Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions[*]

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

Current version November 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 performance of standard approaches to treatment effect estimation. In this paper, we develop a method for de-biasing penalized regression adjustments to allow sparse regression methods like the lasso to be used for $\sqrt{n}$-consistent inference of average treatment effects. Our method works under substantially weaker assumptions than other methods considered in the literature: Unlike high-dimensional doubly robust methods recently developed in econometrics, we do not need to assume that the treatment propensities are estimable, and unlike existing de-biasing techniques from the statistics literature, our method is not limited to considering sparse contrasts of the parameter vector. Instead, in addition standard assumptions used to make lasso regression on the outcome model consistent under 1-norm error, we only require overlap, i.e., that the propensity score be uniformly bounded away from 0 and 1. Procedurally, our 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

[*] We are grateful for detailed comments from Edgar Dobriban, Bryan Graham, Chris Hansen, Nishanth Mundru and Jose Zubizarreta, and for discussions with seminar participants at Columbia University, Cowles Foundation, Stanford University, UNC Chapel Hill, and University of Southern California.

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

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

[§] Assistant Professor of Operations, Information and Technology, Stanford GSB, swager@stanford.edu.

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 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 attempting to control for high-dimensional confounders using a regularized regression adjustment obtained via, e.g., the lasso, can result in substantial biases. 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) and Chernozhukov et al. (2016) discuss 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 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 trying to estimate treatment propensities, our approach seeks to directly de-bias penalized regression adjustments by optimizing bias bounds from linear theory. Given this approach, we show that $\sqrt{n}$-consistent inference of average treatment effects is possible even when the propensity score is not estimable. Instead, 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.

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.

Our goal is to tighten the connection between the estimation strategy and the objective of estimating the average treatment effect. To do so, we study the following two-stage approximate residual balancing algorithm. 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 implied balance and variance provided by the weights, rather than the fit of the propensity score. 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 the bias arising from a potential failure to adjust for a large number of 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 imbalance 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; it 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.

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 5, 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 $i = 1,\, \dots,\, n$, 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 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 conditional average 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); we also assume a linear model for the potential outcomes in both groups.

Assumption 1 (Unconfoundedness)**.**

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

Assumption 2 (Linearity).

$$\mu_c(x) = \mathbb{E}\left[Y_i(0) \,\middle|\, X = x\right] = x \cdot \beta_c, \qquad \mu_t(x) = \mathbb{E}\left[Y_i(1) \,\middle|\, X = x\right] = x \cdot \beta_t, \tag{4}$$

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

$$\tau = \mu_t - \mu_c, \quad \text{where} \quad \mu_t = \overline{X}_t \cdot \beta_t, \quad \mu_c = \overline{X}_t \cdot \beta_c, \quad \text{and} \quad \overline{X}_t = \frac{1}{n_t} \sum_{\{i:W_i=1\}} X_i. \tag{5}$$

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 use 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.

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 $\hat{\mu}_c$ as

$$\hat{\mu}_c = \sum_{\{i:W_i=0\}} \gamma_i \, Y_i^{\text{obs}}. \tag{6}$$

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

$$\gamma_i = \frac{e(X_i)}{1 - e(X_i)} \Bigg/ \sum_{\{i:W_j=0\}} \frac{e(X_j)}{1 - e(X_j)}. \tag{7}$$

To implement these methods researchers typically substitute an estimate of the propensity score into the expression for the weights (7). 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 (7) 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 because estimating the treatment effect then involves dividing by $1 - \hat{e}(X_i)$, small inaccuracies in $\hat{e}(X_i)$ can have large effects, especially when $e(x)$ can be close to one; 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 (6) 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 where $p < n_{\mathrm{c}}$ 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 $\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}}$. 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 the regression adjustment is in general impossible, and methods such as the lasso, ridge regression, or the elastic net may perform poorly when plugged in for β_{c}; 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 $\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}}$, with $\hat{\beta}_{\mathrm{c}}$ estimated by ordinary least squares on a selected variables, is implicitly equivalent to running (6) 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.

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 $\hat{\beta}_{\mathrm{c}}$ 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 the covariatees at least approximately, and apply these weights to the residuals (Cassel et al., 1976; Robins et al., 1994):

$$\hat{\mu}_{\text{c}} = \overline{X}_{\text{t}} \cdot \hat{\beta}_{\text{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_{\text{c}} \right). \tag{8}$$

In other words, we fit a model parametrized by β_{c} to capture some of the strong signals, and then use a non-parametric re-balancing of the control data on the features to extract left-over signal from the residuals $Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_{\text{c}}$. 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 (8) is that it yields a simple and powerful theoretical guarantee, as stated below. Recall that $\mathbf{X}_{\text{c}}$ is the feature matrix for the control units. Consider the difference between $\hat{\mu}_{\text{c}}$ and μ_{c} for our proposed approach: $\hat{\mu}_{\text{c}} - \mu_{\text{c}} = (\overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \gamma) \cdot (\hat{\beta}_{\text{c}} - \beta_{\text{c}}) + \gamma \cdot \varepsilon$, where ε is the intrinsic noise $\varepsilon_i = Y_i(0) - X_i \cdot \beta_{\text{c}}$. With only the regression adjustment and no weighting, the difference would be $\hat{\mu}_{\text{c,reg}} - \mu_{\text{c}} = (\overline{X}_{\text{t}} - \overline{X}_{\text{c}}) \cdot (\hat{\beta}_{\text{c}} - \beta_{\text{c}}) + \mathbf{1} \cdot \varepsilon / n_{\text{c}}$, and with only the weighting the difference would be $\hat{\mu}_{\text{c,weight}} - \mu_{\text{c}} = (\overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \gamma) \cdot \beta_{\text{c}} + \gamma \cdot \varepsilon$. 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 $\hat{\mu}_{\text{c,no-adj}} - \mu_{\text{c}} = (\overline{X}_{\text{t}} - \overline{X}_{\text{c}}) \cdot \beta_{\text{c}} + \mathbf{1} \cdot \varepsilon / n_{\text{c}}$. The regression reduces the bias from $(\overline{X}_{\text{t}} - \overline{X}_{\text{c}}) \cdot \beta_{\text{c}}$ to $(\overline{X}_{\text{t}} - \overline{X}_{\text{c}}) \cdot (\hat{\beta}_{\text{c}} - \beta_{\text{c}})$, which will be substantial reduction if the estimation error $(\hat{\beta}_{\text{c}} - \beta_{\text{c}})$ is small relative to β_{c}. The weighting further reduces this to $(\overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \gamma) \cdot (\hat{\beta}_{\text{c}} - \beta_{\text{c}})$, which may be helpful if there is a substantial difference between $\overline{X}_{\text{t}}$ and $\overline{X}_{\text{c}}$. 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 the appendix.

Proposition 1. *The estimator* (8) *satisfies* $|\hat{\mu}_{\text{c}} - \mu_{\text{c}}| \leq \left\| \overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \gamma \right\|_\infty \left\| \hat{\beta}_{\text{c}} - \beta_{\text{c}} \right\|_1 + \left| \sum_{\{i:W_i=0\}} \gamma_i \, \varepsilon_i \right|$.

This result decomposes the error of $\hat{\mu}_{\text{c}}$ into two parts. The first is the main term, depending on the design $\mathbf{X}_{\text{c}}$, 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, this bound 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 $\hat{\beta}_{\text{c}}$. We define *approximately balancing weights* as

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ (1 - \zeta) \left\| \tilde{\gamma} \right\|_2^2 + \zeta \left\| \overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \tilde{\gamma} \right\|_\infty^2 \text{ subject to } \sum \tilde{\gamma}_i = 1, \ \tilde{\gamma}_i \geq 0 \right\}, \tag{9}$$

for some $\zeta \in (0, 1)$. These weights, which are closely related to a recent proposal by Zubizarreta (2015), are designed to make both terms in the bound from Proposition 1 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 $\mathbf{X}_{\text{c}}^\top \gamma$, match the treated sample mean $\overline{X}_{\text{t}}$ as closely as possible. Below, we show that we can find a ζ that achieves our objective of bounding both terms of the bound; in our simulations we use $\zeta = 1/2$, which balances the square of the bias term and the variance.

