights that approximately balance all the features. Here we follow Zubizarreta [2015] in focusing on the properties of the weights in terms of the implied balance and their variance, rather than the fit of the propensity score, as the formal criterion to choose weights. Approximate balancing on all pretreatment variables (rather than exact balance on a subset of features, as in a regularized regression, or weighting using a regularized propensity model that may not be able to capture all relevant dimensions) allows us to guarantee that bias that might arise from failing to adjust for a large number of potentially weak confounders can be bounded.

In our simulations, we find that three features of the algorithm are important: (i) the direct covariance adjustment based on the outcome data with regularization to deal with the large number of features, (*ii*) the weighting using the relation between the treatment and the features, and (*iii*) the fact that the weights are based on direct measures of balance rather than on estimates of the propensity score, again with regularization to take account of the many features.

The finding that both weighting and regression adjustment are important is similar to conclusions drawn from the earlier literature on doubly robust estimation in low dimensions [Robins and Rotnitzky, 1995, Robins et al., 1995], where combining both techniques was shown to weaken the assumptions required to achieve consistent estimation of average treatment effects. In our setting, this pairing is not just helpful in terms of robustness, and is in fact required for efficiency: Neither regression adjustments nor approximately balanced weighting of the outcomes alone can achieve the optimal rate of convergence. Meanwhile, the finding that weights designed to achieve balance perform better than weights based on the propensity score is consistent with findings in Chan et al. [2015], Graham et al. [2012, 2016], Hainmueller [2012], and Zubizarreta [2015]. The current paper is the first to combine direct covariance adjustment with such balancing weights in a high-dimensional setting where regularization is required.

Note that if we are not willing to assume an approximately sparse propensity model as in Belloni et al. [2014], then the average treatment effect is a generic dense functional of the outcome model. As shown by Cai and Guo [2015], constructing valid confidence intervals of length $\mathcal{O}(1/\sqrt{n})$ for such a dense functional is in general impossible no matter how sparse the outcome model is; see also related results by Collier et al. [2015] for the Gaussian sequence model. We avoid the impossibility result of Cai and Guo [2015] by exploiting the overlap assumption that the propensity score is bounded away from zero and one, which enables us to take advantage of special structure in the causal inference problem.

Our paper is structured as follows. First, in Section 2, we motivate our two-stage procedure using a simple bound for its estimation error. Then, in Section 3, we provide a formal analysis of our procedure under high-dimensional asymptotics, and we identify conditions under which approximate residual balancing is asymptotically Gaussian and allows for practical inference about the average treatment effect with dimension-free rates of convergence. Finally, in Section 4, we conduct a simulation experiment, and find our method to perform well in a wide variety of settings relative to other proposals in the literature.

2 Estimating Average Treatment Effects in High Dimensions

2.1 Setting and Notation

Our goal is to estimate average treatment effects in the potential outcome framework, or Rubin Causal Model [Rubin, 1974, Imbens and Rubin, 2015]. For each unit in a large population there is pair of (scalar) potential outcomes, $(Y_i(0),\, Y_i(1))$. Each unit is assigned to the treatment or not, with the treatment indicator denoted by $W_i \in \{0,\, 1\}$. Each unit is also characterized by a vector of covariates or features $X_i \in \mathbb{R}^p$, with p potentially large, possibly larger than the sample size. For a random sample of size n from this population, we observe the triple $(X_i,\, W_i,\, Y_i^{\text{obs}})$ for $i = 1,\, \ldots,\, n$, where

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1, \end{cases} \tag{1}$$

is the realized outcome, equal to the potential outcome corresponding to the actual treatment received. The total number of treated units is equal to n_{t} and the number of control units equals n_{c}. We will frequently use the short-hand $\mathbf{X}_{\text{c}}$ and $\mathbf{X}_{\text{t}}$ for the feature matrices corresponding only to control or treated units respectively. We write the propensity score, i.e., the conditional probability of receiving the treatment given features, as $e(x) = \mathbb{P}[W_i = 1 | X_i = x]$ [Rosenbaum and Rubin, 1983].

We focus primarily on the expected treatment effect for the treated sample,

$$\tau = \frac{1}{n_{\text{t}}} \sum_{\{i : W_i = 1\}} \mathbb{E}\left[Y_i(0) - Y_i(1) \,\middle|\, X_i\right]. \tag{2}$$

We note that the average treatment effect for the controls can be handled similarly, and the overall average effect is a weighted average of the two.

Throughout the paper we assume unconfoundedness, i.e., that conditional on the pretreatment variables, treatment assignment is as good as random [Rosenbaum and Rubin, 1983]:

Assumption 1 (Unconfoundedness)**.**

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \;\big|\; X_i. \tag{3}$$

Given unconfoundedness, there is a well established semiparametric efficiency theory for estimating average treatment effects. Assuming overlap and some regularity conditions, the semiparametric efficiency bound for regular estimators μ_c with a fixed number of covariates implies that no estimator can improve on an asymptotic variance of $\mathbb{V} = \mathbb{E}\big[\mathbb{P}\left[W=1\right]^{-2}\, e^2(X_i) \cdot \sigma_c^2(X_i) \,/\, (1 - e(X_i)\big]$ [Bickel et al., 1998, Hahn, 1998, Hirano et al., 2003, Robins and Rotnitzky, 1995, Robins et al., 1995]. This asymptotic variance is induced by the influence function

$$\psi(y, x, w) = -(1-w)\, \frac{e(x)}{\mathbb{P}\left[W=1\right](1-e(x))}\, \left(y - \mu_c(x)\right), \tag{4}$$

so that $\mathbb{E}[\psi(Y_i^{\text{obs}}, X_i, W_i)^2] = \mathbb{V}$. Our goal is to construct an estimator that attains this efficiency bound in high dimensions, without requiring special assumptions—such as sparsity—on the propensity model.[1]

For our analysis, we assume a linear model for the potential outcomes in both groups.

Assumption 2 (Linearity)**.**

$$\mu_c(x) = \mathbb{E}\left[Y_i(0) \,\middle|\, X = x\right] = x \cdot \beta_c, \qquad \mu_t(x) = \mathbb{E}\left[Y_i(1) \,\middle|\, X = x\right] = x \cdot \beta_t, \tag{5}$$

for $w \in \{0,\, 1\}$ and $x \in \mathbb{R}^p$.

In fact, we will only use the linear model for the control outcome because we focus on the average effect for the treated units, but if we were interested in the overall average effect we would need linearity in both groups. The linearity assumption is strong, but it may be plausible, especially if the researcher includes transformations of the basic features in the design. Given this linearity, we can write the estimand as

$$\tau = \mu_t - \mu_c, \quad \text{where} \quad \mu_t = \overline{X}_t \cdot \beta_t, \quad \mu_c = \overline{X}_t \cdot \beta_c, \quad \text{and} \quad \overline{X}_t = \frac{1}{n_t} \sum_{\{i : W_i = 1\}} X_i. \tag{6}$$

Here, estimating the first term is easy: $\hat{\mu}_t = \overline{Y}_t = \sum_{\{i:W_i=1\}} Y_i^{\text{obs}} / n_t$ is unbiased for μ_t, and in fact we do not require linearity for the treated outcomes. In contrast, estimating μ_c is a major challenge, especially in settings where p is large, and it is the main focus of the paper.

2.2 Baselines and Background

We begin by reviewing two classical approaches to estimating μ_c, and thus also τ, in the above linear model. The first is a weighting-based approach, which seeks to re-weight the control sample to make it look more like the treatment sample; the second is a regression-based approach, which seeks to adjust for differences in features between treated and control units by fitting an accurate model to the outcomes. Neither approach alone performs well in a high-dimensional setting with a generic propensity score. However, in Section 2.3, we show that these two approaches can be fruitfully combined to obtain better estimators for τ.

[1] We note that the above efficiency bounds are only known to be optimal in a low-dimensional semiparametric setting following, e.g., Newey [1994]. Our goal of reaching the asymptotic rate $\mathbb{V}$ is thus only motivated by an aesthetic or heuristic connection to earlier work, and not by a matching efficiency theory for high-dimensional, linear average treatment effect estimation. Developing such a formal theory would be of considerable interest.

2.2.1 Weighted Estimation

A first approach is to re-weight the control dataset using weights γ_i to make the weighted covariate distribution mimic the covariate distribution in the treatment population. Given the weights we estimate $\hat{\mu}_{\mathrm{c}}$ as

$$\hat{\mu}_{\mathrm{c}} = \sum_{\{i:W_i=0\}} \gamma_i \, Y_i^{\mathrm{obs}}. \tag{7}$$

The standard way of selecting weights γ_i uses the propensity score:

$$\gamma_i = \frac{e(X_i)}{1-e(X_i)} \Bigg/ \sum_{\{i:W_j=0\}} \frac{e(X_j)}{1-e(X_j)}, \tag{8}$$

where $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ is the propensity score [Rosenbaum and Rubin, 1983]. To implement these methods researchers typically substitute an estimate of the propensity score into the expression for the weights (8). Such inverse-propensity weights with a flexibly estimated propensity score have desirable asymptotic properties [Hirano et al., 2003] in settings where the asymptotics is based on a fixed number of covariates.

The finite-sample performance of methods based on (8) can be poor, however, both in settings with limited overlap in covariate distributions and in settings with many covariates. In the latter case recently proposed methods include (regularized) logistic regression, boosting, support vector machines, neural networks, and classification trees [McCaffrey et al., 2004, Westreich et al., 2010]. But since estimating the treatment effect then involves dividing by $1-\hat{e}(X_i)$, small inaccuracies in $\hat{e}(X_i)$ can have large effects; this problem is often quite severe in high dimensions. To our knowledge, methods based on inverse propensity weighting are not known to have good asymptotic properties in high-dimensional settings.

