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

Susan Athey† Guido W. Imbens‡ Stefan Wager§

Current version February 2018

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 in high-dimensional linear models. Given linearity, we do not need to assume that the treatment propensities are estimable, or that the average treatment effect is a sparse contrast of the outcome model parameters. Rather, 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 order to identify causal effects in observational studies, practitioners may assume treatment assignments to be as good as random (or unconfounded) conditional on observed features of the units; see Rosenbaum and Rubin (1983) and Imbens and Rubin (2015) for general discussions. Motivated by this setup, there is a large literature on how to adjust for differences in observed features between the treatment and control groups; 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 et al., 1994, 1995, 2017; Rosenbaum, 2002; Tan, 2010; Tsiatis, 2007; Van Der Laan and Rubin, 2006).

In practice, researchers sometimes need to account for a substantial number of features to make this assumption of unconfoundedness plausible. For example, in an observational study of the effect of flu vaccines on hospitalization, we may be concerned that only controlling for differences in the age and sex distribution between controls and treated may not be sufficient to eliminate biases. In contrast, controlling for detailed medical histories and personal characteristics may make unconfoundedness more

*We are grateful for detailed comments from Jelena Bradic, Edgar Dobriban, Bryan Graham, Chris Hansen, Nishanth Mundru, Jamie Robins and José Zubizarreta, and for discussions with seminar participants at the Atlantic Causal Inference Conference, Boston University, the Columbia Causal Inference Conference, Columbia University, Cowles Foundation, the Econometric Society Winter Meeting, the EGAP Standards Meeting, the European Meeting of Statisticians, ICML, INFORMS, Stanford University, UNC Chapel Hill, University of Southern California, and the World Statistics Congress.

†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 and of Statistics (by courtesy), Stanford Graduate School of Business, swager@stanford.edu.

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, 2017) 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. Belloni et al. (2014) propose an augmented variable selection scheme to avoid this effect, while Belloni et al. (2017), Chernozhukov et al. (2017), Farrell (2015), and Van der Laan and Rose (2011) build on the work of Robins et al. (1994, 1995) and discuss how a doubly robust approach to average treatment effect estimation in high dimensions can also be used to compensate for the bias of regularized regression adjustments. Despite the breadth of research on the topic, all the above papers rely crucially on the existence of a consistent estimator of the propensity score, i.e., the conditional probability of receiving treatment given the features, in order to yield $\sqrt{n}$-consistent estimates of the average treatment effect in high dimensions.[1]

In this paper, we show that in settings where we are willing to entertain a sparse, well-specified linear model on the outcomes, efficient inference of average treatment effects in high-dimensions is possible under more general assumptions than suggested by the literature discussed above. Given linearity assumptions, we show that it is not necessary to consistently estimate treatment propensities; rather, it is enough to rely on de-biasing techniques building on recent developments in the high-dimensional inference literature (Javanmard and Montanari, 2014, 2015; Van de Geer et al., 2014; Zhang and Zhang, 2014). In particular, in sparse linear models, we show that $\sqrt{n}$-consistent inference of average treatment effects is possible provided we simply 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. We do not need to assume the existence of a consistent estimator of the propensity scores, or any form of sparsity on the propensity model.

The starting point behind both our method and the doubly robust methods of Belloni et al. (2017), Chernozhukov et al. (2017), Farrell (2015), Van der Laan and Rose (2011), etc., is a recognition that high dimensional regression adjustments (such as the lasso) always shrink estimated effects, and that ignoring this shrinkage may result in prohibitively biased treatment effect estimates. The papers on doubly robust estimation then proceed to show that propensity-based adjustments can be used to compensate for this bias, provided we have a consistent propensity model that converges fast enough to the truth. Conceptually, this work builds on the result of Rosenbaum and Rubin (1983), who showed that controlling for the propensity score is sufficient to remove all biases associated with observed covariates, regardless of their functional form.

If we are willing to focus on high-dimensional linear models, however, it is possible to tighten the connection between the estimation strategy and the objective of estimating the average treatment effect and, in doing so, extend the number of settings where $\sqrt{n}$-consistent inference is possible. The key insight is that, in a linear model, propensity-based methods are attempting to solve a needlessly difficult task when they seek to eliminate biases of any functional form. Rather, in linear models, it is enough to correct for linear biases. In high dimensions, this can still be challenging; however, we find that it is possible to approximately correct for such biases whenever we assume overlap.