Meanwhile, for estimating $\hat{\beta}_{\text{c}}$ 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 $\hat{\beta}_{\text{c}}$ as

$$\hat{\beta}_{\text{c}} = \arg\min_{\beta_{\text{c}}} \left\{ \sum_{\{i:W_i=0\}} (Y_i^{\text{obs}} - X_i^\top \beta_{\text{c}})^2 + \lambda \left((1 - \alpha) \left\| \beta_{\text{c}} \right\|_2^2 + \alpha \left\| \beta_{\text{c}} \right\|_1 \right). \right\}. \tag{10}$$

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

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta) \left\|\tilde{\gamma}\right\|_2^2 + \zeta \left\|\overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \tilde{\gamma}\right\|_\infty^2 \text{ s.t.} \sum_{\{i:W_i=0\}} \tilde{\gamma}_i = 1 \text{ and } \tilde{\gamma}_i \geq 0 \right\}. \tag{11}$$

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

$$\hat{\beta}_{\text{c}} = \text{argmin}_{\beta} \left\{ \sum_{\{i:W_i=0\}} \left(Y_i^{\text{obs}} - X_i \cdot \beta\right)^2 + \lambda \left((1-\alpha) \left\|\beta\right\|_2^2 + \alpha \left\|\beta\right\|_1 \right) \right\}. \tag{12}$$

3. Estimate the average treatment effect τ as

$$\hat{\tau} = \overline{Y}_{\text{t}} - \left(\overline{X}_{\text{t}} \cdot \hat{\beta}_{\text{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_{\text{c}}\right) \right). \tag{13}$$

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, $e(x) = \exp(x \cdot \theta)/(1 + \exp(x \cdot \theta))$. 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, $\hat{\theta}_{\text{ml}} = \arg\max_\theta \sum_{i=1}^n \{W_i X_i \cdot \theta - \ln(1 + \exp(X_i \cdot \theta))\}$. An alternative, less efficient, estimator for θ is the method of moments estimator $\hat{\theta}_{\text{mm}}$ that balances the covariates exactly: $\overline{X}_{\text{t}} = \sum_{\{i:W_i=0\}} X_i \exp(X_i \cdot \theta) \, / \sum_{\{j:W_j=0\}} \exp(X_j \cdot \theta)$, 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) in the context of a linear outcome model.

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). 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). Belloni et al. (2016), Chernozhukov et al. (2016) and Farrell (2015) study 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.

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), Farrell (2015), and Van der Laan and Rose (2011).

Our work differs from the "double machine learning" approach to treatment effect estimation studied by Belloni et al. (2016), Chernozhukov et al. (2016) and Farrell (2015) in that these methods all require the treatment propensities $e(x)$ estimable at an $n^{-1/4}$ rate; and then consider various methods that can be used for estimation $e(x)$, ranging from penalized regression (Farrell, 2015) to boosting (Chernozhukov et al., 2016). Here, by specifying our weights γ_i directly using moment constraints, we are able to sidestep any estimability requirements on the propensities; and simply assuming overlap is sufficient. From a mathematical perspective, our work is more closely related to recent advances in de-biased linear inference (Cai and Guo, 2015; Javanmard and Montanari, 2014, 2015; Ning and Liu, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014), as discussed further in Section 3.

Our approximately balancing weights (9) 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 (6) with weights

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty \leq t \right\}, \tag{14}$$

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 (14) 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 estimators have been considered by Graham et al. (2012, 2016) and Hainmueller (2012) in a setting where exact balancing is possible, with slightly different objection functions. For example, Hainmueller (2012) uses $-\sum_i \ln(\gamma_i)$ instead of $\sum_i \gamma_i^2$, leading to

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ -\sum_{\{i : W_i = 0\}} \log\left(\tilde{\gamma}_i\right) \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \ \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty = 0 \right\}. \tag{15}$$

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 (15).

3 Asymptotics of Approximate Residual Balancing

As we have already emphasized, approximate residual balancing is a method that enables us to do inference average treatment effects without needing to estimate treatment propensities as nuisance parameters. The method compensates for the weaker available assumptions on the treatment propensity function by relying more explicitly on linearity of the outcome function, as in Proposition 1.

This trade-off is also mirrored in our theoretical development. Unlike Belloni et al. (2016), Chernozhukov et al. (2016) or Farrell (2015) whose analysis builds on the semiparametric efficiency literature for treatment effect estimation (Bickel et al., 1998; Hahn, 1998; Hirano et al., 2003; Robins and Rotnitzky, 1995; Robins et al., 1995), our theory falls more naturally under the purview of the recent literature on inference in high-dimensional linear models (Cai and Guo, 2015; Javanmard and Montanari, 2014, 2015; Ning and Liu, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014).

3.1 Approximate Residual Balancing as Debiased Linear Estimation

Our goal is to understand the asymptotics our estimates for $\mu_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}$. In the interest of generality, however, we begin by considering a broader problem. Given an arbitrary linear contrast $\theta = \xi \cdot \beta_{\mathrm{c}}$, we define an "approximate residual balancing" estimator $\hat{\theta}$ for θ, and study conditions under which $\sqrt{n}(\hat{\theta} - \theta)$ has a Gaussian limit under $p \gg n$ asymptotics. This detour via linear theory will help highlight the statistical phenomena that make approximate residual balancing work, and explain why—unlike the methods of Belloni et al. (2016), Chernozhukov et al. (2016) or Farrell (2015)—our method does not require $n^{-1/4}$-rate estimability of the treatment propensity function $e(x)$.

The problem of estimating linear contrasts $\xi \cdot \beta_{\mathrm{c}}$ in high-dimensional regression problems has received considerable attention recently, including notable contributions by Javanmard and Montanari (2014, 2015), Van de Geer et al. (2014), and Zhang and Zhang (2014). This line of work, however, exclusively considers the setting where ξ is a sparse vector; in particular, these papers focus on the case where ξ is the j-th basis vector e_j, i.e., the target estimand is the j-th coordinate of β_{c}. Furthermore, Cai and Guo (2015) showed that $\sqrt{n}$-consistent inference about generic dense contrasts of β_{c} is in general impossible. In our setting, however, the contrast-defining vector $\overline{X}_{\mathrm{t}}$ is random and thus generically dense; moreover, we are interested in applications where $m_t = \mathbb{E}[\overline{X}_{\mathrm{t}}]$ itself may also be dense. Thus, an alternative analysis will be required.

Given these preliminaries, we study estimators for $\theta = \xi \cdot \beta_{\mathrm{c}}$ obtained by simply replacing $\overline{X}_{\mathrm{t}}$ with ξ in our approximate residual balancing algorithm, or, in other words, by pretending that the treated class is centered at ξ rather than $\overline{X}_{\mathrm{t}}$:[1]

$$\gamma = \operatorname{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \text{ subject to } \left\| \xi - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma} \right\|_\infty \leq K \sqrt{\frac{\log(p)}{n_{\mathrm{c}}}},\ \max_i |\tilde{\gamma}_i| \leq n_{\mathrm{c}}^{-2/3} \right\}, \tag{16}$$

$$\hat{\theta} = \xi \cdot \hat{\beta}_{\mathrm{c}} + \sum_{\{i : W_i = 0\}} \gamma_i \left(Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_{\mathrm{c}} \right), \tag{17}$$

where $\hat{\beta}_{\mathrm{c}}$ is a properly tuned sparse linear estimator and K is a tuning parameter discussed below. In the classical parameter estimation setting, i.e., with $\xi = e_j$, the above procedure is algorithmically equivalent to the one proposed by Javanmard and Montanari (2014, 2015); however, as discussed above, the focus of our analysis is different from theirs. Javanmard and Montanari (2014, 2015) study the consistency of the parameter estimator $\hat{\beta}_{\mathrm{c}}$ under more general conditions than us and, in particular, consider the use of fixed designs; meanwhile, our main interest is with dense rather than sparse contrast vectors ξ.