Recently, there have been proposals to select weights γ_i by focusing on balance directly, rather than on fit of the propensity score [Deville and Särndal, 1992, Chan et al., 2015, Graham et al., 2012, 2016, Hainmueller, 2012, Hellerstein and Imbens, 1999, Imai and Ratkovic, 2014, Zhao, 2016, Zubizarreta, 2015]. This is a subtle but important improvement. The motivation behind this approach is that, in a linear model, the bias for estimators based on (7) depends solely on $\overline{X}_{\mathrm{t}} - \sum_{\{i:W_i=0\}} \gamma_i \, X_i$. Therefore getting the propensity model exactly right is less important than accurately matching the moments of $\overline{X}_{\mathrm{t}}$.

In high dimensions, however, exact balancing weights do not in general exist. When $p \gg n_{\mathrm{c}}$, there will in general be no weights γ_i for which $\overline{X}_{\mathrm{t}} - \sum_{\{i:W_i=0\}} \gamma_i \, X_i = 0$, and even in settings with $p < n_{\mathrm{c}}$ but p is large such estimators would not have good properties. Zubizarreta [2015] extends the balancing weights approach to allow for weights that achieve approximate balance instead of exact balance, and considers the tradeoff between precision of the resulting estimators and the bias from lack of balance. We find, however, that only achieving approximate balancing leads to estimators for τ that still have substantial bias in many settings.

2.2.2 Regression Adjustments

A second approach is to compute an estimator $\hat{\beta}_{\mathrm{c}}$ for β_{c} using the n_{c} control observations, and then estimate μ_{c} as

$$\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}}. \tag{9}$$

In a low-dimensional regime with $p \ll n_{\mathrm{c}}$, the ordinary least squares estimator for β_{c} is a natural choice, and yields an accurate and unbiased estimate of μ_{c}. In high dimensions, however, the problem is more delicate: accurate unbiased estimation of β_{c} is in general impossible, and methods such as the lasso, ridge regression, or the elastic net may perform poorly when plugged into (9), in particular when $\overline{X}_{\mathrm{t}}$ is far away from $\overline{X}_{\mathrm{c}}$, the average covariate values for the controls.

As stressed by Belloni et al. [2014, 2016], the problem with plain lasso regression adjustments is that features with a substantial difference in average values between the two treatment arms can generate large biases even if the coefficients on these features in the outcome regression are small. Thus, a regularized regression that has been tuned to optimize goodness of fit on the outcome model is not appropriate whenever bias in the treatment effect estimate due to failing to control for potential confounders is of concern. To address this problem, Belloni et al. [2014] propose running least squares regression on the union of two sets of selected variables, one selected by a lasso regressing the outcome on the covariates, and the other selected by a lasso logistic regression for the treatment assignment. We note that estimating μ_c by a regression adjustment $\hat{\mu}_c = \overline{X}_t \cdot \hat{\beta}_c$, with $\hat{\beta}_c$ estimated by ordinary least squares on a selected variables, is implicitly equivalent to running (7) with weights γ chosen to balance the selected features.

The Belloni et al. [2014] approach works well in settings where both the outcome regression and the treatment regression are at least approximately sparse. However, when the propensity is not sparse, we find that the performance of such double-selection methods is often poor. Moreover, Cai and Guo [2015] formally establish that minimax rates for inference about $a \cdot \beta_c$ for generic dense vectors a are quite poor, even when β_c is very sparse. Thus, methods relying on regression adjustments alone may not be able deliver accurate inference about τ for generic propensity models.

2.3 Approximate Residual Balancing

Here we propose a new method combining weighting and regression adjustments to overcome the limitations of each method. In the first step of our method, we use a regularized linear model, e.g., the lasso or the elastic net, to obtain a first pilot estimate of the treatment effect. In the second step, we do "approximate balancing" of the regression residuals to estimate treatment effects: that is, we weight the residuals using weights that achieve approximate balance of the covariate distribution between treatment and control groups. This step compensates for the potential bias of the pilot estimator that arises due to confounders that may be weakly correlated with the outcome but are important due to their correlation with the treatment assignment. We find that the regression adjustment is effective at capturing strong effects; the weighting on the other hand is effective at capturing small effects. The combination leads to an effective and simple-to-implement estimator for average treatment effects in a wide variety of settings with many features.

We focus on a meta-algorithm that first computes an estimate $\hat{\beta}_c$ of β_c, using the full sample of control units. This estimator may take a variety of forms, but typically it will involve some form of regularization to deal with the number of features. Second we compute weights γ_i that balance, at least approximately, covariate distributions. Then our estimator for μ_c has the general form

$$\hat{\mu}_c = \overline{X}_t \cdot \hat{\beta}_c + \sum_{\{i : W_i = 0\}} \gamma_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_c \right). \tag{10}$$

In other words, we fit a model parametrized by β_c to capture some of the big signals, and then use a non-parametric re-balancing of the control data on the features to extract left-over signal from the residuals $Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_c$. Ideally, we would hope for the first term to take care of any strong effects, while the re-balancing of the residuals can efficiently take care of the small spread-out effects. Our theory and experiments will verify that this is in fact the case.

A major advantage of the functional form in (10) is that it yields a simple and powerful theoretical guarantee, as stated below. Recall that $\mathbf{X}_c$ is the feature matrix for the control units. Consider the difference between $\hat{\mu}_c$ and μ_c for our proposed approach:

$$\hat{\mu}_c - \mu_c = \left(\overline{X}_t - \mathbf{X}_c^\top \gamma\right) \cdot \left(\hat{\beta}_c - \beta_c\right) + \gamma \cdot \varepsilon,$$

where ε is the intrinsic noise $\varepsilon_i = Y_i(0) - X_i \cdot \beta_c$. With only the regression adjustment and no weighting, the difference would be

$$\hat{\mu}_{c,\text{reg}} - \mu_c = \left(\overline{X}_t - \overline{X}_c\right) \cdot \left(\hat{\beta}_c - \beta_c\right) + \mathbf{1} \cdot \varepsilon / n,$$

and with only the weighting the difference would be

$$\hat{\mu}_{c,\text{weight}} - \mu_c = \left(\overline{X}_t - \mathbf{X}_c^\top \gamma\right) \cdot \beta_c + \gamma \cdot \varepsilon.$$

Without any adjustment, just using the average outcome for the controls as an estimator for μ_{c}, the difference between the estimator for μ_{c} and its actual value would be

$$\hat{\mu}_{\mathrm{c,no-adj}} - \mu_{\mathrm{c}} = \left(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}\right) \cdot \beta_{\mathrm{c}} + \mathbf{1} \cdot \varepsilon / n.$$

The regression reduces the first term, capturing part of the bias, from $(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}) \cdot \beta_{\mathrm{c}}$ to $(\overline{X}_{\mathrm{t}} - \overline{X}_{\mathrm{c}}) \cdot (\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}})$, which will be substantial reduction if $(\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}})$ is small relative to β_{c}. The weighting further reduces this to $(\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^{\top}\gamma) \cdot (\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}})$, which may be helpful if there is a substantial difference between $\overline{X}_{\mathrm{t}}$ and $\overline{X}_{\mathrm{c}}$. This argument shows the complimentary nature of the regression adjustment and the weighting.

The following result formalizes the notion that the combination of regression and weighting can improve the properties of the estimators substantially. All proofs are given in Section 5.

Proposition 1. *The estimator defined in* (10) *satisfies*

$$\left|\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}\right| \leq \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^{\top}\gamma\right\|_{\infty} \left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_{1} + \left|\sum_{\{i:W_i=0\}} \gamma_i \, \varepsilon_i\right|. \tag{11}$$

The result (11) decomposes the error of $\hat{\mu}_{\mathrm{c}}$ into two parts. The first is the main term, depending on the design $\mathbf{X}_{\mathrm{c}}$, and affected by the dimension of the covariates; the second term is a variance term that does not depend on the dimension of the covariates. The upshot is that the main term, which encodes the high-dimensional nature of the problem, involves a product of two factors that can both other be made reasonably small; more specifically, we will focus on regimes where the first term scales as $\mathcal{O}(\sqrt{\log(p)/n})$, while the second term scales as $\mathcal{O}(k\sqrt{\log(p)/n})$ where k is the sparsity of the outcome model. Thus, (11) will often enable us us to control high-dimensional bias effects better than only weighting or only estimation of β_{c}.

In order to exploit Proposition 1, we need to make concrete choices for the weights γ and the parameter estimates $\hat{\beta}_{\mathrm{c}}$. We define *approximately balancing weights* as

$$\gamma = \mathrm{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta) \left\|\tilde{\gamma}\right\|_2^2 + \zeta \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^{\top}\tilde{\gamma}\right\|_{\infty}^2 \text{ subject to } \sum \tilde{\gamma}_i = 1, \ \tilde{\gamma}_i \geq 0 \right\}, \tag{12}$$

for some $\zeta \in (0, 1)$. When we wish to stress the dependence of these weights on ζ, we will write $\gamma(\zeta)$. Most of the time we omit this dependence to simplify the exposition. These weights, which are closely related to a recent proposal by Zubizarreta [2015], are designed to make both terms of (11) small. In contrast, the inverse propensity score weights do not take the variance component into account at all. We refer to these weights as approximately balancing since they seek to make the mean of the re-weighted control sample, namely $\mathbf{X}_{\mathrm{c}}^{\top}\gamma$, match the treated sample mean $\overline{X}_{\mathrm{t}}$ as closely as possible. Below, we show that we can find a ζ that achieves our objective of bounding both terms of (11); in our simulations we use $\zeta = 1/2$, which balances the square of the bias term and the variance.