Concretely, 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 between the treatment and control groups. Here we follow Zubizarreta (2015), and optimize 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

[1] Some of the above methods assume that the propensity scores can be consistently estimated using a sparse logistic model, while others allow for the use of more flexible modeling strategies following, e.g., McCaffrey et al. (2004), Van der Laan et al. (2007), or Westreich et al. (2010).

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. Formally, this second step of re-weighting residuals using the weights proposed by Zubizarreta (2015) is closely related to de-biasing corrections studied in the high-dimensional regression literature (Javanmard and Montanari, 2014, 2015; Van de Geer et al., 2014; Zhang and Zhang, 2014); we comment further on this connection in Section 3.

This approach also bears a close conceptual connection to work by Chan et al. (2015), Deville and Särndal (1992), Graham et al. (2012, 2016), Hainmueller (2012), Hellerstein and Imbens (1999), Imai and Ratkovic (2014) and Zhao (2016), who fit propensity models to the data under a constraint that the resulting inverse-propensity weights exactly balance the covariate distributions between the treatment and control groups, and find that these methods out-perform propensity-based methods that do not impose balance. Such an approach, however, is only possible in low dimensions; in high dimensions where there are more covariates than samples, achieving exact balance is in general impossible. One of the key findings of this paper is that, in high dimensions, it is still often possible to achieve approximate balance under reasonable assumptions and that—when combined with a lasso regression adjustment—approximate balance suffices for eliminating bias due to regularization.

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. The finding that both weighting and regression adjustment are important is similar to conclusions drawn from the earlier literature on doubly robust estimation (e.g., Robins and Ritov, 1997), where combining both techniques was shown to extend the set of problems where efficient treatment effect estimation is possible. 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), Imai and Ratkovic (2014), and Zubizarreta (2015).

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. A software implementation for `R` is available at `https://github.com/swager/balanceHD`.

2 Estimating Average Treatment Effects in High Dimensions

2.1 Setting and Notation

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

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

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,

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

We note that the average treatment effect for the controls and the overall average effect can be handled similarly. 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)**.** $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i$.

Assumption 2 (Linearity)**.** The conditional response functions satisfy $\mu_{\mathrm{c}}(x) = \mathbb{E}\left[Y_i(0) \,\middle|\, X = x\right] = x \cdot \beta_{\mathrm{c}}$ and $\mu_{\mathrm{t}}(x) = \mathbb{E}\left[Y_i(1) \,\middle|\, X = x\right] = x \cdot \beta_{\mathrm{t}}$, for all $x \in \mathbb{R}^p$.

Here, 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 in high dimensions some strong structural assumption is in general needed for inference to be possible. Then, given linearity, we have

$$\tau = \mu_{\mathrm{t}} - \mu_{\mathrm{c}}, \quad \text{where} \quad \mu_{\mathrm{t}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{t}}, \quad \mu_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \beta_{\mathrm{c}}, \quad \text{and} \quad \overline{X}_{\mathrm{t}} = \frac{1}{n_{\mathrm{t}}} \sum_{i=1}^{n} \mathbf{1}\left(\{W_i = 1\}\right) X_i. \tag{3}$$

Estimating the first term is easy: $\hat{\mu}_{\mathrm{t}} = \overline{Y}_{\mathrm{t}} = \sum_{\{i:W_i=1\}} Y_i^{\mathrm{obs}} / n_{\mathrm{t}}$ is unbiased for μ_{t}. 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. Though neither approach alone performs well in a high-dimensional setting with a generic propensity score, we find 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 $\hat{\mu}_{\mathrm{c}}$ as a weighted average $\hat{\mu}_{\mathrm{c}} = \sum_{\{i:W_i=0\}} \gamma_i Y_i^{\mathrm{obs}}$. The standard way of selecting weights γ_i uses the propensity score: $\gamma_i = e(X_i)/(1-e(X_i)) \,/\, (\sum_{\{i:W_j=0\}} e(X_j)/(1-e(X_j)))$. To implement these methods researchers typically substitute an estimate of the propensity score into this expression. Such inverse-propensity weights with non-parametric propensity score estimates have desirable asymptotic properties in settings with a small number of covariates (Hirano et al., 2003). The finite-sample performance of methods based on inverse-propensity weighting can be poor, however, both in settings with limited overlap in covariate distributions and in settings with many covariates. A key difficulty is that estimating the treatment effect then involves dividing by $1-\hat{e}(X_i)$, and so 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.