We start our analysis in Section 3.2 by considering the estimation of $\theta = \xi \cdot \beta_{\mathrm{c}}$ for potentially dense contrast ξ, and find conditions under which $\sqrt{n}$-consistent inference is possible provided that $\xi^\top \Sigma_c^{-1} \xi = \mathcal{O}(1)$, where Σ_c is the covariance of $\mathbf{X}_{\mathrm{c}}$. We note that, whenever Σ_c has latent correlation structure, it is possible to have $\xi^\top \Sigma_c^{-1} \xi = \mathcal{O}(1)$ even when ξ is dense and $\|\xi\|_2 \gg 1$. To our knowledge, this is the first sparsity-adaptive inference result about dense contrasts of β_{c}. Interestingly, the original debiased

[1] The optimization program (16) differs slightly from Procedure 1. We have written the problem in constraint form rather than in Lagrange form, and also added a requirement that $|\gamma_i| \leq n_{\mathrm{c}}^{-2/3}$. The motivation for the first change is that, although there is a one-to-one mapping between γ-solutions obtained in Lagrange versus constraint forms, the former problem is easier to tune in practice while the latter allows for a more transparent theoretical discussion. Meanwhile, the new condition $|\gamma_i| \leq n_{\mathrm{c}}^{-2/3}$ appears to hold in practice even if we do not explicitly enforce it; and a further analysis may find that this condition is redundant. We will revisit the constraints that $\sum \gamma_i = 1$ and $\gamma_i \geq 0$ from Procedure 1 in the following section.

lasso estimates $\hat{\beta}_{\mathrm{c}}^{(\text{debiased})}$ cannot be used for efficient inference about θ, and $\hat{\theta} = \xi \cdot \hat{\beta}_{\mathrm{c}}^{(\text{debiased})}$ would be a potentially inconsistent point estimate for θ. Rather, as our analysis makes clear, we must specify the contrast ξ we are interested when choosing how to debias the lasso.

Given this general result, we then move to our main goal, i.e., the estimation of $\mu_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}$. The key difficulty is that, due to randomness in $\overline{X}_{\mathrm{t}}$, the quantity $\overline{X}_{\mathrm{t}}^{\top} \Sigma_c^{-1} \overline{X}_{\mathrm{t}}$ will in general be much larger than 1. We propose two possible analyses: First, in Section 3.3, we extend our linear theory analysis, while Section 3.4 develops a simpler asymptotic theory that obtains slightly looser performance guarantees in exchange for making substantially weaker assumptions on the data-generating mechanism. Finally, in Section 3.5, we discuss practical, heteroskedasticity-robust inference. Through our analysis, we assume that $\hat{\beta}_{\mathrm{c}}$ is obtained via the lasso; however, we could just as well consider, e.g., the square-root lasso (Belloni et al., 2011) or sorted L_1-penalized regression (Su and Candes, 2016).

3.2 Debiasing Dense Contrasts

As we begin our analysis of $\hat{\theta}$ defined in (17), it is first important to note that the optimization program (16) is not always feasible. 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 ξ consists of n times "1" followed by n times "-1"; then $\left\| \xi - \mathbf{X}_{\mathrm{c}}^{\top} \gamma \right\|_{\infty} \geq 1$ for any $\gamma \in \mathbb{R}^{n_{\mathrm{c}}}$, and the approximation error does not improve as n_{c} and p both get large. Thus, our first task is to identify a class of problems for which the problem (16) has a solution with high probability. The following lemma establishes such a result for random designs, in the case of vectors ξ for which $\xi^{\top} \Sigma_c^{-1} \xi$ is bounded; here $\Sigma_c = \mathrm{Var}\left[X_i \,\middle|\, W_i = 0\right]$ denotes the population variance of control features. We also rely on the following regularity condition, which will be needed for an application of the Hanson-Wright concentration bound for quadratic forms following Rudelson and Vershynin (2013).

Assumption 3 (Transformed Independence Design)**.** Suppose that we have a sequence of random design problems with[2] $\mathbf{X}_{\mathrm{c}} = Q\, \Sigma_c^{\frac{1}{2}}$, where $\mathbb{E}\left[Q_{ij}\right] = 0$, $\mathrm{Var}\left[Q_{ij}\right] = 1$, for all indices i and j, and the individual entries Q_{ij} are all independent. Moreover suppose that the Q-matrix is sub-Gaussian for some $\varsigma > 0$, $\mathbb{E}\left[\exp\left[t\left(Q_{ij} - \mathbb{E}\left[Q_{ij}\right]\right)\right]\right] \leq \exp\left[\varsigma^2 t^2/2\right]$ for any $t > 0$, and that $(\Sigma_c)_{jj} \leq S$ for all $j = 1, \ldots, p$.

Lemma 2. *Suppose that we have a sequence of problems for which Assumption 3 holds and, moreover, $\xi^{\top} \Sigma_c^{-1} \xi \leq V$ for some constant $V > 0$. Then, there is a universal constant $C > 0$ such that, setting $K = C\varsigma^2 \sqrt{VS}$, the optimization problem (16) is feasible with probability tending to 1; and, in particular, the constraints are satisfied by*

$$\gamma_i^* = \frac{1}{n_{\mathrm{c}}} \xi^{\top} \Sigma_c^{-1} X_i. \tag{18}$$

The above lemma is the key to our analysis of approximate residual balancing. Because, with high probability, the weights γ^* from (18) provide one feasible solution to the constraint in (16); we conclude that, again with high probability, the actual weights we use for approximate residual balancing must satisfy $\|\gamma\|_2^2 \leq \|\gamma^*\|_2^2 \approx n_{\mathrm{c}}^{-1} \xi^{\top} \Sigma_c^{-1} \xi$. In order to turn this insight into a formal result, we need assumptions on both the sparsity of the signal and the covariance matrix Σ_c.

Assumption 4 (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., $\|\beta_{\mathrm{c}}\|_0 \leq k$, and that $k \log(p)/\sqrt{n} \to 0$.[3]

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}; Belloni et al. (2014) use a similar assumption where they allow for additional non-zero components, but they assume that beyond the largest

[2] In order to simplify our exposition, this assumption implicitly rules out the use of an intercept. Our analysis would go through verbatim, however, if we added an intercept $X_1 = 1$ to the design.

[3] In recent literature, there has been some interest in methods that require only require approximate, rather than exact, k-sparsity. We emphasize that our results also hold with approximate rather than exact sparsity, as we only use our sparsity assumption to get bounds on $\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\|_1$ that can be used in conjunction with Proposition 1. For simplicity of exposition, however, we restrict our present discussion to the case of exact sparsity.

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. 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.[4]

Next, 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); and, following Bickel et al. (2009), these results usually require that $\mathbf{X}_c$ satisfy a form of the restricted eigenvalue condition (e.g., Belloni et al., 2014; Meinshausen and Yu, 2009; Negahban et al., 2012). Below, we make a restricted eigenvalue assumption on $\Sigma_c^{1/2}$; then, we will use results from Rudelson and Zhou (2013) to verify that this also implies a restricted eigenvalue condition on $\mathbf{X}_c$.

Assumption 5 (Well-Conditioned Covariance)**.** Given the sparsity level k specified above, the covariance matrix $\Sigma_c^{1/2}$ of the control features satisfies the $\{k,\, 2\omega,\, 10\}$-restricted eigenvalue defined as follows, for some $\omega > 0$. For $1 \leq k \leq p$ and $L \geq 1$, define the set $\mathcal{C}_k(L)$ as

$$\mathcal{C}_k(L) = \left\{ \beta \in \mathbb{R}^p : \left\lVert \beta \right\rVert_1 \leq L \sum_{j=1}^{k} \left|\beta_{i_j}\right| \quad \text{for some} \quad 1 \leq i_1 < \ldots < i_j \leq p \right\}. \tag{19}$$

Then, $\Sigma_c^{1/2}$ satisfies the $\{k,\, \omega,\, L\}$-restricted eigenvalue condition if $\beta^\top \Sigma_c \beta \geq \omega \left\lVert \beta \right\rVert_2^2$ for all $\beta \in \mathcal{C}_k(L)$.