Meanwhile, for estimating $\hat{\beta}_{\mathrm{c}}$ there are a number of possibilities. One is to use the lasso [Chen et al., 1998, Tibshirani, 1996] as there are several well-known results that let us control its 1-norm error [Hastie et al., 2015]. In our simulations we use the elastic net [Zou and Hastie, 2005] to estimate β_{c}. Note that we do not need to select a sparse model, we just need to regularize the estimator. Using a combination of L_1 and L_2 regularization may therefore work well in practice. So, specifically, we calculate $\hat{\beta}_{\mathrm{c}}$ as

$$\hat{\beta}_{\mathrm{c}} = \arg\min_{\beta_{\mathrm{c}}} \left\{ \sum_{\{i:W_i=0\}} (Y_i^{\mathrm{obs}} - X_i^{\top}\beta_{\mathrm{c}})^2 + \lambda \left((1-\alpha) \left\|\beta_{\mathrm{c}}\right\|_2^2 + \alpha \left\|\beta_{\mathrm{c}}\right\|_1 \right) . \right\}. \tag{13}$$

For some of the theoretical analysis we focus on the lasso case with $\alpha = 1$. Our complete algorithm is described in Procedure 1.

One question is why the balancing weights perform better than the propensity score weights, a finding that is also reported in Chan et al. [2015], Hainmueller [2012], and Zubizarreta [2015]. To gain intuition for this issue in a simple parametric context, suppose the propensity score has the following logistic form,

$$e(x) = \frac{\exp(x \cdot \theta)}{1 + \exp(x \cdot \theta)}.$$

Procedure 1. Approximately Residual Balancing with Elastic Net

The following algorithm estimates the average treatment effect on the treated by approximately balanced residual weighting. Here, $\zeta \in (0,\, 1)$, $\alpha \in (0,\, 1)$ and $\lambda > 0$ are tuning parameters. This procedure is implemented in our `R` package `balanceHD`; we default to $\zeta = 0.5$ and $\alpha = 0.9$, and select λ by cross-validation using the `lambda.1se` rule from the `glmnet` package [Friedman et al., 2010].

1. Compute positive approximately balancing weights γ as

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta) \left\|\tilde{\gamma}\right\|_2^2 + \zeta \left\| \overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \tilde{\gamma} \right\|_\infty^2 \text{ s.t.} \sum_{\{i:W_i=0\}} \tilde{\gamma}_i = 1 \text{ and } \tilde{\gamma}_i \geq 0 \right\}. \tag{14}$$

2. Fit β_{c} in the linear model using an elastic net,

$$\hat{\beta}_{\text{c}} = \text{argmin}_{\beta} \left\{ \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\text{obs}} - X_i \cdot \beta \right)^2 + \lambda \left((1-\alpha) \left\|\beta\right\|_2^2 + \alpha \left\|\beta\right\|_1 \right) \right\}. \tag{15}$$

3. Estimate the average treatment effect τ as

$$\hat{\tau} = \overline{Y}_{\text{t}} - \left(\overline{X}_{\text{t}} \cdot \hat{\beta}_{\text{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_{\text{c}} \right) \right). \tag{16}$$

In that case the inverse propensity score weights would be proportional to $\gamma_i \propto \exp(x \cdot \theta)$. The efficient estimator for θ is the maximum likelihood estimator,

$$\hat{\theta}_{\text{ml}} = \arg\max_{\theta} \sum_{i=1}^{n} \left\{ W_i \, X_i \cdot \theta - \ln(1 + \exp(X_i \cdot \theta)) \right\}.$$

An alternative, less efficient, estimator for θ is the method of moments estimator $\hat{\theta}_{\text{mm}}$ that balances the covariates exactly:

$$\overline{X}_{\text{t}} = \sum_{\{i:W_i=0\}} X_i \frac{\exp(X_i \cdot \theta)}{\sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)},$$

with implied weights $\gamma_i \propto \exp(X_i \cdot \hat{\theta}_{\text{mm}})$. The weights are very similar to those based on the estimated propensity score, with the only difference that the parameter estimates $\hat{\theta}$ differ. The estimator $\hat{\theta}_{\text{mm}}$ leads to weights that achieve exact balance on the covariates, in contrast to either the true value θ, or the maximum likelihood estimator $\hat{\theta}_{\text{ml}}$. The point of this discussion is that the goal of balancing (leading to $\hat{\theta}_{\text{mm}}$) is different from the goal of estimating the propensity score (for which $\hat{\theta}_{\text{ml}}$ is optimal).

2.4 Related Work

The idea of combining weighted and regression-based approaches to treatment effect estimation has a long history in the causal inference literature. In a low-dimensional setting where both methods are already consistent on their own, they can be combined to get "doubly robust" estimates of τ [Robins and Rotnitzky, 1995, Robins et al., 1995, Van der Laan and Robins, 2003]. These methods, which first calculate the weights based on propensity score estimates and then estimate β_{c} by weighted least squares, are guaranteed to be consistent if either the outcome model or the propensity model is well specified, although they do not always have good properties when the estimated propensity score is close to zero or one [Hirano et al., 2003, Kang and Schafer, 2007]. Farrell [2015] studies the behavior of doubly robust estimators in high dimensions, and establishes conditions under which they can reach efficiency when both the propensity function and the outcome model are consistently estimable.

Our method has strong parallels with the double-robustness approach, although we modify it in two aspects: We estimate the regression function $\mu_{\mathrm{c}}(x) = x \cdot \beta_{\mathrm{c}}$ without weights, and we construct the weights by balancing arguments, rather than by estimating the propensity score. These differences enable us to provide efficient estimates of τ in high-dimensional linear models under weaker conditions on the propensity score; however, this gain in efficiency comes at the cost of losing doubly robust consistency guarantees.

Our method is also related to the semiparametric efficiency bound literature. In low dimensions, i.e., with a fixed number of features, efficient estimators of μ_c are asymptotically equivalent to

$$\tilde{\mu}_c = \frac{1}{N_t} \sum_{i=1}^{N} \left\{ e(X_i)(\mu_{\mathrm{c}}(X_i) - \mu_{\mathrm{c}}) - (1 - W_i) \frac{e(X_i)}{1 - e(X_i)} (Y_i - \mu_{\mathrm{c}}(X_i)) \right\}; \tag{17}$$

see, e.g., Hahn [1998]. Bickel et al. [1998] suggest estimating the components of this asymptotic linear representation (i.e., $e(x)$ and $\mu_{\mathrm{c}}(x)$), and substituting them into the function as a general approach to obtaining an efficient estimator for the parameter of interest. A key component of our modification, motivated by the difficulty of accurately estimating inverse propensity weights in some settings, is that we use balancing weights instead of the propensity score weights $e(X_i)/(1-e(X_i))$. A second modification is that we estimate the regression function using regularization methods.

Intriguingly, in low dimensions, doubly robust methods are not necessary for achieving semiparametric efficiency; this rate can be achieved by either non-parametric inverse-propensity weighting or non-parametric regression adjustments on their own [Chen et al., 2008, Hirano et al., 2003]. At best, the use of non-parametric doubly robust methods can only improve on the second-order properties of the the average treatment effect estimate [Rothe and Firpo, 2013]. Conversely, in high-dimensions, we have found that both weighting and regression adjustments are required for $\sqrt{n}$-consistency; this finding mirrors the results of Belloni et al. [2016] and Farrell [2015].

Our approximately balancing weights (12) are inspired by the work of Chan et al. [2015], Graham et al. [2012, 2016], Hainmueller [2012], Hirano et al. [2001], Imai and Ratkovic [2014], and Zubizarreta [2015]. Most closely related, Zubizarreta [2015] proposes estimating τ using the re-weighting formula (7) with weights

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty \leq t \right\}, \tag{18}$$

where the tuning parameter is t; he calls these weights *stable balancing weights*. The main conceptual difference between our setting and that of Zubizarreta [2015] is that he considers problem settings where $p < n_{\mathrm{c}}$, and then considers t to be a practically small tuning parameter, e.g., $t = 0.1\sigma$ or $t = 0.001\sigma$. However, in high dimensions, the optimization problem (18) will not in general be feasible for small values of t; and in fact the bias term $\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty$ becomes the dominant source of error in estimating τ. We call our our weights γ "approximately" balancing in order to remind the reader of this fact. Similar estimator have been considered in Graham et al. [2012, 2016] and Hainmueller [2012] in a setting where exact balancing is possible, with slightly different objection function; for example, Hainmueller [2012] uses $-\sum_i \ln(\gamma_i)$ instead of $\sum_i \gamma_i^2$, leading to

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ -\sum_{\{i : W_i = 0\}} \log\left(\tilde{\gamma}_i\right) \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty = 0 \right\}. \tag{19}$$

This estimator has attractive conceptual connections to logistic regression and maximum entropy estimation. In particular, in a low dimensional setting where $W|X$ admits a well-specified logistic model, the results of Owen [2007] imply that the methods of Graham et al. [2012, 2016] and Hainmueller [2012] are doubly robust; see also Newey and Smith [2004], Imbens et al. [1998], and Hirano et al. [2001]. In terms of our immediate concerns, however, the variance of $\hat{\tau}$ depends on γ through $\|\gamma\|_2^2$ and not $-\sum \log(\gamma_i)$, so our approximately balancing weights should be more efficient than those defined in (19).

Finally, we note that while both our method and that of Belloni et al. [2014] rely on the lasso, we use the lasso in conceptually different ways. Belloni et al. [2014] use the lasso purely as a variable selection method to find the features that matter for estimating τ; they then run a linear regression without any shrinkage on the union of the two sets of selected variables. In contrast, our residual re-weighting scheme enables us to simply use the lasso as an accurate predictive tool, and our analysis does not rely on recovering a decent approximation to the full set of important pretreatment variables.