As discussed in the introduction, if the control potential outcomes $Y_i(0)$ have a linear dependence on X_i, then using weights γ_i that explicitly seek to balance the features X_i is often advantageous (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 weighted averaging 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; however, directly using his approach does not allow for $\sqrt{n}$-consistent estimation in a regime where p is much larger than n.

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 $\overline{X}_{\mathrm{c}}$. As stressed by Belloni et al. (2014), 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 selected variables, is implicitly equivalent to using a weighted averaging estimator with weights γ chosen to balance the selected features (Robins et al., 2007). 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 pilot estimate of the treatment effect. In the second step, we do "approximate balancing" of the regression residuals to estimate treatment effects: 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 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):

$$\hat{\mu}_{\mathrm{c}} = \overline{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_{\mathrm{c}} \right). \tag{4}$$

In other words, we fit a model parametrized by β_{c} to capture some of the strong signals, and then use direct numerical re-balancing of the control data on the features to extract left-over signal from the residuals $Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_{\mathrm{c}}$. 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 (4) is that it yields a simple and powerful theoretical guarantee, as stated below. Recall that $\mathbf{X}_c$ is the feature matrix for the control units. Consider the difference between $\hat{\mu}_c$ and μ_c for our proposed approach: $\hat{\mu}_c - \mu_c = (\overline{X}_t - \mathbf{X}_c^\top \gamma) \cdot (\hat{\beta}_c - \beta_c) + \gamma \cdot \varepsilon$, where ε is the intrinsic noise $\varepsilon_i = Y_i(0) - X_i \cdot \beta_c$. With only the regression adjustment and no weighting, the difference would be $\hat{\mu}_{c,\text{reg}} - \mu_c = (\overline{X}_t - \overline{X}_c) \cdot (\hat{\beta}_c - \beta_c) + \mathbf{1} \cdot \varepsilon / n_c$, and with only the weighting the difference would be $\hat{\mu}_{c,\text{weight}} - \mu_c = (\overline{X}_t - \mathbf{X}_c^\top \gamma) \cdot \beta_c + \gamma \cdot \varepsilon$. Without any adjustment, just using the average outcome for the controls as an estimator for μ_c, the difference between the estimator for μ_c and its actual value would be $\hat{\mu}_{c,\text{no-adj}} - \mu_c = (\overline{X}_t - \overline{X}_c) \cdot \beta_c + \mathbf{1} \cdot \varepsilon / n_c$. The regression reduces the bias from $(\overline{X}_t - \overline{X}_c) \cdot \beta_c$ to $(\overline{X}_t - \overline{X}_c) \cdot (\hat{\beta}_c - \beta_c)$, which will be substantial reduction if the estimation error $(\hat{\beta}_c - \beta_c)$ is small relative to β_c. The weighting further reduces this to $(\overline{X}_t - \mathbf{X}_c^\top \gamma) \cdot (\hat{\beta}_c - \beta_c)$, which may be helpful if there is a substantial difference between $\overline{X}_t$ and $\overline{X}_c$. This result, formalized below, shows the complimentary nature of the regression adjustment and the weighting. All proofs are given in the appendix.

Proposition 1. *The estimator (4) satisfies* $|\hat{\mu}_c - \mu_c| \leq \left\|\overline{X}_t - \mathbf{X}_c^\top \gamma\right\|_\infty \left\|\hat{\beta}_c - \beta_c\right\|_1 + \left|\sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i\right|$.