Theorem 3. *Under the conditions of Lemma 2 hold, suppose that the control outcomes $Y_i(0)$ are drawn from a sparse, linear model as in Assumptions 1, 2, 3 and 4, that $\Sigma_c^{1/2}$ satisfies the restricted eigenvalue property (Assumption 5), and that we have a minimum estimand size*[5] $\left\lVert \xi \right\rVert_\infty \geq \kappa > 0$*. Suppose, moreover, that we have homoskedastic noise:* $\mathrm{Var}[\varepsilon_i(0) \,\big|\, X_i] = \sigma^2$ *for all $i = 1, \ldots, n$, and also that the response noise $\varepsilon_i(0) := Y_i(0) - \mathbb{E}[Y_i(0) \,\big|\, X_i]$ is uniformly sub-Gaussian with parameter $v^2 S > 0$. Finally, suppose that we estimate $\hat\theta$ using* (17)*, with the optimization parameter K selected as in Lemma 2 and the lasso penalty parameter set to $\lambda_n = 5\varsigma^2 v \sqrt{\log(p) / n_c}$. Then, $\hat\theta$ is asymptotically Gaussian,*

$$\left(\hat\theta - \theta\right) \,/\, \left\lVert \gamma \right\rVert_2 \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right), \quad n_c \left\lVert \gamma \right\rVert_2^2 \,/\, \xi^\top \Sigma_c^{-1} \xi \leq 1 + o_p(1). \tag{20}$$

The statement of Theorem 3 suggests an intriguing connection between our debiased estimator (17), and the ordinary least-squares (OLS) estimator. Under classical large-sample asymptotics with $n \gg p$, it is well known that the OLS estimator, $\hat\theta^{(OLS)} = \xi^\top (\mathbf{X}_c^\top \mathbf{X}_c)^{-1} \mathbf{X}_c^\top Y$, satisfies

$$\sqrt{n_c}\left(\hat\theta^{(OLS)} - \theta\right) \,/\, \sqrt{\xi^\top \Sigma_c^{-1} \xi} \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right), \quad \text{and} \ \sqrt{n_c}\left(\hat\theta^{(OLS)} - \theta - \sum_{\{i : W_i = 0\}} \gamma_i^* \varepsilon_i(0)\right) \to_p 0, \tag{21}$$

where γ_i^* is as defined in (18). By comparing this characterization to our result in Theorem 3, it becomes apparent that our debiased estimator $\hat\theta$ has been able to recover the large-sample qualitative behavior of $\hat\theta^{(OLS)}$, despite being in a high-dimensional $p \gg n$ regime.

The connection between debiasing and OLS ought not appear too surprising. After all, under classical assumptions, $\hat\theta^{(OLS)}$ is known to be the minimum variance unbiased linear estimator for θ; while the weights γ in (16) were explicitly chosen to minimize the variance of $\hat\theta$ subject to the estimator being nearly unbiased. Developing a deeper understanding of the connection between debiased prediction and OLS would be of considerable interest.

[4] We are only aware of two exceptions to this assumption. In recent work, Javanmard and Montanari (2015), 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; meanwhile, Wager et al. (2016) show that lasso regression adjustments allow for efficient average treatment effect estimation in randomized trials even when $k \ll n \,/\, \log(p)$. The point in common between both results is that they let us weaken the sparsity requirements at the expense of considerably strengthening our knowledge of the X-distribution.

[5] The minimum estimand size assumption is needed to rule out pathological superefficient behavior. As a concrete example, suppose that $X_i \sim \mathcal{N}(0,\, I_{p \times p})$, and that $\xi_j = 1/\sqrt{p}$ for $j = 1, \ldots, p$ with $p \gg n_c$. Then, with high probability, the optimization problem (16) will yield $\gamma = 0$. This leaves us with a simple lasso estimator $\hat\theta = \xi \cdot \hat\beta_c$ whose risk scales as $\mathbb{E}[(\hat\theta - \theta)^2] = \mathcal{O}(k^2 \log(p)/(p n_c)) \ll 1/n_c$. The problem with this superefficient estimator is that it is not necessarily asymptotically Gaussian.

3.3 Application to Treatment Effect Estimation

The previous section developed a fairly general theory of debiased estimation of contrasts of the form $\xi \cdot \beta_{\mathrm{c}}$ for sparse β_{c}, under the assumption that $\xi^\top \Sigma_c^{-1} \xi$ remains bounded. Unfortunately, however, this result does not directly apply to our main problem of interest, namely estimating $\mu_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}$; the problem is that, in general, $\overline{X}_{\mathrm{t}}^\top \Sigma^{-1} \overline{X}_{\mathrm{t}}$ is on the order of p/n due to the randomness in $\overline{X}_{\mathrm{t}}$, thus directly violating our main assumption. In this section, we show how to get around this problem under the weaker assumption that $m_t^\top \Sigma_c^{-1} m_t$ is bounded, where ; i.e., we show that the stochasticity $\overline{X}_{\mathrm{t}}$ does not invalidate our result.

The following result also immediately implies a central limit theorem for $\hat{\tau} = \overline{Y}_t - \hat{\mu}_{\mathrm{c}}$ where $\overline{Y}_t$ is the average of the treated outcomes, since $\overline{Y}_t$ is uncorrelated with $\hat{\mu}_{\mathrm{c}}$ conditionally on $\overline{X}_{\mathrm{t}}$.

Corollary 4. *Under the conditions of Theorem 3, suppose that we want to estimate $\mu_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}$ by replacing ξ with $\overline{X}_{\mathrm{t}}$ in (17), and let $m_t = \mathbb{E}\left[X \,\middle|\, W = 1\right]$. Suppose, moreover, that we replace all the assumptions made about ξ in Theorem 3 with the following assumptions: throughout our sequence of problems, the vector m_t satisfies $m_t \Sigma_c^{-1} m_t \leq V$ and $\|m_t\|_\infty \geq \kappa$. Suppose, finally, that $(X_i - m_t)_j \,\big|\, W_i = 1$ is sub-Gaussian with parameter $\nu^2 > 0$, and that the overall odds of receiving treatment $\mathbb{P}\left[W = 1\right] / \mathbb{P}\left[W = 0\right]$ tend to a limit ρ bounded away from 0 and infinity. Then, setting the tuning parameter in (16) as $K = C\varsigma^2\sqrt{VS} + \nu\sqrt{2.1\,\rho}$, we get*

$$\left(\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}\right) \Big/ \left\|\gamma\right\|_2 \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right), \quad n_{\mathrm{c}} \left\|\gamma\right\|_2^2 \Big/ m_t^\top \Sigma_c^{-1} m_t \leq 1 + o_p(1). \tag{22}$$

The asymptotic variance bound $m_t^\top \Sigma_c^{-1} m_t$ is exactly the Mahalanobis distance between the mean treated and control subjects with respect to the covariance of the control sample. Thus, our 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 only to a requirement on the Mahalanobis distance between the treated and control classes, and 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, 2016), Chernozhukov et al. (2016) 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. In other words, by relying more heavily on linearity of the outcome function, we can considerably relax the assumptions required to get $\sqrt{n}$-consistent treatment effect estimation.

3.4 A Direct Analysis with Overlap

Our discussion so far, leading up to Corollary 4, gives a characterization of when and why we should expect approximate residual balancing to work. However, from a practical perspective, the assumptions used in our derivation were somewhat stronger than ones we may feel comfortable making in applications; the transformed independence design assumption being perhaps the most problematic one.