3 Theoretical Analysis

In this section, we provide a brief discussion of the theoretical properties of approximately balanced estimation. We find that our method admits strong asymptotic performance guarantees that can be verified using mostly elementary proof techniques paired with some well understood results about the lasso as discussed by, e.g., Hastie et al. [2015]. We begin by studying regimes under which we can achieve approximate balance using our weighting scheme, i.e., our procedure succeeds in making both $\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty$ and $\|\gamma\|_2^2$ small; we then use these results to provide study the behavior of $\hat{\tau}$ in Sections 3.2 and 3.3.

In this section, we consider a slightly more careful form of approximately balanced weighting, and also require that the largest weight γ_i must be smaller than $n_{\mathrm{c}}^{-2/3}$:

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta)\,\|\tilde{\gamma}\|_2^2 + \zeta \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty^2 \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \max_i \tilde{\gamma}_i \leq n_{\mathrm{c}}^{-2/3} \right\}. \quad (20)$$

This additional constraint will enable us to ensure asymptotic normality of our estimator. We do not impose this additional constraint in our software, since it is often already satisfied by the solution to (12).

3.1 Achieving Approximate Balance

In high-dimensions, it is not in general possible to find a re-weighting vector γ for which $\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty$ is small. For example suppose that $p = 2n_{\mathrm{c}}$, that $\mathbf{X}_{\mathrm{c}} = (I_{n_{\mathrm{c}} \times n_{\mathrm{c}}}\ I_{n_{\mathrm{c}} \times n_{\mathrm{c}}})$, and that $\overline{X}_{\mathrm{t}}$ consists of n times "1" followed by n times "-1"; then $\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty \geq 1$ for any $\gamma \in \mathbb{R}^{n_{\mathrm{c}}}$, and the approximation error does not improve as n_{c} and p both get large. We will find, however, that there exist fairly natural conditions under which we can achieve approximate balance with high probability even when $p \gg n_{\mathrm{c}}$.

The first such condition is to assume overlap, as is common in the literature on causal inference from observational studies [Crump et al., 2009, Imbens and Rubin, 2015]. Informally, overlap requires that each unit have a positive probability of receiving each of the treatment and control conditions, and thus that the treatment and control populations cannot be too dissimilar. Without overlap, estimation of average treatment effects relies fundamentally on extrapolation beyond the support of the features, and thus makes estimation inherently sensitive to functional form assumptions.

Assumption 3 (Overlap)**.** There is a constant $0 < \eta$ such that $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathbb{R}^p$.

For estimation of the average effect for the treated we in fact only need the propensity score to be bounded from above by $1 - \eta$, but for estimation of the overall average effect we would require both the lower and upper bound on the propensity score.

In order to guarantee approximate balance, we also need to bound the tail behavior of the features X_i. Assumption 4 ensures that the features are sub-Gaussian.

Assumption 4 (Sub-Gaussian Features)**.** There is a constant $\nu > 0$ such that the distribution of X_j conditional on $W = w$ is sub-Gaussian with parameter ν^2 after re-centering. In other words, $\mathbb{E}\left[\exp[t(X_{ij} - \mathbb{E}\left[X_{ij} \,\middle|\, W_i = w\right])] \,\middle|\, W_i = w\right] \leq \exp[\nu^2 t^2/2]$ for any $t > 0$, $j \in \{1, ..., p\}$ and $w \in \{0, 1\}$.

Lemma 2. *Suppose that we have n independent and identically distributed examples (X_i, Y_i, W_i) drawn from a joint distribution satisfying Assumptions 3 and 4. Then, for any $\delta > 0$, with probability at least $1 - \delta$, there exists a $\zeta \in (0, 1)$ for which the weights γ defined in (20) satisfy*

$$\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty \leq \nu \sqrt{2 \log\left(\frac{10\, p}{\delta}\right) \left(\frac{1}{n_{\mathrm{t}}} + \frac{(1-\eta)^2}{n_{\mathrm{c}}\, \eta^2}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right), \quad \textit{and} \tag{21}$$

$$n_{\mathrm{t}}\, \left\|\gamma\right\|_2^2 \leq \frac{1}{\rho^2}\, \mathbb{E}\left[\left(\frac{e(X_i)}{1 - e(X_i)}\right)^2 \middle| W_i = 0\right] + \frac{(1-\eta)^2\,(2 - 2\eta + \rho\eta)}{\rho^3\, \eta^3} \sqrt{\frac{1}{2n_{\mathrm{c}}} \log\left(\frac{10\, p}{\delta}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right), \tag{22}$$

where $\rho = \mathbb{P}\left[W_i = 1\right] / \mathbb{P}\left[W_i = 0\right]$ is the odds ratio.

Lemma 2 shows that the “Overlap” and “Sub-Gaussian Features” assumptions allows us to provide bounds on the imbalance for approximately balancing weights γ and the variance of the weights. These bounds are essentially tight for the true propensity score weights, $\gamma_i^* \propto e(X_i)/(1 - e(X_i))$; thus, using γ^* instead of γ for approximate residual balancing would not improve on the guarantees we get from Proposition 1.

Note that the above result only guarantees the existence of a $\zeta \in (0, 1)$ that achieves good balance and satisfies a bound on the variance, but does not tell us what the value of that ζ is. However, because the right-hand side terms of (21) and (22) depend on observables, we can simply try many different values of ζ to find one that yields weights γ satisfying both conditions.

3.2 Gaussian Asymptotics

The main goal of our paper is to use the approximately balancing weights discussed above to construct efficient estimators for average treatment effects in high-dimensional linear models, where $\mu_{\mathrm{c}}(x) = x \cdot \beta_{\mathrm{c}}$ for some β_{c} in $\mathbb{R}^p$. Building on Proposition 1, we focus on the problem of finding estimators $\hat{\beta}_{\mathrm{c}}$ for which $\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\|_1$ is small. This is a well-studied problem in the mathematical statistics literature, and we will be able to leverage recent theoretical results. In order to accurately estimate β_{c} under L_1 norm, we make a sparsity assumption on β_{c}.

Assumption 5 (Sparsity)**.** We have a sequence of problems indexed by n, p, and k such that the parameter vector β_{c} is k-sparse, i.e., $\left\|\beta_{\mathrm{c}}\right\|_0 \leq k$, and that $k\, \log(p)/\sqrt{n} \to 0$.

The above sparsity requirement is quite strong. However, many analyses that seek to establish asymptotic normality in high dimensions rely on such an assumption. For example, Javanmard and Montanari [2014], Van de Geer et al. [2014], and Zhang and Zhang [2014] all make this assumption when seeking to provide confidence intervals for individual components of β_{c}; furthermore, Cai and Guo [2015] show that efficient inference about the entries of β_{c} is in general impossible unless $k \ll \sqrt{n}\, /\, \log(p)$, or the sparsity level k is known a-priori. Belloni et al. [2014] use a similar assumption where they allow for additional non-zero components, but they assume that beyond the largest k components with k satisfying the same sparsity condition, the remaining non-zero elements of β_{c} are sufficiently small that they can be ignored, in what they refer to as approximate sparsity. The only exception to this assumption we are aware of is in recent work by Javanmard and Montanari [2015], who show that inference of β_{c} is possible even when $k \ll n\, /\, \log(p)$ in a setting where X is a random Gaussian matrix with either a known or extremely sparse population precision matrix.

We also require that the design $\mathbf{X}_{\mathrm{c}}$ satisfy a form of the restricted eigenvalue condition [Bickel et al., 2009]. This type of assumption is standard in the literature on high-dimensional estimation [e.g., Belloni et al., 2014, Meinshausen and Yu, 2009, Negahban et al., 2012, Van De Geer and Bühlmann, 2009]. Below, we follow the framework of Negahban et al. [2012].

Assumption 6 (Restricted Eigenvalue)**.** For $1 \leq k \leq p$, define the set $\mathcal{C}_k$ as

$$\mathcal{C}_k = \left\{\beta \in \mathbb{R}^p : \left\|\beta\right\|_1 \leq 4 \sum_{j=1}^{k} \left|\beta_{i_j}\right| \quad \text{for some} \quad 1 \leq i_1 < ... < i_j \leq p\right\}. \tag{23}$$

Then, $\mathbf{X}_{\mathrm{c}}$ satisfies the $\{k,\, \omega\}$-restricted eigenvalue condition for $\omega > 0$ if

$$\frac{1}{n_{\mathrm{c}}} \left\|\mathbf{X}_{\mathrm{c}}\beta\right\|_2^2 \geq \omega \left\|\beta\right\|_2^2 \ \text{ for all } \ \beta \in \mathcal{C}_k. \tag{24}$$

Our analysis builds on well-known bounds on the estimation error of the lasso [Bickel et al., 2009, Candès and Tao, 2007, Meinshausen and Yu, 2009]; the following result depends directly on Corollary 2 of Negahban et al. [2012].