This result decomposes the error of $\hat{\mu}_c$ into two parts. The first is a bias term reflecting the dimension p of the covariates; the second term is a variance term that does not depend on it. The upshot is that the bias term, which encodes the high-dimensional nature of the problem, involves a product of two factors that can both other be made reasonably small; we will focus on regimes where the first term should be expected to scale 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. If we are in a sparse enough regime (i.e., k is small enough), Proposition 1 implies that our procedure will be variance dominated; and, under these conditions, we also show that it is $\sqrt{n}$-consistent.

In order to exploit Proposition 1, we need to make concrete choices for the weights γ and the parameter estimates $\hat{\beta}_c$. First, just like Zubizarreta (2015), we choose our weights γ to directly optimize the bias and variance terms in Proposition 1; the functional form of γ is given in (5), where $\zeta \in (0, 1)$ is a tuning parameter. We refer to them as *approximately balancing weights* since they seek to make the mean of the re-weighted control sample, namely $\mathbf{X}_c^\top \gamma$, match the treated sample mean $\overline{X}_t$ as closely as possible. The positivity constraint on γ_i in (5) aims to prevent the method from extrapolating too aggressively, while the upper bound is added for technical reasons discussed in Section 3. Meanwhile, for estimating $\hat{\beta}_c$, we simply need to use an estimator that achieves good enough risk bounds under L_1-risk. In our analysis, we focus on the lasso (Chen et al., 1998; Tibshirani, 1996); however, in experiments, we use the elastic net for additional stability (Zou and Hastie, 2005). Our complete algorithm is described in Procedure 1.

Finally, although we do use this estimator in the present paper, we note that an analogous estimator for the average treatment effect $\mathbb{E}\left[Y(1) - Y(0)\right]$ can also be constructed:

$$\begin{aligned} \hat{\tau}_{ATE} &= \overline{X}\left(\hat{\beta}_t - \hat{\beta}_c\right) + \sum_{\{i:W_i=1\}} \gamma_{t,i}\left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_t\right) - \sum_{\{i:W_i=0\}} \gamma_{c,i}\left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_c\right), \text{ where} \\ \gamma_t &= \text{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta)\left\|\tilde{\gamma}\right\|_2^2 + \zeta\left\|\overline{X} - \mathbf{X}_t^\top \tilde{\gamma}\right\|_\infty^2 \text{ subject to } \sum_{\{i:W_i=1\}} \tilde{\gamma}_i = 1 \text{ and } 0 \leq \tilde{\gamma}_i \leq n_t^{-2/3} \right\}, \end{aligned} \tag{8}$$

and γ_c is constructed similarly. This method can be analyzed using the same tools developed in this paper, and is available in our software package `balanceHD`. The conditions required for $\sqrt{n}$-consistent inference of τ_{ATE} using (8) directly mirror the conditions (sparsity, overlap, etc.) listed in Section 3 for inference about the average treatment effect on the treated.

2.4 Connection to Doubly Robust Estimation

The idea of combining weighted and regression-based approaches to treatment effect estimation has a long history in the causal inference literature. Given estimated propensity scores $\hat{e}(X_i)$, Cassel et al. (1976)

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). The optimization problem (5) is a quadratic program, and so can be solved using off-the-shelf convex optimization software; we use the interior point solver `mosek` by default (MOSEK, 2015).

1. Compute positive approximately balancing weights γ as

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ (1-\zeta) \left\| \tilde{\gamma} \right\|_2^2 + \zeta \left\| \overline{X}_\text{t} - \mathbf{X}_\text{c}^\top \tilde{\gamma} \right\|_\infty^2 \text{ s. t. } \sum_{\{i:W_i=0\}} \tilde{\gamma}_i = 1 \text{ and } 0 \leq \tilde{\gamma}_i \leq n_\text{c}^{-2/3} \right\}. \quad (5)$$

2. Fit β_c in the linear model using a lasso or an elastic net,

$$\hat{\beta}_\text{c} = \text{argmin}_\beta \left\{ \sum_{\{i:W_i=0\}} \left(Y_i^\text{obs} - X_i \cdot \beta \right)^2 + \lambda \left((1-\alpha) \left\| \beta \right\|_2^2 + \alpha \left\| \beta \right\|_1 \right) \right\}. \quad (6)$$