In this section, we propose an alternative analysis of approximate residual balancing that sheds many of the more delicate assumptions made above, and replaces them with overlap. 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; and, for this reason, overlap has become a common assumption in the literature on causal inference from observational studies (Crump et al., 2009; Imbens and Rubin, 2015). 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.

Assumption 6 (Overlap)**.** There is a constant $0 < \eta$ such that $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathbb{R}^p$.

Given these assumptions, we can replace all the assumptions made previously about X with the following technical conditions. Note that the requirements below are essentially necessary for our argument to make sense: In order for the lasso regression adjustment to be useful, we need $\mathbf{X}_{\mathrm{c}}$ to satisfy a restricted eigenvalue condition with high probability; and for the ∞-norm of the between class distances to concentrate, we need the features X_{ij} to have rapidly decaying tails.

Assumption 7 (Design)**.** Our design X satisfies the following two conditions. First, the design is sub-Gaussian, i.e., 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. Second, we assume that $\mathbf{X}_{\mathrm{c}}$ satisfies the $\{k,\, \omega,\, 4\}$-restricted eigenvalue condition as defined in Assumption 5, with probability tending to 1.

Given these conditions, we study the following estimator of $\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}$:

$$\gamma = \mathrm{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \ : \ \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \tilde{\gamma}\right\|_\infty \leq K \sqrt{\frac{\log(p)}{n_{\mathrm{c}}}},\ \sum_{\{i:W_i=0\}} \tilde{\gamma}_i = 1,\ 0 \leq \tilde{\gamma}_i \leq n_{\mathrm{c}}^{-2/3} \right\}, \tag{23}$$

$$\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_{\mathrm{c}}\right). \tag{24}$$

Note that, here, we have re-incorporated the positivity and sum constraints on γ. The positivity constraint stops us from interpolating outside of the support of the data, and appears to improve robustness to model misspecification. Meanwhile, the requirement that $\sum_{\{i:W_i=0\}} \gamma_i = 1$ is a practical trick that is comparable to not penalizing the intercept term in a penalized regression.

Following Lemma 2, our analysis again proceeds by guessing a feasible solution to (23), and then using it to bound the variance of our estimator. Here, however, we using inverse-propensity weights as our guess: $\gamma_i^* \propto e(X_i)/(1-e(X_i))$. Our proof hinges on showing that the actual weights we get from (23) are at least as good as these inverse-propensity weights, and thus our method will be at most as variable as one that used true inverse-propensity residual weighting.

Theorem 5. *Suppose that we have n independent and identically distributed training examples satisfying Assumptions 1, 2, 4, 6, 7, and that the treatment odds $\mathbb{P}[W=1]/\mathbb{P}[W=0]$ converge to ρ with $0 < \rho < \infty$. Suppose, moreover, that we have homoskedastic noise: $\mathrm{Var}[\varepsilon_i(w) \,|\, X_i] = \sigma^2$ for all $i = 1,\, ...,\, n$, and also that the response noise $\varepsilon_i(w) := Y_i(w) - \mathbb{E}[Y_i(w) \,|\, X_i]$ is uniformly sub-Gaussian with parameter $\upsilon^2 > 0$. Finally, suppose that we use (24) with $K = \nu\sqrt{2.1(\rho + (\eta^{-1}-1)^2}$ for estimation, with the lasso penalty parameter set to $\lambda_n = 5\nu\upsilon\sqrt{\log(p)/n_{\mathrm{c}}}$ instead of selecting λ_n by cross-validation. Then,*

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

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

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

The rate over convergence guaranteed by (26) is the same as what we would get if we actually knew the true propensities and could use them for weighting (Hahn, 1998). Here, we achieve this rate although we have no guarantees that the true propensities $e(X_i)$ are consistently estimable. Finally, we note that, when the assumptions to Corollary 4 hold, the bound (22) is stronger than (26); however, there exist designs where the bounds match.

3.5 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 6. *Under the conditions of Theorems 3 or 5, suppose instead that we have heteroskedastic noise $\upsilon_{\min}^2 \leq \mathrm{Var}\left[\varepsilon_i(W_i) \,\middle|\, X_i,\, W_i\right] \leq \upsilon^2$ for all $i = 1,\, ...,\, n$. Then, the following holds:*

$$\left(\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}\right) \Big/ \sqrt{\widehat{V}_c} \Rightarrow \mathcal{N}(0,\, 1), \quad \widehat{V}_c = \sum_{\{i:W_i=0\}} \gamma_i^2 \left(Y_i - X_i \cdot \hat{\beta}_{\mathrm{c}}\right)^2. \tag{27}$$

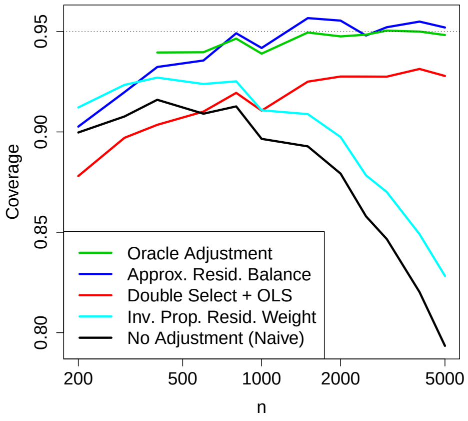


Figure 1: Finite sample coverage of the average treatment effect on the treated for different estimators, aggregated over 2,000 replications. The target coverage rate, 0.95, is denoted with a dotted line.

In order to provide inference about τ, we also need error bounds for $\hat{\mu}_{\text{t}}$. Under sparsity assumptions comparable to those made for β_{c} in Theorem 5, we can verify that

$$\left(\hat{\mu}_{\text{t}} - \mu_{\text{t}}\right) \Big/ \sqrt{\widehat{V}_t} \Rightarrow (0,\, 1)\,, \quad \widehat{V}_t = \frac{1}{n_{\text{t}}^2} \sum_{\{i : W_i = 1\}} \left(Y_i - X_i \hat{\beta}_{\text{t}}\right)^2, \tag{28}$$

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

4 Application: The Efficacy of Welfare-to-Work Programs

Starting in 1986, California implemented the Greater Avenues to Independence (GAIN) program, with an aim to reduce dependence on welfare and promote work among disadvantaged households. The GAIN program provided its participats with a mix of educational resources such as English as a second language courses and vocational training, and job search assistance. This program is described in detail by Hotz et al. (2006). In order to evaluate the effect of GAIN, the Manpower Development Research Corporation conducted a randomized study between 1988 and 1993, where a random subset of GAIN registrants were eligible to receive GAIN benefits immediately, whereas others were embargoed from the program until 1993 (after which point they were allowed to participate in the program). All experimental subjects were followed for a 3-year post-randomization period.

The randomization for the GAIN evaluation was conducted separately by county; following Hotz et al. (2006), we consider data from Alameda, Los Angeles, Riverside and San Diego counties. As discussed in detail in Hotz et al. (2006), the experimental conditions differed noticeably across counties, both in terms of the fraction of registrants eligible for GAIN, i.e., the treatment propensity, and in terms of the subjects participating in the experiment. For example, the GAIN programs in Riverside and San Diego counties sought to register all welfare cases in GAIN, while the programs in Alameda and Los Angeles counties focused on long-term welfare recipients.

The fact that the randomization of the GAIN evaluation was done at the county level rather than at the state level presents us with a natural opportunity to test our method, as follows. We seek to estimate the average treatment effect of GAIN on the treated; however, we hide the county information from our procedure, and instead try to compensate for sampling bias by controlling for a large amount of covariates. We used spline expansions of age and prior income, indicators for race, family status, etc., for a total of $p = 93$ covariates. Meanwhile, we can check our performance against a gold standard estimate of the average treatment effect that is stratified by county and thus guaranteed to be unbiased.[6]

We compare the behavior of different methods for estimating the average treatment effect on the treated using randomly drawn subsamples of the original data (the full dataset has $n = 19,170$). In addition to approximate residual balancing, we only consider baselines with formal inferential guarantees in high dimensions, namely double selection (Belloni et al., 2014), and inverse-propensity residual weighting (Belloni et al., 2016; Farrell, 2015). In addition, we also show the behavior of an "oracle" procedure that gets to observe the hidden county information and then simply estimates treatment effects for each county separately, and the "naive" difference-in-means estimator that ignores the features X. In very small samples, the oracle procedure was not always well defined because some samples may result in counties where either everyone or no one is treated.