Theorem 3. *Suppose that we have n independent and identically distributed training examples satisfying Assumptions 1–6, and write $\rho = \mathbb{P}\left[W = 1\right] / \mathbb{P}\left[W = 0\right]$. Suppose, moreover, that we have homoskedastic noise: $\mathrm{Var}[\varepsilon_i(w) \,\big|\, X_i] = \sigma^2$ for all $i = 1, \ldots, n$, and also that the response noise $\varepsilon_i(w) := Y_i(w) - \mathbb{E}[Y_i(w) \,\big|\, X_i]$ is uniformly sub-Gaussian with parameter $\upsilon^2 > 0$. Finally, suppose that we use Procedure 1 for estimation, with the lasso penalty parameter set to $\lambda_n = 5\nu\upsilon\sqrt{\log\left(p\right)/n_{\mathrm{c}}}$ instead of selecting λ_n by cross-validation. Then, there exists a sequence $\zeta_n \in (0,\, 1)$ for which the weights $\gamma = \gamma(\zeta_n)$ defined in (20) satisfy (21) and (22) with $\delta = 1/p$, and yield*

$$\frac{\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}}{\left\|\gamma\right\|_2} \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right) \quad \textit{and} \quad \frac{\hat{\tau} - \tau}{\sqrt{n_{\mathrm{t}}^{-1} + \left\|\gamma\right\|_2^2}} \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right), \tag{25}$$

where τ is the expected treatment effect on the treated (2). Moreover,

$$\limsup_{n \to \infty} \ n_{\mathrm{c}} \left\|\gamma\right\|_2^2 \ \leq \ \rho^{-2}\, \mathbb{E}\left[\left(\frac{e\left(X_i\right)}{1 - e\left(X_i\right)}\right)^2 \middle| \, W_i = 0\right]. \tag{26}$$

This result shows that we can achieve asymptotic inference about τ with a $1/\sqrt{n}$ rate of convergence, irrespective of the dimension of the features, subject to the effectively the same sparsity assumptions on the Y-model as used by the rest of the high-dimensional inference literature, including Belloni et al. [2014] and Farrell [2015]. However, unlike this literature, we make no assumptions on the propensity model beyond overlap, and do not require it to be estimated consistently. We also note that, recently, Bloniarz et al. [2015] proved an analogue to Theorem 3 for lasso regression adjustments to randomized controlled trials. Our result shows that, in observational studies, lasso regression adjustments are still helpful—but now need to be paired with an approximate residual balancing step.

The rate over convergence implied by (26) is the same as what we would get if we actually knew the true propensities and could use them for weighting; we achieve this rate although we have no guarantees that our balancing weights γ_i approximate the underlying propensity weights $e(X_i)/(1 - e(X_i))$. We also note that the rate in Theorem 3 matches the semiparametric efficiency bound for estimating average treatment effects given homoskedasticity [Hahn, 1998, Hirano et al., 2003], i.e., the upper bound (26) is equivalent to the efficiency bound (4) for inference of average treatment effects in a low-dimensional semiparametric setting.

3.3 Inference under Heteroskedasticity

The previous section established that, in the homoskedastic setting, approximate residual balancing has a Gaussian limit distribution. This result naturally suggests that our method should also allow for asymptotic inference about τ. Here, we verify that this is in fact the case; and, moreover, we show that our proposed confidence intervals are heteroskedaticity robust.

Corollary 4. *Under the conditions of Theorem 3, suppose instead that we have heteroskedastic noise*

$$\upsilon_{\min}^2 \leq \mathrm{Var}\left[\varepsilon_i(W_i) \,\big|\, X_i,\, W_i\right] \leq \upsilon^2 \quad \textit{for all} \ \ i = 1,\, \ldots,\, n. \tag{27}$$

Then, the following holds:

$$\left(\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}\right) \Big/ \sqrt{\widehat{V}_c} \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad \widehat{V}_c = \sum_{\{i : W_i = 0\}} \gamma_i^2 \left(Y_i - X_i \cdot \hat{\beta}_{\mathrm{c}}\right)^2. \tag{28}$$

In order to provide inference about τ, we also need error bounds for $\hat{\mu}_{\mathrm{t}}$. Under sparsity assumptions comparable to those made for β_{c} in Theorem 3, we can verify that

$$(\hat{\mu}_{\mathrm{t}} - \mu_{\mathrm{t}}) \Big/ \sqrt{\widehat{V}_t} \Rightarrow (0,\,1)\,, \quad \widehat{V}_t = \frac{1}{n_{\mathrm{t}}^2} \sum_{\{i:W_i=1\}} \left(Y_i - X_i \hat{\beta}_{\mathrm{t}}\right)^2, \tag{29}$$

where $\hat{\beta}_{\mathrm{t}}$ is obtained using the lasso with $\lambda_n = 5\nu\upsilon\sqrt{\log\left(p\right)/n_{\mathrm{c}}}$. Moreover, $\hat{\mu}_{\mathrm{c}}$ and $\hat{\mu}_{\mathrm{t}}$ are independent conditionally on X and W, thus implying that $(\hat{\tau} - \tau)\,/\,(\widehat{V}_c + \widehat{V}_t)^{1/2} \Rightarrow \mathcal{N}\left(0,\,1\right)$. This last expression is what we use for building confidence intervals for τ.

3.4 Approximate Residual Balancing as Debiased Estimation

Much of the existing literature on high-dimensional inference relies a 2-step debiasing algorithms. For example, in order to provide confidence intervals for individual parameters β_i in high-dimensional regression problems, Javanmard and Montanari [2014] center their confidence intervals at

$$\hat{\beta}^{(\text{debiased})} = \hat{\beta} + MX^\top\left(Y - X\hat{\beta}\right), \tag{30}$$

where $\hat{\beta}$ is the plain lasso estimate, and M is the solution to a quadratic program chosen such as to make the second term acts as an approximate Newton step; see also Van de Geer et al. [2014]. In this context, we can also understand the residual balancing step of our algorithm as an analogous debiasing step for the plain lasso-based estimator $\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}}$.

Now, despite the cosmetic similarity, we emphasize that the conceptual reasons for why both methods work are quite different. For example, trying to estimate μ_{c} as $\overline{X}_{\mathrm{t}} \cdot \hat{\beta}^{(\text{debiased})}$ would not work well at all. Moreover, the minimax results of Cai and Guo [2015] imply that no method that simply seeks to debias estimates of β can achieve efficient inference about τ. Here, our overlap assumption has enabled us to obtain qualitatively different guarantees from the existing literature on high dimensional inference [Javanmard and Montanari, 2014, Van de Geer et al., 2014, Zhang and Zhang, 2014]; and, to our knowledge, our result is the first to provide sparsity-adaptive inference about a dense functional of β_{c}, namely in our case τ.

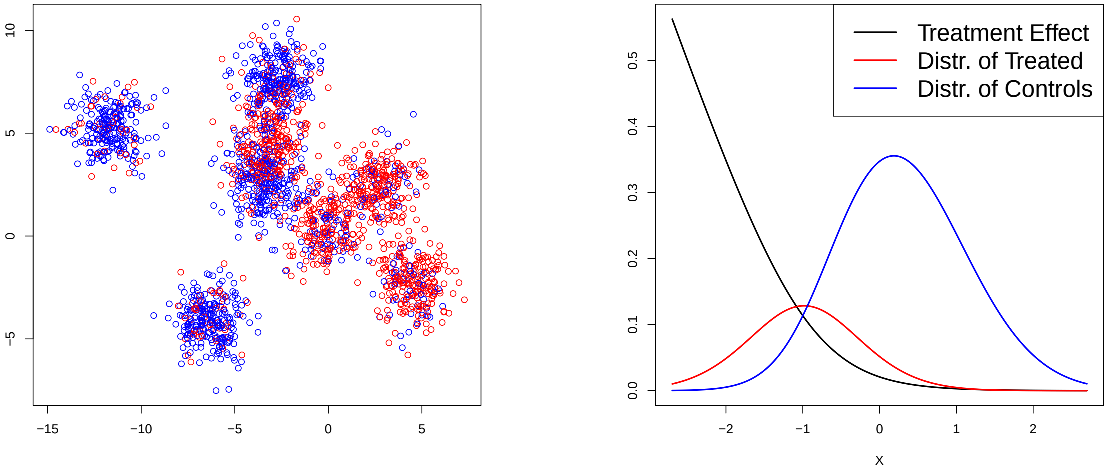
4 Simulation Experiments

In order to evaluate the finite-sample performance of our method, we first compare its performance in estimating τ to several other proposals available in the literature. After that, we consider the coverage of our confidence intervals as proposed in Section 3.3. All numbers reported in Tables 1–5 are averaged over 1000 simulation replications.

4.1 Methods under Comparison

In addition to **approximate residual balancing** as described in Procedure 1, the methods we use as baselines are as follows: **naive**, or difference-in-means estimation $\hat{\tau} = \overline{Y}_{\mathrm{t}} - \overline{Y}_{\mathrm{c}}$, which simply ignores the covariate information X; **elastic net** estimation [Zou and Hastie, 2005], or equivalently, Procedure 1 with trivial weights $\gamma_i = 1/n_{\mathrm{c}}$; **approximately balanced** estimation [Zubizarreta, 2015], or equivalently, Procedure 1 with trivial parameter estimates $\hat{\beta}_{\mathrm{c}} = 0$; **inverse-propensity weighting**, which uses (7) and (8), together with propensity estimates $\hat{e}(X_i)$ obtained by elastic net logistic regression, with the propensity scores trimmed at 0.05 and 0.95; **inverse-propensity residual weighting**, which pairs elastic net regression adjustments with the above inverse-propensity weights by plugging both into (10) [Farrell, 2015]; and **ordinary least squares after model selection** where, in the spirit of Belloni et al. [2014], we run lasso linear regression for $Y \,\big|\, X,\, W = 0$ and lasso logistic regression for $W \,\big|\, X$, and then compute the ordinary least squares estimate for τ on the union of the support of the three lasso problems.

Whenever there is a “λ” regularization parameter to be selected, we use cross validation with the `lambda.1se` rule from the `glmnet` package [Friedman et al., 2010]. In Belloni et al. [2014], the authors



(a) Low-dimensional version of the many clusters simulation setting. The blue and red dots denote control and treated X-observations respectively.

(b) Schematic of misspecified simulation setting, along the first covariate $(X_i)_1$. The "treatment effect" curve is not to scale along the Y-axis.

Figure 1: Illustrating simulation designs.

recommend selecting λ using more sophisticated methods, such as the square-root lasso [Belloni et al., 2011]. However, in our simulations, our implementation of Belloni et al. [2014] still attains excellent performance in the regimes the method is designed to work in. Similarly, our confidence intervals for τ are built using a cross-validated choice of λ instead of the fixed choice assumed by Corollary 4, as motivated by recent results [Chatterjee and Jafarov, 2015, Reid et al., 2016]. Our implementation of approximate residual balancing, as well as all the discussed baselines, is available in the R-package `balanceHD`.