3. Estimate the average treatment effect τ as

$$\hat{\tau} = \overline{Y}_\text{t} - \left(\overline{X}_\text{t} \cdot \hat{\beta}_\text{c} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^\text{obs} - X_i \cdot \hat{\beta}_\text{c} \right) \right). \quad (7)$$

and Robins et al. (1994) propose using an augmented inverse-propensity weighted (AIPW) estimator,

$$\hat{\mu}_\text{c}^{(AIPW)} = \overline{X}_\text{t} \cdot \hat{\beta}_\text{c} + \sum_{\{i:W_i=0\}} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \left(Y_i^\text{obs} - X_i \cdot \hat{\beta}_\text{c} \right) \Big/ \sum_{\{i:W_i=0\}} \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}; \quad (9)$$

the difference between this estimator and ours is that they obtain their weights γ_i for (4) via propensity-score modeling instead of quadratic programming. Estimators of this type have several desirable aspects: they are "doubly robust" in the sense that they are consistent whenever either the propensity fit $\hat{e}(\cdot)$ or the outcome fit $\hat{\beta}_\text{c}$ is consistent, and they are asymptotically efficient in a semiparametric specification (Hahn, 1998; Hirano et al., 2003; Robins and Rotnitzky, 1995; Tsiatis, 2007). However, a practical concern with this class of methods is that they may perform less well when $1 - \hat{e}(X_i)$ is close to 0 (Hirano et al., 2003; Kang and Schafer, 2007). Several higher-order refinements to the simple AIPW estimator (9) have also been considered in the literature. In particular, Kang and Schafer (2007) use the inverse-propensity weights $\hat{e}(X_i)/(1-\hat{e}(X_i))$ as sample weights when estimating $\hat{\beta}_\text{c}$, while Scharfstein et al. (1999) and Van Der Laan and Rubin (2006) consider adding these weights as features in the outcome model; see also Robins et al. (2007) and Tan (2010) for further discussion.

Belloni et al. (2017) and Farrell (2015) study the behavior of AIPW in high dimensions, and establish 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); doubly robust methods can then be used to relax the regularity conditions needed for efficiency (Farrell, 2015; Robins et al., 2017). Conversely, in high-dimensions, both weighting and regression adjustments are required for $\sqrt{n}$-consistency (Belloni et al., 2017; Farrell, 2015; Robins and

Ritov, 1997).

Although our estimator (4) is cosmetically quite closely related to the AIPW estimator (9), the motivation behind it is quite different. A common description of the AIPW estimator is that it tries to estimate two different nuisance components, i.e., the outcome model $\hat{\mu}_c$ and the propensity model $\hat{e}$; it then achieves consistency if either of these components is itself estimated consistently, and efficiency if both components are estimated at fast enough rates. In contrast, our approximate residual balancing estimator bets on linearity twice: once in fitting the outcome model via the lasso, and once in de-biasing the lasso via balancing weights (5).

By relying more heavily on linearity, we can considerably extend the set of problem under which $\sqrt{n}$-consistent is possible (assuming linearity in fact holds). As a concrete example, a simple analysis of AIPW estimation in high-dimensional linear models would start by assuming that the lasso is $o_P(n^{-1/4})$ consistent in root-mean squared error,[2] which can be attained via the lasso assuming a k-sparse true model with sparsity level $k \ll \sqrt{n}/\log(p)$; and this is, in fact, exactly the same condition we assume in Theorem 5. Then, in addition to this requirement on the outcome model, AIPW estimators still need to posit the existence of an $o_P(n^{-1/4})$ consistent estimator of the treatment propensities, whereas we do not need to assume anything about the treatment assignment mechanism beyond overlap. The reason for this phenomenon is that the task of balancing (which is all that is needed to correct for the bias of the lasso in a linear model) is different from the task of estimating the propensity score—and is in fact often substantially easier.[3]

2.5 Related Work

Our approximately balancing weights (5) are inspired by the recent litearture on balancing weights (Chan et al., 2015; Deville and Särndal, 1992; Graham et al., 2012, 2016; Hainmueller, 2012; Hellerstein and Imbens, 1999; Hirano et al., 2001; Imai and Ratkovic, 2014; Zhao, 2016). Most closely related, Zubizarreta (2015) proposes estimating τ using the re-weighting formula as in Section 2.2.1 with weights