Figure 1 compares the coverage of the different methods. Here, we see that approximate residual balancing achieves excellent performance in moderately large samples, and effectively gets nominal coverage. Double selection get reasonable coverage and improves with n; whereas inverse-propensity residual weighting barely improves over the naive difference-in-means estimator for the sample sizes under consideration. Meanwhile, in terms of MSE, double selection and approximate residual balancing both also perform well: approximate residual balancing has a comparable MSE to the oracle adjustment, while double selection can do 5-10% better in moderately large samples. It appears that double selection is effectively shrinking its predictions in a way that hurts coverage but improves MSE.

5 Simulation Experiments

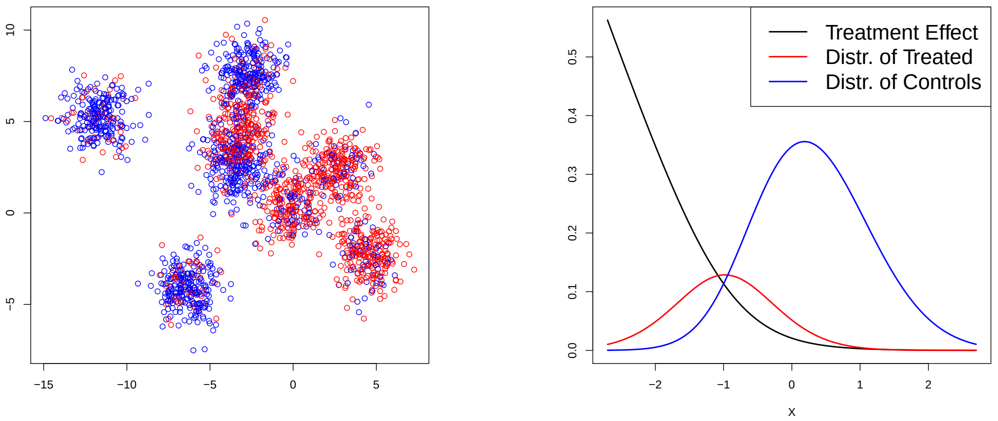
In order to evaluate the finite-sample performance of our method, we first compare its performance in estimating τ to several other proposals available in the literature. After that, we consider the coverage of our confidence intervals as proposed in Section 3.5. All numbers reported in Tables 1–5 are averaged over 1000 simulation replications.

5.1 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 $\hat{\tau} = \overline{Y}_{\text{t}} - \overline{Y}_{\text{c}}$, which simply ignores the covariate information X; **elastic net** estimation (Zou and Hastie, 2005), or equivalently, Procedure 1 with trivial weights $\gamma_i = 1/n_{\text{c}}$; **approximately balanced** estimation (Zubizarreta, 2015), or equivalently, Procedure 1 with trivial parameter estimates $\hat{\beta}_{\text{c}} = 0$; **inverse-propensity weighting**, which uses (6) and (7), 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 (8) (Belloni et al., 2016; 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 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 6. Our

[6] More formally, in our experiments, we set the gold standard using the county-stratified oracle estimator on bootstrap samples of the full $n = 19,170$ sample. We use bootstrap samples to correct for the correlation of estimators $\hat{\tau}$ obtained using the full dataset and subsamples of it. We also note that, given this setup, the quantity we are using as our goal standard is not and estimate of τ, i.e., the conditional average treatment effect on the treated sample, and should rather be thought of as an estimate of $\mathbb{E}\left[\tau\right]$, i.e., the average treatment effect on the treated population. Since we are in a setting with a fairly weak signal, this should not make a noticeable difference in practice.



(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 2: Illustrating simulation designs.

implementation of approximate residual balancing, as well as all the discussed baselines, is available in the R-package `balanceHD`.

5.2 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 $\delta_j = 40/\sqrt{n}\,1(\{j = 1 \text{ modulo } 10\})$. 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 2a; 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 2b; 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 $j = 1, \ldots, p$. 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.

5.3 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 (9) 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.

Meanwhile, 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.

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. Coverage is generally better with more overlap ($\eta = 0.25$) rather than less ($\eta = 0.1$), 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 6 are beginning to apply.

References

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

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

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

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation with high-dimensional data. *Econometrica, forthcoming*, 2016.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag, 1998.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

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

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

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	$\lVert\beta_P\rVert_2 = 1$		$\lVert\beta_P\rVert_2 = 4$		$\lVert\beta_P\rVert_2 = 1$		$\lVert\beta_P\rVert_2 = 4$	
Structure Sig. Strength ($\lVert\beta\rVert_2$)	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 3.5, with data generated as in the many cluster setting. The target coverage is 0.95.

C. M. Cassel, C. E. Särndal, and J. H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *JRSS-B*, 2015.

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

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

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, page asn055, 2009.

J.-C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *JASA*, 87(418):376–382, 1992.

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

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

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

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

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

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

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 2015.

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

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

K. Hirano, G. Imbens, G. Ridder, and D. Rubin. Combining panels with attrition and refreshment samples. *Econometrica*, pages 1645–1659, 2001.

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

V. J. Hotz, G. W. Imbens, and J. A. Klerman. Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the california GAIN program. *Journal of Labor Economics*, 24(3), 2006.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *JRSS-B*, 76(1):243–263, 2014.

G. Imbens, R. Spady, and P. Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 1998.

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

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

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

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

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

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

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

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

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

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765*, 2014.

A. B. Owen. Infinitely imbalanced logistic regression. *JMLR*, 8:761–773, 2007.

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

J. Robins, A. Rotnitzky, and L. 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.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

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

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

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

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

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

W. Su and E. Candes. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSS-B*, pages 267–288, 1996.

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

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 2016. doi: 10.1073/pnas.1614732113.

D. Westreich, J. Lessler, and M. 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.

C.-H. Zhang and S. 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.

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

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSS-B*, 67(2):301–320, 2005.

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

A Proofs

Proof of Proposition 1

First, we can write

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

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

For any $j = 1, \ldots, p$, write

$$\begin{aligned}\left(\mathbf{X}_{\mathrm{c}}^{\top}\gamma^*\right)_j &= \frac{1}{n_{\mathrm{c}}} e_j^{\top}\mathbf{X}_{\mathrm{c}}^{\top}\mathbf{X}_{\mathrm{c}}\Sigma_c^{-1}\xi\\ &= \frac{1}{n_{\mathrm{c}}}\sum_i Q_i^{\top}A_jQ_i, \quad A_j := \Sigma_c^{-\frac{1}{2}}\xi e_j^{\top}\Sigma_c^{\frac{1}{2}},\end{aligned}$$

where e_j is the j-th basis vector, and Q_i denotes the i-th row of the Q matrix as a column vector. Here, A_j is a rank-1 matrix, with Frobenius norm

$$\|A_j\|_F^2 = \operatorname{tr}\left(\Sigma_c^{\frac{1}{2}} e_j\xi^{\top}\Sigma_c^{-1}\xi e_j^{\top}\Sigma_c^{\frac{1}{2}}\right) = (\Sigma_c)_{jj}\,\xi^{\top}\Sigma_c^{-1}\xi \leq VS.$$