4.2 Simulation Designs

We consider four different simulation settings. Our first setting is a **two-cluster** layout, where half the data is drawn as $X_i \sim \mathcal{N}(C_i,\, I_{p\times p})$, while $C_i \in \{0,\, \delta\}$ such that $\mathbb{P}\left[C_i = 0 \,\middle|\, W_i = 0\right] = 0.8$ and $\mathbb{P}\left[C_1 = 0 \,\middle|\, W_i = 1\right] = 0.2$. We consider two settings for the between-cluster vector δ: a "dense" setting where $\delta = 4/\sqrt{n}\ \mathbf{1}$, and a "sparse" setting where $\delta_j = 40/\sqrt{n}\ 1\left(\{j = 1 \text{ modulo } 10\}\right)$. We generated our data as $Y_i = X_i \cdot \beta + 10\, W_i + \varepsilon_i$ with $W_i = \text{Bernoulli}(0.5)$ and $\varepsilon_i \sim \mathcal{N}(0,\, 1)$, where β is one of:

$$\text{dense} : \beta \propto (1,\, 1/\sqrt{2},\, ...,\, 1/\sqrt{p}), \quad \text{harmonic} : \beta \propto (1/10, 1/11, ..., 1/\left(p+9\right)),$$
$$\text{moderately sparse} : \beta \propto (\underbrace{10,\, ...,\, 10}_{10},\, \underbrace{1,\, ...,\, 1}_{90},\, \underbrace{0,\, ...,\, 0}_{p-100}),\ \text{and very sparse} : \beta \propto (\underbrace{1,\, ...,\, 1}_{10},\, \underbrace{0,\, ...,\, 0}_{p-10}).$$

In each case we scaled β such that $\|\beta\|_2 = 10$. Finally, we set $n = 300$ and $p = 800$.

Our second **many-cluster** layout is closely related to the first, except now we have 20 cluster centers $C_i \in \{c_1,\, ...,\, c_{20}\}$, where all the cluster centers are independently generated as $c_k \sim \mathcal{N}(0,\, I_{p\times p})$. To generate the data, we first draw C_i uniformly at random from one of the 20 cluster centers and then set $W_i = 1$ with probability η for the first 10 clusters and $W_i = 1$ with probability $1-\eta$ for the last 10 clusters; we tried both $\eta = 0.1$ and $\eta = 0.25$. We used the same choices of β as above, except now we normalized them to $\|\beta\|_2 = 18$. We again used $n = 300$ and $p = 800$. We illustrate this simulation concept in Figure 1a; we purposefully chose a treatment assignment mechanism where log-linear propensity estimators may not perform well to highlight the fact our method only relies on overlap.

To test the robustness of all considered methods, we also ran a **misspecified** simulation. Here, we first drew $X_i \sim \mathcal{N}(0,\, I_{p\times p})$, and defined latent parameters $\theta_i = \log(1 + \exp(-2 - 2 * (X_i)_1))/0.915$.

We then drew $W_i \sim \text{Bernoulli}(1 - e^{-\theta_i})$, and finally $Y_i = (X_i)_1 + \cdots + (X_i)_{10} + \theta_i(2W_i - 1)/2 + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, 1)$. We varied n and p. This simulation setting, loosely inspired by the classic program evaluation dataset of LaLonde [1986], is illustrated in Figure 1b; note that the average treatment effect on the treated is much greater than the overall average treatment effect here.

Finally, we considered a **two-stage** setting closely inspired by an experiment of Belloni et al. [2014]. Here $X_i \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$, and $\theta_i = X_i \cdot \beta_1 + \varepsilon_{i1}$. Then, $W_i \sim \text{Bernoulli}(1/(1 + e^{\theta_i}))$, and finally $Y_i = X_i \cdot \beta_2 + 0.5\, W_i + \varepsilon_{i2}$ where ε_{i1} and ε_{i2} are independent standard Gaussian. Following Belloni et al. [2014], we set the structure model as $(\beta)_j \propto 1/j^2$ for $j = 1, ..., p$. However, for the propensity model, we once follow their paper and use a "sparse" propensity model $(\beta_P)_j \propto 1/j^2$, but also try a "dense" propensity model $(\beta_P)_j \propto 1$. We set $n = 100$ and $p = 200$. Note that this sparse setting is in fact very sparse; adjusting for differences in the two most important covariates removses 95% of the bias associated with all the covariates. In contrast, for example, in the first and fifth columns in Table 1 it would require adjusting for differences in the 700 or 90 most important covariates to remove 95% of the bias associated with all the covariates.

4.3 Accuracy Results

In the first two experiments, for which we report results in Tables 1 and 2, the outcome model $Y|X$ is reasonably sparse, while the propensity model has overlap but is not in general sparse. In fact, for Table 2, the propensity model does not even have a linear log odds ratio. Here approximate residual balancing does well, while none of the other methods can successfully fit large effects while mitigating bias due to small effects. When β is very sparse, methods that only seek to fit β—namely the elastic net and least squares with model selection—do quite well. We find that in general the balancing performs substantially better than propensity score weighting, with or without direct covariate adjustment. We also find that combining direct covariate adjustment with weighting does better than weighting on its own, irrespective of whether the weighting is based on balance or on the propensity score.

Encouragingly, approximate residual balancing also does a good job in the misspecified setting from Table 3. It appears that our stipulation that the approximately balancing weights (12) must be non-negative (i.e., $\gamma_i \geq 0$) helps prevent our method from interpolating too aggressively. Conversely, least squares with model selection does not perform well despite both the outcome and propensity models being sparse; apparently, it is more sensitive to the misspecification here.

Finally, in Table 4, we find that the method of Belloni et al. [2014] has excellent performance—as expected—when both the propensity and outcome models are sparse. However, if we make the propensity model dense, its performance decays substantially, and both approximate residual balancing and the elastic net do better.

4.4 Coverage Results

We evaluate coverage of confidence intervals in the "many-cluster" setting for different choices of β, n, and p; results are given in Table 5. We see that coverage is generally better with more ($\eta = 0.25$) rather than less ($\eta = 0.1$) overlap, and with sparser choices of β. Moreover, coverage rates appear to improve as n increases, suggesting that we are in a regime where the asymptotics from Corollary 4 are beginning to apply.

5 Proofs

Proof of Proposition 1

First, we can write

$$\begin{aligned}\hat{\mu}_c &= \overline{X}_t^\top \hat{\beta} + \gamma^\top \left(\mathbf{Y}_c - \mathbf{X}_c \hat{\beta}\right) \\ &= \overline{X}_t^\top \hat{\beta} + \gamma^\top \mathbf{X}_c \left(\beta - \hat{\beta}\right) + \gamma^\top \varepsilon_c.\end{aligned}$$

Beta Model	dense		harmonic		moderately sparse		very sparse	
Propensity Model	dense	sparse	dense	sparse	dense	sparse	dense	sparse
Naive	2.847	3.158	1.920	2.136	0.817	0.812	0.453	0.456
Elastic Net	1.822	0.445	1.127	0.304	0.296	0.113	0.034	0.029
Approximate Balance	1.670	0.621	1.133	0.442	0.499	0.224	0.289	0.182
Approx. Resid. Balance	**1.576**	**0.207**	**0.973**	**0.183**	**0.243**	**0.080**	**0.027**	**0.024**
Inverse Prop. Weight	2.368	1.511	1.594	1.029	0.686	0.415	0.384	0.251
Inv. Prop. Resid. Weight	2.234	1.610	1.458	1.107	0.534	0.426	0.239	0.231
Double-Select + OLS	1.814	0.228	1.126	0.209	0.290	0.096	0.034	**0.024**

Table 1: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}/\tau$ in the two-cluster setting.

Beta Model	dense		harmonic		moderately sparse		very sparse	
Overlap (η)	0.1	0.25	0.1	0.25	0.1	0.25	0.1	0.25
Naive	0.672	0.498	0.688	0.484	0.686	0.484	0.714	0.485
Elastic Net	0.451	0.302	0.423	0.260	0.181	0.114	0.031	**0.021**
Approximate Balance	0.470	0.317	0.498	0.292	0.489	0.302	0.500	0.302
Approx. Resid. Balance	**0.412**	**0.273**	**0.399**	**0.243**	**0.172**	**0.111**	**0.030**	**0.021**
Inverse Prop. Weight	0.491	0.396	0.513	0.376	0.513	0.388	0.533	0.380
Inv. Prop. Resid. Weight	0.463	0.352	0.479	0.326	0.389	0.273	0.363	0.248
Double-Select + OLS	0.679	0.368	0.595	0.329	0.239	0.145	0.047	0.023

Table 2: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}/\tau$ in the many-cluster setting.

n	400					1000				
p	100	200	400	800	1600	100	200	400	800	1600
Naive	1.72	1.73	1.73	1.72	1.74	1.71	1.70	1.72	1.70	1.72
Elastic Net	0.44	0.46	0.50	0.51	0.54	0.37	0.39	0.39	0.40	0.42
Approximate Balance	0.48	0.55	0.61	0.63	0.70	0.24	0.30	0.38	0.40	0.45
Approx. Resid. Balance	**0.24**	**0.26**	**0.28**	**0.29**	**0.32**	**0.16**	**0.17**	**0.18**	**0.19**	**0.20**
Inverse Prop. Weight	1.04	1.07	1.11	1.13	1.18	0.82	0.84	0.88	0.89	0.94
Inv. Prop. Resid. Weight	1.29	1.30	1.31	1.31	1.33	1.25	1.25	1.26	1.25	1.28
Double-Select + OLS	0.28	0.29	0.31	0.31	0.34	0.24	0.25	0.25	0.25	0.26

Table 3: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}/\tau$ in the misspecified setting.