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \left\|\overline{X}_t - \mathbf{X}_c^\top \tilde{\gamma}\right\|_\infty \leq t \right\}, \tag{10}$$

where the tuning parameter is t; he calls these weights *stable balancing weights*. These weights are of course equivalent to ours, the only difference being that Zubizarreta bounds imbalance in constraint form whereas we do so in Lagrange form. The main conceptual difference between our setting and that of Zubizarreta (2015) is that he considers problem settings where $p < n_c$, 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 (10) will not in general be feasible for small values of t; and in fact the bias term $\left\|\overline{X}_t - \mathbf{X}_c^\top \gamma\right\|_\infty$ becomes the dominant source of error in estimating τ. We call our weights γ "approximately" balancing in order to remind the reader of this fact. In settings where it is only possible to achieve approximate balance, weighting alone as considered in Zubizarreta (2015) will not yield a $\sqrt{n}$-consistent estimate of the average treatment effect, and it is necessary to use a regularized regression adjustment as in (4).

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,

[2] A more careful analysis of AIPW estimators can trade off the accuracy of the propensity and main effect models and, instead of requiring that both the propensity and outcome models can be estimated at $o_P(n^{-1/4})$ rates, only assumes that the product of the two rates be bounded as $o_P(n^{-1/2})$; see, e.g., Farrell (2015). In high dimensions, this amounts to assuming that the outcome and propensity models are both well specified and sparse, with respective sparsity levels k_β and k_e satisfying $k_\beta k_e \ll n/\log(p)^2$. AIPW can thus be preferable to ARB given sparse enough and well specified propensity models, with $k_e \ll \sqrt{n}/\log(p)$.

[3] The above distinctions are framed are in a situation where the statistician starts with a set of high-dimensional covariates, and needs to find a way to control for all of them at once. In this setting, linearity is a strong assumption, and so it is not surprising that making this assumption lets us considerably weaken requirements on other aspects of the problem. In other applications, however, the statistician may have started with low-dimensional data, but then created a high-dimensional design by listing series expansions of the original data, interactions, etc. In this setting, linearity is replaced with smoothness assumptions on the outcome model (since any smooth function can be well approximated using a large enough number of terms from an appropriately chosen series expansion). Here, variants of our procedure can be more directly compared with doubly robust methods and, in particular, the γ_i in fact converge to the oracle inverse-propensity weights $e(X_i)/(1-e(X_i))$; see Hirshberg and Wager (2017) and Wang and Zubizarreta (2017) for a discussion and further results.

Hainmueller (2012) uses $\sum_i \gamma_i \log(\gamma_i)$ instead of $\sum_i \gamma_i^2$, leading to

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ \sum_{\{i:W_i=0\}} \tilde{\gamma}_i \log(\tilde{\gamma}_i) \text{ subject to } \sum \tilde{\gamma}_i = 1,\ \tilde{\gamma}_i \geq 0,\ \left\| \overline{X}_{\text{t}} - \mathbf{X}_{\text{c}}^\top \tilde{\gamma} \right\|_\infty = 0 \right\}. \quad (11)$$

This estimator has attractive conceptual connections to logistic regression and maximum entropy estimation, and in a low dimensional setting where $W|X$ admits a well-specified logistic model the methods of Graham et al. (2012, 2016) and Hainmueller (2012) are doubly robust (Zhao and Percival, 2017); see also Hirano et al. (2001), Imbens et al. (1998), and Newey and Smith (2004). In terms of our immediate concerns, however, the variance of $\hat{\tau}$ depends on γ through $\|\gamma\|_2^2$ and not $\sum \gamma_i \log(\gamma_i)$, so our approximately balancing weights are more directly induced by our statistical objective than those defined in (11).

Finally, in this paper, we have emphasized an asymptotic analysis point of view, where we evaluate estimators via their large sample accuracy. From this perspective, our estimator—which combines weighting with a regression adjustment as in (4)—appears to largely dominate pure weighting estimators; in particular, in high dimensions, we achieve $\sqrt{n}$-consistency whereas pure weighting estimators do not. On the other hand, stressing practical concerns, Rubin (2008) strongly argues that "designed based" inference leads to more credible conclusions in applications by better approximating randomized experiments. In our context, design based inference amounts to using a pure weighting estimator of the form $\sum \gamma_i Y_i$ where the γ_i are chosen without looking at the Y_i. The methods considered by Chan et al. (2015), Graham et al. (2012), Hainmueller (2012), Zubizarreta (2015), etc., all fit within this design-based paradigm, whereas ours does not.