We can now apply the Hanson-Wright inequality, as presented in Theorem 1.1 of Rudelson and Vershynin (2013). Given our assumptions on Q_i—namely that it have independent, standardized, and sub-Gaussian entries—the Hanson-Wright inequality implies that $Q_i^{\top}A_jQ_i$ is sub-Exponential; more specifically, there exist universal constants C_1 and C_2 such that

$$\mathbb{E}\left[e^{t\left(Q_i^{\top}A_jQ_i - \mathbb{E}\left[Q_i^{\top}A_jQ_i\right]\right)}\right] \leq \exp\left[C_1t^2\varsigma^4VS\right] \text{ for all } t \leq \frac{C_2}{\varsigma^2\sqrt{VS}}.$$

Thus, noting that $\mathbb{E}\left[\mathbf{X}_{\mathrm{c}}^{\top}\gamma^*\right] = \xi$, we find that for any sequence $t_n > 0$ with $t_n^2/n \to 0$, the following relation holds for large enough n:

$$\mathbb{E}\left[\exp\left[\sqrt{n}\,t_n\left(\mathbf{X}_{\mathrm{c}}^{\top}\gamma^* - \xi\right)_j\right]\right] \leq \exp\left[C_1t_n^2\,\varsigma^4\,VS\right].$$

We can turn the above moment bound into a tail bound by applying Markov's inequality. Plugging in $t := \sqrt{\log(p/2\delta)}\,/\,(\varsigma^2\sqrt{C_1VS})$ and also applying a symmetric argument to $(-\mathbf{X}_{\mathrm{c}}^{\top}\gamma^* + \xi)_j$, we find that for large enough n and any $\delta > 0$,

$$\mathbb{P}\left[\left|\sqrt{n}\,\left(X^{\top}\gamma^* - \xi\right)_j\right| > 2\varsigma^2\sqrt{C_1VS\log\left(\frac{p}{2\delta}\right)}\right] \leq \frac{\delta}{p}.$$

The desired result then follows by applying a union bound, and noting that $\|\gamma^*\|_{\infty} \leq n^{-2/3}$ with probability tending to 1 by sub-Gaussianity of Q_{ij}.

Proof of Theorem 3

We start by mimicking Proposition 1, and write

$$\begin{aligned}\hat{\theta} - \theta &= \xi \cdot \left(\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right) + \sum_{\{i : W_i = 0\}} \gamma_i \left(Y_i - X_i \cdot \hat{\beta}_{\mathrm{c}}\right) \\ &= \sum_{\{i : W_i = 0\}} \gamma_i \varepsilon_i(0) + \left(\xi - \mathbf{X}_{\mathrm{c}}^\top \gamma\right) \cdot \left(\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right) \\ &= \sum_{\{i : W_i = 0\}} \gamma_i \varepsilon_i(0) + \mathcal{O}\left(\left\|\xi - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty \left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1\right)\end{aligned} \tag{29}$$

The proof of our main result now follows by analyzing the above bound using Lemma 2 from the main text, as well as technical results proved below in Lemmas 7 and 8.

We first consider the error term. On the event that (16) is feasible—which, by Lemma 2 will occur with probability tending to 1—we know that $\left\|\xi - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty = \mathcal{O}(\sqrt{\log(p)/n_{\mathrm{c}}})$. Meanwhile, given our assumptions, we can obtain an L_1-risk bound for the lasso that scales as $\mathcal{O}(k\sqrt{\log(p)/n_{\mathrm{c}}})$; see Lemma 7. Taken together, these results imply that

$$\left\|\xi - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty \left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 = \mathcal{O}\left(\frac{k \log(p)}{n_{\mathrm{c}}}\right), \tag{30}$$

which, by Assumption 4, decays faster than $1/\sqrt{n_{\mathrm{c}}}$.

Next, to rule out superefficiency, we need a lower bound on $\|\gamma\|_2^2$. By our minimum estimand size assumption we know that there exists an index $j \in \{1, \ldots, p\}$ with $|\xi_j| \geq \kappa$; and thus, any feasible solution to (16) must eventually satisfy $(\mathbf{X}_{\mathrm{c}}^\top \gamma)_j^2 \geq \kappa^2/2$. By Cauchy-Schwarz, this implies

$$\|\gamma\|_2^2 \geq \kappa^2 \Bigg/ \left(2 \sum_{\{i : W_i = 0\}} \mathbf{X}_{ij}^2\right) = \Theta_P\left(\frac{1}{n_{\mathrm{c}}}\right),$$

as desired. Given this result, a standard application of Lyapunov's central limit theorem (Lemma 8) paired with the bound (30) implies that, by Slutsky's theorem,

$$\left(\hat{\theta} - \theta\right) / \|\gamma\|_2^2 \Rightarrow \mathcal{N}\left(0, \sigma^2\right),$$

which was the first part of our desired conclusion.

Finally, we need to characterize the scale of the main term. To do so, consider the weights γ^* defined in (18). The concentration bound from Theorem 2.1 in Rudelson and Vershynin (2013) implies that $n_{\mathrm{c}} \|\gamma\|_2 / (\xi^\top \Sigma_c^{-1} \xi) \to_p 1$; thus, Lemma 2 implies that, with probability tending to 1, the optimization program for γ is feasible and

$$n_{\mathrm{c}} \|\gamma\|_2^2 \Big/ \left(\xi^\top \Sigma_c^{-1} \xi\right) \leq 1 + o_p(1),$$

thus concluding the proof.

Lemma 7. *Under the conditions of Theorem 3, the lasso satisfies*

$$\left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 \leq \frac{5\varsigma^2}{4} \frac{24\, \upsilon}{\omega} k \sqrt{\frac{\log p}{n_{\mathrm{c}}}}. \tag{31}$$

Proof. Given our well-conditioning assumptions on the covariance Σ_c, Theorem 6 of Rudelson and Zhou (2013) implies that the matrix $\mathbf{X}_{\mathrm{c}}$ will also satisfy the a weaker restricted eigenvalue property with high probability. Specifically, in our setting Assumption 4 implies that $\log(p) \ll \sqrt{n_{\mathrm{c}}}$, and so we can use the work of Rudelson and Zhou (2013) to conclude that $n_{\mathrm{c}}^{-1/2} \mathbf{X}_{\mathrm{c}}$ satisfies the $\{k, \omega, 4\}$-restricted eigenvalue condition with high probability.

Next, given Assumption 3, we can use Theorem 2.1 of Rudelson and Vershynin (2013) to verify that the design matrix is $\mathbf{X}_\mathrm{c}$ column standardized with high probability in the sense that, with probability tending to 1,

$$n_\mathrm{c}^{-1} \sum_{\{i:W_i=0\}} (\mathbf{X}_\mathrm{c})_{ij}^2 \leq (5/4)^2 \varsigma^4 S \text{ for all } j = 1, \ldots, p.$$

Thus, pairing these two fact about $\mathbf{X}_\mathrm{c}$ with sparsity as in Assumption 4 and the sub-Gaussianity of the noise smash$\varepsilon_i(w)$, we can use the results of Negahban et al. (2012) to bound the L_1-risk of the lasso. Specifically, their Corollary 2 implies that, if we obtain $\hat{\beta}_\mathrm{c}$ by running the lasso with $\lambda = 5\,\varsigma^2\,\upsilon\,S\,\sqrt{\log(p)/n_\mathrm{c}}$, then, with probability tending to 1, (31) holds. Formally, to get this result, we first scale down the design by a factor $5\varsigma^2/4$, and then apply the cited result verbatim; note that we also need to re-scale the restricted eigenvalue parameter ω. □

Lemma 8. *Under the setting of Theorem 3, suppose that* $\max_i |\gamma_i| \leq n_\mathrm{c}^{-2/3}$ *and* $\|\gamma\|_2^2 = \Omega_p(1/n_\mathrm{c})$. *Then, we obtain a central limit theorem*

$$\frac{1}{\|\gamma\|_2} \sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i(0) \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right). \tag{32}$$

Proof. The proof follows Lyapunov's method. Since the optimization program for γ did not consider the outcomes Y_i, unconfoundedness (Assumption 1) implies that $\varepsilon_i(0)$ is independent of γ_i conditionally on X_i, and so