Propensity Model	sparse				dense			
First Stage Sig. Strength	$\Vert\beta_P\Vert_2 = 1$		$\Vert\beta_P\Vert_2 = 4$		$\Vert\beta_P\Vert_2 = 1$		$\Vert\beta_P\Vert_2 = 4$	
Structure Sig. Strength ($\Vert\beta\Vert_2$)	1	4	1	4	1	4	1	4
Naive	1.998	7.761	3.466	13.605	0.692	2.264	0.782	2.577
Elastic Net	1.272	1.376	2.755	3.165	**0.529**	0.645	**0.590**	0.774
Approximate Balance	1.017	3.570	1.785	6.552	0.705	2.063	0.811	2.505
Approx. Resid. Balance	0.775	0.874	1.550	1.959	0.563	**0.637**	0.634	**0.765**
Inverse Prop. Weight	1.692	6.454	2.591	10.080	0.690	2.217	0.774	2.495
Inv. Prop. Resid. Weight	1.449	4.325	2.434	7.666	0.601	1.381	0.670	1.591
Double-Select + OLS	**0.608**	**0.703**	**0.985**	**1.223**	0.634	0.695	1.366	1.323

Table 4: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}/\tau$ in the two-stage setting of Belloni et al. [2014].

		$\beta_j \propto 1(\{j \leq 10\})$		$\beta_j \propto 1/j^2$		$\beta_j \propto 1/j$	
n	p	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$
200	400	0.90	0.84	0.94	0.88	0.84	0.71
200	800	0.86	0.76	0.92	0.85	0.82	0.71
200	1600	0.84	0.74	0.93	0.85	0.85	0.73
400	400	0.94	0.90	0.97	0.93	0.90	0.78
400	800	0.93	0.91	0.95	0.90	0.88	0.76
400	1600	0.93	0.88	0.94	0.90	0.86	0.76
800	400	0.96	0.95	0.98	0.96	0.96	0.90
800	800	0.96	0.94	0.97	0.96	0.94	0.90
800	1600	0.95	0.92	0.97	0.95	0.93	0.86

Table 5: Coverage for approximate residual balancing confidence intervals as constructed in Section 3.3, with data generated as in the many cluster setting. The target coverage is 0.95.

Thus,

$$\begin{aligned}\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}} &= \overline{X}_{\mathrm{t}}^\top \left(\hat{\beta} - \beta\right) + \gamma^\top \mathbf{X}_{\mathrm{c}} \left(\beta - \hat{\beta}\right) + \gamma^\top \varepsilon_c \\ &= \left(\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right)^\top \left(\hat{\beta} - \beta\right) + \gamma^\top \varepsilon_c,\end{aligned}$$

and so the desired conclusion follows by Hölder's inequality.

Proof of Lemma 2

In order to verify this claim it suffices to show that, with probability $1 - \delta$, there exists a weight vector γ satisfying both (21) and (22), as well as the equality constraint $\sum_{\{i : W_i = 0\}} \gamma_i = 1$. Here, we take a constructive approach, and define

$$\gamma_i^* = \frac{e\left(X_i\right)}{1 - e\left(X_i\right)} \Big/ \sum_{\{i : W_i = 0\}} \frac{e\left(X_i\right)}{1 - e\left(X_i\right)}$$

for all i for which $W_i = 0$, and 0 else. Our goal is to verify that our construction succeeds with probability at least $1 - \delta$. Notice that these weights also trivially satisfy $\gamma_i^* \leq n_{\mathrm{c}}^{-2/3}$ once n_{c} is large enough.

To do so, first note that because $\sum \gamma_i^* = 1$, our main quantity of interest $\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}} \gamma^*$ is translation invariant (i.e., we can map $X_i \to X_i + c$ for any $c \in \mathbb{R}^p$ without altering the quantity). Thus, we can without loss of generality re-center our problem such that $\mathbb{E}\left[X_i \,\middle|\, W_i = 1\right] = 0$. Given this re-centering, we use standard manipulations of sub-Gaussian random variables to check that, conditionally on n_{c} and n_{t} and for every $j = 1, \ldots, p$:

- $\overline{X}_{\mathrm{t},j} = n_{\mathrm{t}}^{-1} \sum_{\{i : W_i = 1\}} X_{ij}$ is sub-Gaussian with parameter ν^2 / n_{t} by Assumption 4.
- $A_j := n_{\mathrm{c}}^{-1} \sum_{\{i : W_i = 0\}} X_{ij}\, e(X_i)/(1 - e(X_i))$ is sub-Gaussian with parameter $\nu^2 (1 - \eta)^2 / (n_{\mathrm{c}}\, \eta^2)$ by Assumption 4 and because $e(X_i) \leq 1 - \eta$. Note that, by construction $\mathbb{E}\left[A_j\right] = \mathbb{E}\left[X_j \,\middle|\, W = 1\right]$, and so given our re-centering $\mathbb{E}\left[A_j\right] = 0$.
- $D := n_{\mathrm{c}}^{-1} \sum_{\{i : W_i = 0\}} e(X_i)/(1 - e(X_i)) - \rho$ is sub-Gaussian with parameter $(1 - \eta)^2 / (4 n_{\mathrm{c}}\, \eta^2)$, where $\rho = \mathbb{P}\left[W = 1\right] / \mathbb{P}\left[W = 0\right]$ denotes the odds ratio.
- $V := n_{\mathrm{c}}^{-1} \sum_{\{i : W_i = 0\}} (e(X_i)/(1 - e(X_i)))^2$ is sub-Gaussian with parameter $(1 - \eta)^4 / (4 n_{\mathrm{c}}\, \eta^4)$ after re-centering.

Next, we apply a union bound, by which, for any $\delta > 0$, the following event $\mathcal{E}_\delta$ occurs with probability at least $1 - \delta$:

$$\begin{aligned}
&\|A\|_\infty \le \nu\,(1-\eta)\,/(\eta\,\sqrt{n_{\mathrm{c}}})\,\sqrt{2\log(10\,p\,\delta^{-1})},\\
&\left\|\overline{X}_{\mathrm{t}} - A\right\|_\infty \le \nu\,\sqrt{1/n_{\mathrm{t}} + (1-\eta)^2\,/\,(n_{\mathrm{c}}\,\eta^2)}\,\sqrt{2\log(10\,p\,\delta^{-1})},\\
&|D| \le (1-\eta)\,/(2\eta\,\sqrt{n_{\mathrm{c}}})\,\sqrt{2\log(10\,\delta^{-1})},\ \text{ and}\\
&V \le \mathbb{E}\left[V\right] + (1-\eta)^2\,/(2\eta^2\,\sqrt{n_{\mathrm{c}}})\,\sqrt{2\log(10\,\delta^{-1})}.
\end{aligned}$$

We then see that on the event $\mathcal{E}_\delta$,

$$\begin{aligned}
\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top\gamma^*\right\|_\infty &= \left\|\overline{X}_{\mathrm{t}} - (\rho+D)^{-1}\,A\right\|_\infty \le \left\|\overline{X}_{\mathrm{t}} - A\right\|_\infty + \left|\frac{D}{\rho+D}\right|\,\|A\|_\infty\\
&\le \nu\sqrt{\frac{1}{n_{\mathrm{t}}} + \frac{(1-\eta)^2}{n_{\mathrm{c}}\,\eta^2}}\,\sqrt{2\log\left(\frac{10\,p}{\delta}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right).
\end{aligned}$$

Moreover, noting that

$$\mathbb{E}\left[V\right] = \mathbb{E}\left[\frac{e(X_i)^2}{(1-e(X_i))^2}\,\middle|\,W_i = 0\right] \le \frac{(1-\eta)^2}{\eta^2},$$

we see that on ε_δ,

$$n_{\mathrm{c}}\,\|\gamma^*\|_2^2 = \frac{V}{(\rho+D)^2} \le \frac{\mathbb{E}\left[V\right]}{\rho^2} + \left(\frac{1}{2} + \frac{1-\eta}{\rho\,\eta}\right)\,\frac{(1-\eta)^2}{\rho^2\eta^2}\,\sqrt{\frac{2}{n_{\mathrm{c}}}\log\left(\frac{10}{\delta}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right).$$

Thus, there exists a γ satisfying all desired constraints; thus, there must also be some $\zeta \in (0,\,1)$ for which (20) yields such a solution.

Proof of Theorem 3

First, note that by Assumptions 4 and 5, the design matrix is column standardized in the sense that, with probability tending to 1, $n_{\mathrm{c}}^{-1}\sum_{\{i:W_i=0\}} X_{ij}^2 \le (5/4)^2\nu^2$ for all $j = 1,\,\ldots,\,p$. Thus, given Assumption 5, 6, and the sub-Gaussianity of the noise $\varepsilon_i(w)$, Corollary 2 of Negahban et al. [2012] shows that if we obtain $\hat{\beta}_{\mathrm{c}}$ by running the lasso with $\lambda = 5\,\nu\,\upsilon\,\sqrt{\log(p)/n_{\mathrm{c}}}$, then, with probability tending to 1,

$$\left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 \le \frac{5\nu}{4}\,\frac{24\,\upsilon}{\omega}\,k\,\sqrt{\frac{\log p}{n_{\mathrm{c}}}}. \tag{31}$$

Formally, to get this result, we first scale down the design by a factor $5\nu/4$, and then apply the cited result verbatim; note that we also need to re-scale the restricted eigenvalue parameter ω.