3 Asymptotics of Approximate Residual Balancing

3.1 Approximate Residual Balancing as Debiased Linear Estimation

As we have already emphasized, approximate residual balancing is a method that enables us to makes inferences about average treatment effects without needing to estimate treatment propensities as nuisance parameters; rather, we build on recent developments 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). Our main goal is to understand the asymptotics our estimates for $\mu_{\text{c}} = \overline{X}_{\text{t}} \cdot \beta_{\text{c}}$. In the interest of generality, however, we begin by considering a broader problem, namely that of estimating generic contrasts $\xi \cdot \beta_{\text{c}}$ in high-dimensional linear models. 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. (2017), Chernozhukov et al. (2017) or Farrell (2015)—our method does not require consistent estimability of the treatment propensity function $e(x)$.

The problem of estimating *sparse* linear contrasts $\xi \cdot \beta_{\text{c}}$ in high-dimensional regression problems has received considerable attention, including notable recent contributions by Javanmard and Montanari (2014, 2015), Van de Geer et al. (2014), and Zhang and Zhang (2014). These papers, however, exclusively consider the setting where ξ is a sparse vector, and, in particular, 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}. But, in our setting, the contrast vector $\overline{X}_{\text{t}}$ defining our estimand $\mu_{\text{c}} = \overline{X}_{\text{t}} \cdot \beta_{\text{c}}$ is random and thus generically dense; moreover, we are interested in applications where $m_{\text{t}} = \mathbb{E}[\overline{X}_{\text{t}}]$ itself may also be dense. Thus, a direct application of these method is not appropriate in our problem.[4]

An extension of this line of work to the problem of estimating dense, generic contrasts $\theta = \xi \cdot \beta_{\text{c}}$ turns out to be closely related to our approximate residual balancing method for treatment effect estimation.

[4] As a concrete example, Theorem 6 of Javanmard and Montanari (2014) shows that their debiased estimator $\hat{\beta}_{\text{c}}^{(\text{debiased})}$ satisfies $\sqrt{n_{\text{c}}}(\hat{\beta}_{\text{c}}^{(\text{debiased})} - \beta_{\text{c}}) = Z + \Delta$, where Z is a Gaussian random variable with desirable properties and $\|\Delta\|_\infty = o(1)$. If we simply consider sparse contrasts of $\hat{\beta}_{\text{c}}$, then this error term Δ is negligible; however, in our setting, we would have a prohibitively large error term $\overline{X}_{\text{t}} \cdot \Delta$ that may grow polynomially in p.

To make this connection explicit, define the following estimator:

$$\hat{\theta} = \xi \cdot \hat{\beta}_c + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{obs} - X_i \cdot \hat{\beta}_c\right), \text{ where} \tag{12}$$

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \text{ subject to } \left\|\xi - \mathbf{X}_c^\top \tilde{\gamma}\right\|_\infty \leq K\sqrt{\frac{\log(p)}{n_c}}, \max_i |\tilde{\gamma}_i| \leq n_c^{-2/3} \right\}, \tag{13}$$

$\hat{\beta}_c$ is a properly tuned sparse linear estimator, and K is a tuning parameter discussed below. If we set $\xi \leftarrow \overline{X}_t$, then this estimator is nothing but our treatment effect estimator from Procedure 1.[5] Conversely, in the classical parameter estimation setting with $\xi \leftarrow e_j$, the above procedure is algorithmically equivalent to the one proposed by Javanmard and Montanari (2014, 2015). Thus, the estimator (12) can be thought of as an adaptation of the method of Javanmard and Montanari (2014, 2015) that debiases $\hat{\beta}_c$ specifically along the direction of interest ξ.