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

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 (16). Thus,

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

because, by assumption, $\|\gamma\|_2^{-1} = \mathcal{O}_P\left(\sqrt{n_\mathrm{c}}\right)$. Thus Lyapunov's theorem implies that the central limit theorem (32). □

Proof of Corollary 4

The key idea in establishing this result is that we need to replace the "oracle" weights defined in (18) with

$$\gamma_i^{**} = \frac{1}{n_\mathrm{c}} m_t \Sigma_c^{-1} X_i. \tag{33}$$

Once we have verified that, with high probability, these candidate weights γ^{**} satisfy the constraint from (16), i.e., $\left\|\overline{X}_\mathrm{t} - \mathbf{X}_\mathrm{c}^\top \gamma^{**}\right\|_\infty \leq K\sqrt{\log(p)/n_\mathrm{c}}$, we can establish the result (22) by replicating the proof of Theorem 3 verbatim. Now, by Lemma 2, we know that with probability tending to 1,

$$\left\|m_t - \mathbf{X}_\mathrm{c}^\top \gamma^{**}\right\|_\infty \leq C\varsigma^2 \sqrt{V S \log(p)\,/\,n_\mathrm{c}}.$$

Meanwhile, a standard Hoeffding bound together with the fact that $n_\mathrm{t}/n_\mathrm{c} \to_p \rho$ establishes that, with probability tending to 1,

$$\left\|\overline{X}_\mathrm{t} - m_t\right\|_\infty \leq \nu\sqrt{2.1\rho}\sqrt{\log(p)\,/\,n_\mathrm{c}}.$$

Combining these two bounds yields the desired result.

Proof of Theorem 5

Our proof mirrors the one used for Theorem 3. We again start by proposing a class of candidate weights γ^* that satisfy the constraints (23); except, this time, we motivate our candidate weights using the overlap assumption:

$$\gamma_i^* = \frac{e\left(X_i\right)}{1 - e\left(X_i\right)} \bigg/ \sum_{\{i : W_i = 0\}} \frac{e\left(X_i\right)}{1 - e\left(X_i\right)}. \tag{34}$$

We start by characterizing the behavior of these weights below; we return to verify these bounds at the end of the proof. We also note that these weights also trivially satisfy $\gamma_i^* \leq n_{\mathrm{c}}^{-2/3}$ once n_{c} is large enough.

Lemma 9. *Under the conditions of Theorem 5, the weights γ^* defined in (34) satisfy the the following bounds with probability at least $1 - \delta$, for any $\delta > 0$:*

$$\left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma^*\right\|_\infty \leq \nu \sqrt{2 \log\left(\frac{10\, p}{\delta}\right) \left(\frac{1}{n_{\mathrm{t}}} + \frac{\left(1 - \eta\right)^2}{n_{\mathrm{c}}\, \eta^2}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right), \quad \textit{and} \tag{35}$$

$$\begin{aligned} n_{\mathrm{t}}\, \left\|\gamma^*\right\|_2^2 \leq \frac{1}{\rho_n^2}\, \mathbb{E}\left[\left(\frac{e(X_i)}{1 - e(X_i)}\right)^2 \middle| W_i = 0\right] \\ + \frac{\left(1 - \eta\right)^2 \left(2 - 2\eta + \rho_n \eta\right)}{\rho_n^3\, \eta^3} \sqrt{\frac{1}{2 n_{\mathrm{c}}} \log\left(\frac{10\, p}{\delta}\right)} + \mathcal{O}\left(\frac{1}{n_{\mathrm{c}}}\right), \end{aligned} \tag{36}$$

where $\rho_n = \mathbb{P}\left[W_i = 1\right] / \mathbb{P}\left[W_i = 0\right]$ is the odds ratio for the n-th problem.

Given these preliminaries, we can follow the proof of Theorem 3 closely. First, by the same argument as used to prove Lemma 7, we can verify 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,

$$\left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 \leq \frac{5\nu}{4}\, \frac{24\, \upsilon}{\omega}\, k \sqrt{\frac{\log p}{n_{\mathrm{c}}}}. \tag{37}$$

Thus, we again find that

$$\sqrt{n}\, \left\|\overline{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^\top \gamma\right\|_\infty \left\|\hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}}\right\|_1 = \mathcal{O}_P\left(\frac{k\, \log(p)}{\sqrt{n}}\right), \tag{38}$$

and so, thanks to our sparsity assumption, we can us Proposition 1 to show that the error in $\hat{\beta}_{\mathrm{c}}$ does not affect the asymptotic distribution of our estimator at the $\sqrt{n}$-scale; provided the problem (23) is feasible.

Next, thanks to Lemma 9, we know that the problem (23) is feasible with high probability. Moreover, because the weights γ obtained via (23) satisfy $\sum \gamma = 1$, we trivially find that $\|\gamma\|_2^2 \geq 1/n_{\mathrm{c}}$ and can apply Lemma 8 to get a central limit result for $\hat{\mu}_{\mathrm{c}} - \hat{\mu}_{\mathrm{c}}$. Finally, invoking (36) and the fact that $\|\gamma\|_2 \leq \|\gamma^*\|_2$ with probability tending to 1, we obtain the desired rate bound (26).

Proof of Lemma 9

To verify our desired result, 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 sub-Gaussianity of X_{ij} as in Assumption 7.

- $A_j := n_{\mathrm{c}}^{-1} \sum_{\{i : W_i = 0\}} X_{ij}\, e(X_i)/(1 - e(X_i))$ is sub-Gaussian with parameter $\nu^2 (1-\eta)^2 / (n_{\mathrm{c}}\, \eta^2)$ by sub-Gaussianity of X_{ij} 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 $\mathbb{E}\left[A_j\right] = 0$.
- $D := n_{\mathrm{c}}^{-1} \sum_{\{i : W_i = 0\}} e(X_i)/(1 - e(X_i)) - \rho_n$ is sub-Gaussian with parameter $(1-\eta)^2/(4 n_{\mathrm{c}}\, \eta^2)$, where $\rho_n = \mathbb{P}\left[W = 1\right] / \mathbb{P}\left[W = 0\right]$ 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 $1 - \delta$:

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

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

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

Moreover, noting that

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

we see that on ε_δ,

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

Thus, there exists a γ satisfying all desired constraints; thus, there must also be some $\zeta \in (0, 1)$ for which (16) yields such a solution.

Proof of Corollary 6

We prove the result in the setting of Theorem 5. First of all, we can use the argument of Theorem 5 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 (38) 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 (27), we only need to show that $\widehat{V}_c / V_c \to_p 1$; the desired conclusion then follows from Slutsky's theorem. Now, to verify this latter result, it suffices to check that

$$\frac{1}{V_c} \sum_{\{i : W_i = 0\}} \gamma_i^2 \left(Y_i - X_i \cdot \beta_{\mathrm{c}}\right)^2 \to_p 1, \text{ and} \tag{39}$$

$$\frac{1}{V_c} \sum_{\{i : W_i = 0\}} \gamma_i^2 \left(X_i \cdot \left(\beta_{\mathrm{c}} - \hat{\beta}_{\mathrm{c}}\right)\right)^2 \to_p 0. \tag{40}$$

To show the first convergence result, we can proceed as in the proof of Lemma 8 to verify that there is a universal constant C_4 for which

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

and so (39) holds by Markov's inequality. Meanwhile, to establish (40), 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

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

where the second step uses Hölder's inequality as in the proof of Proposition 1. Then, thanks to the assumed upper and lower bounds on the conditional variance of $\varepsilon_i(W_i)$ given X_i and W_i, we only need to check that

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

We can use sub-Gaussianity of X_i (Assumption 7) and the bound (37) on the L_1-error of $\hat{\beta}_{\mathrm{c}}$ to find a constant $C(\nu,\, \omega,\, \upsilon)$ for which

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

with probability tending to 1. Then, noting our sparsity condition on k (Assumption 4), we find that

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

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