Next, given our hypotheses and picking ζ according to Lemma 2 with $\delta = p^{-1}$, we find that with probability tending to 1,

$$\begin{aligned}
&\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top\gamma\right\|_\infty \le \nu\sqrt{5\,\log\left(p\right)\left(\frac{1}{n_{\mathrm{t}}} + \frac{(1-\eta)^2}{n_{\mathrm{c}}\,\eta^2}\right)} \quad \text{and}\\
&n_{\mathrm{c}}\,\left\|\gamma\right\|_2^2 \le \frac{1}{\rho^2}\,\mathbb{E}\left[\left(\frac{e(X_i)}{1-e(X_i)}\right)^2\,\middle|\,W_i = 0\right] + \frac{(1-\eta)^2\,(2-2\eta+\rho\eta)}{2\rho^3\eta^3}\,\sqrt{\frac{5}{n_{\mathrm{c}}}}\,\log\left(p\right).
\end{aligned}$$

Given these inequalities, we work from the representation in Proposition 1 to establish our desired asymptotic normality result. First, given the above scalings, we see that

$$\sqrt{n}\,\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top\gamma\right\|_\infty\,\left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 = \mathcal{O}_P\left(\frac{k\,\log(p)}{\sqrt{n}}\right).$$

Thus, thanks to Assumption 5 and Proposition 1, we see that

$$\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}} = \sum_{\{i : W_i = 0\}} \gamma_i \, \varepsilon_i(0) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{32}$$

Given this expression, we only need to verify that

$$\frac{1}{\|\gamma\|_2} \sum_{\{i : W_i = 0\}} \gamma_i \, \varepsilon_i(0) \Rightarrow \mathcal{N}\left(0, \, \sigma^2\right). \tag{33}$$

Because $\sum \gamma_i = 1$ and $\|\gamma\|_0 \leq n_{\mathrm{c}}$, we immediately see that $\|\gamma\|_2^{-1} \leq 1/\sqrt{n_{\mathrm{c}}}$; thus, Slutsky's inequality with (32) and (33) implies the first part of (25). Meanwhile, the second part of (25) follows immediately by noting that $n_{\mathrm{t}}(\hat{\mu}_{\mathrm{t}} - \mu_{\mathrm{t}})$ is asymptotically normal with variance σ^2, and is uncorrelated with $\hat{\mu}_{\mathrm{c}}$. Finally, (26) follows directly from Lemma 2.

We now turn to verifying (33) using Lyapunov's method; our goal is to establish a central limit theorem conditionally on on γ. First, note that by unconfoundedness (Assumption 1), γ_i is independent of $\varepsilon_i(0)$ conditional on X_i. Thus, given our homoskedasticity assumption,

$$\mathbb{E}\left[\sum_{\{i : W_i = 0\}} \gamma_i \, \varepsilon_i(0) \,\middle|\, \gamma\right] = 0 \text{ and } \quad \mathrm{Var}\left[\sum_{\{i : W_i = 0\}} \gamma_i \, \varepsilon_i(0) \,\middle|\, \gamma\right] = \sigma^2 \, \|\gamma\|_2^2 \, .$$

Next, we can again use unconfoundedness to verify that

$$\mathbb{E}\left[\sum_{\{i : W_i = 0\}} \left(\gamma_i \, \varepsilon_i(0)\right)^3 \,\middle|\, \gamma\right] = \sum_{\{i : W_i = 0\}} \gamma_i^3 \; \mathbb{E}\left[\left(\varepsilon_i(0)\right)^3 \,\middle|\, X_i\right] \leq C_3 \, \upsilon^3 \sum_{\{i : W_i = 0\}} \gamma_i^3 \leq C_3 \, \upsilon^3 \, n_{\mathrm{c}}^{-2/3} \, \|\gamma\|_2^2$$

for some universal constant C_3, where the last inequality follows by sub-Gaussianity of ε and by noting the upper bound on γ_i in (20). Thus,

$$\mathbb{E}\left[\sum_{\{i : W_i = 0\}} \left(\gamma_i \, \varepsilon_i(0)\right)^3 \,\middle|\, \gamma\right] \Big/ \; \mathrm{Var}\left[\sum_{\{i : W_i = 0\}} \gamma_i \, \varepsilon_i(0) \,\middle|\, \gamma\right]^{3/2} = \mathcal{O}\left(n_{\mathrm{c}}^{-2/3} \, \|\gamma\|_2^{-1}\right) = o_P(1),$$

and so Lyapunov's theorem implies that (33) holds.

Proof of Corollary 4

First of all, we can use the argument of Theorem 3 verbatim to show that

$$\left(\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}\right) \Big/ \sqrt{V_c} \Rightarrow \mathcal{N}\left(0, \, 1\right), \quad V_c = \sum_{\{i : W_i = 0\}} \gamma_i^2 \; \mathrm{Var}\left[\varepsilon_i(0) \,\middle|\, X_i\right].$$

To establish this claim, note that our bias bound (32) did not rely on homskedasticity, and the Lyapunov central limit theorem remains valid as long as the conditional variance of $\varepsilon_i(0)$ remains bounded from below. Thus, in order to derive the pivot (28), we only need to show that $\widehat{V}_c / V_c \rightarrow_p 1$; the desired conclusion then follows from Slutsky's theorem. Now, to verify this latter result, it suffices to check that

$$\frac{1}{V_c} \sum_{\{i : W_i = 0\}} \gamma_i^2 \left(Y_i - X_i \cdot \beta_{\mathrm{c}}\right)^2 \rightarrow_p 1, \text{ and} \tag{34}$$

$$\frac{1}{V_c} \sum_{\{i : W_i = 0\}} \gamma_i^2 \left(X_i \cdot \left(\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right)\right)^2 \rightarrow_p 0. \tag{35}$$

To show the first convergence result, we can proceed as in the proof of Theorem 3 to verify that there is a universal constant C_4 for which

$$\operatorname{Var}\left[\sum_{\{i:W_i=0\}} \gamma_i^2 \left(Y_i - X_i \cdot \beta_{\mathrm{c}}\right)^2 \,\middle|\, \gamma\right] \leq C_4\, \upsilon^4\, \left\|\gamma\right\|_4^4 \leq C_4\, \upsilon^4\, n_{\mathrm{c}}^{-4/3}\, \left\|\gamma\right\|_2^2,$$

and so (34) holds by Markov's inequality. Meanwhile, to establish (35), we focus on the case $\liminf \log(p)/\log(n) > 0$. We omit the argument in the ultra-low dimensional case since, when $p \ll n^{0.01}$, there is no strong reason to run our method instead of classical methods based on ordinary least squares. Now, we first note the upper bound

$$\sum_{\{i:W_i=0\}} \gamma_i^2 \left(X_i \cdot \left(\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right)\right)^2 \leq \left\|\gamma\right\|_2^2 \left\|\mathbf{X}_{\mathrm{c}} \left(\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right)\right\|_\infty^2 \leq \left\|\gamma\right\|_2^2 \left\|\mathbf{X}_{\mathrm{c}}\right\|_\infty^2 \left\|\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right\|_1^2,$$

where the second step uses Hölder's inequality as in the proof of Proposition 1. Then, thanks to (27), we only need to check that

$$\left\|\mathbf{X}_{\mathrm{c}}\right\|_\infty^2 \left\|\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right\|_1^2 \to_p 0.$$

We can use sub-Gaussianity of X_i (Assumption 4) and the bound (31) on the L_1-error of $\hat{\beta}_{\mathrm{c}}$ to find a constant $C(\nu,\, \omega,\, \upsilon)$ for which

$$\left\|\mathbf{X}_{\mathrm{c}}\right\|_\infty^2 \left\|\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right\|_1^2 \leq C(\nu,\, \omega,\, \upsilon)\, \log\left(p\, n_{\mathrm{c}}\right)\; k^2\, \frac{\log\left(p\right)}{n_{\mathrm{c}}}$$

with probability tending to 1. Then, noting our sparsity condition on k (Assumption 5), we find that

$$\log\left(p\, n_{\mathrm{c}}\right)\; k^2\, \frac{\log\left(p\right)}{n_{\mathrm{c}}} \ll \frac{\log\left(p\, n_{\mathrm{c}}\right)}{\log(p)},$$

which is bounded from above whenever $\liminf \log(p)/\log(n) > 0$.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation with high-dimensional data. *Econometrica, forthcoming*, 2016.

Peter Bickel, Chris Klaassen, Yakov Ritov, and Jon Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon, and Bin Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *arXiv preprint arXiv:1507.03652*, 2015.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *arXiv preprint arXiv:1506.05539*, 2015.

Emmanuel Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*, 2015.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, pages 808–843, 2008.

Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv preprint arXiv:1502.00665*, 2015.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, page asn055, 2009.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Bryan Graham, Christine Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, pages 1053–1079, 2012.

Bryan Graham, Christine Pinto, and Daniel Egel. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, pages –, 2016.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

Judith Hellerstein and Guido Imbens. Imposing moment restrictions by weighting. *Review of Economics and Statistics*, 81(1):1–14, 1999.

Keisuke Hirano, Guido Imbens, Geert Ridder, and Donald Rubin. Combining panels with attrition and refreshment samples. *Econometrica*, pages 1645–1659, 2001.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Guido Imbens, Richard Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 1998.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press, 2015.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard and Andrea Montanari. De-biasing the lasso: Optimal sample size for gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

Joseph Kang and Joseph Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–529, 2007.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403, 2004.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4): 538–557, 2012.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Whitney K Newey and Richard J Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

Art B Owen. Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research*, 8: 761–773, 2007.

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67, 2016.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(1):122–129, 1995.

James Robins, Andrea Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(1):106–121, 1995.

Paul R Rosenbaum. *Observational Studies.* Springer, 2002.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Christoph Rothe and Sergio Firpo. Semiparametric estimation and inference using doubly robust moment conditions. *IZA discussion paper*, 2013.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Mark J Van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media, 2003.

Daniel Westreich, Justin Lessler, and Michele J Funk. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, 2010.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *arXiv preprint arXiv:1601.05890*, 2016.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.