We begin our analysis in Section 3.2 by considering a general version of (12) under fairly strong "transformed independence design" generative assumptions on $\mathbf{X}_c$. Although these assumptions may be too strong to be palatable in practical data analysis, this result lets us make a crisp conceptual link between approximate residual balancing and the debiased lasso. In particular, we find that (Theorem 3), $\hat{\theta}$ from (12) is $\sqrt{n}$-consistent for θ provided $\xi^\top \Sigma_c^{-1} \xi = \mathcal{O}(1)$, where Σ_c is the covariance of $\mathbf{X}_c$. Interestingly, if $\Sigma_c = I_{p\times p}$, then in general $\xi^\top \Sigma_c^{-1} \xi = \|\xi\|_2^2 = \mathcal{O}(1)$ if and only if ξ is very sparse, and so the classical de-biased lasso theory reviewed above is essentially sharp despite only considering the sparse-ξ case (see also Cai and Guo, 2015). On the other hand, 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 $\|\xi\|_2 \gg 1$, provided that ξ is aligned with the large latent components of Σ_c. We also note that, in the application to treatment effect estimation, $\overline{X}_t^\top \Sigma_c^{-1} \overline{X}_t$ will in general be much larger than 1; however, in Corollary 4 we show how to overcome this issue.

To our knowledge, this was the first result for $\sqrt{n}$-consistent inference about dense contrasts of β_c at the time we first circulated our manuscript. We note, however, simultaneous and independent work by Zhu and Bradic (2016), who developed a promising method for testing hypotheses of the form $\xi \cdot \beta_c = 0$ for potentially dense vectors ξ; their approach uses an orthogonal moments construction that relies on regressing $\xi \cdot X_i$ against a $p-1$ dimensional design that captures the components of X_i orthogonal to ξ.

Finally, in Section 3.3, we revisit the specific problem of high-dimensional treatment effect estimation via approximate residual balancing under substantially weaker assumptions on the design matrix $\mathbf{X}_c$: Rather than assuming a generative "transformed independence design" model, we simply require overlap and standard regularity conditions. The cost of relaxing our assumptions on $\mathbf{X}_c$ is that we now get slightly looser performance guarantees; however, our asymptotic error rates are still in line with those we could get from doubly robust methods. We also discuss practical, heteroskedasticity-robust confidence intervals for τ. Through our analysis, we assume that $\hat{\beta}_c$ is obtained via the lasso; however, we could just as well consider, e.g., the square-root lasso (Belloni et al., 2011), sorted L_1-penalized regression (Bogdan et al., 2015; Su and Candes, 2016), or other methods with comparable L_1-risk bounds.

3.2 Debiasing Dense Contrasts

As we begin our analysis of $\hat{\theta}$ defined in (12), it is first important to note that the optimization program (13) is not always feasible. For example suppose that $p = 2n_c$, that $\mathbf{X}_c = (I_{n_c\times n_c} \; I_{n_c\times n_c})$, and that ξ consists of n times "1" followed by n times "-1"; then $\left\|\xi - \mathbf{X}_c^\top \gamma\right\|_\infty \geq 1$ for any $\gamma \in \mathbb{R}^{n_c}$, 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 (13) 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 $\Sigma_c = \text{Var}\left[X_i \,\middle|\, W_i = 0\right]$ denotes the population variance of control features. We also rely on the following

[5] Here, we phrased the imbalance constraint in constraint form rather than in Lagrange form; the reason for this is that, although there is a 1:1 mapping between these two settings, we found the former easier to work with formally whereas the latter appears to yield more consistent numerical performance. We also dropped the constraints $\sum \gamma_i = 1$ and $\gamma_i \geq 0$ for now, but will revisit them in Section 3.3.

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[6] $\mathbf{X}_c = Q\,\Sigma_c^{\frac{1}{2}}$, where $\mathbb{E}[Q_{ij}] = 0$, $\mathrm{Var}[Q_{ij}] = 1$, 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 $j = 1, ..., p$.

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 (13) is feasible with probability tending to 1; and, in particular, the constraints are satisfied by $\gamma_i^* = \frac{1}{n_c}\xi^\top \Sigma_c^{-1} X_i$.*

The above lemma is the key to our analysis of approximate residual balancing. Because, with high probability, the weights γ^* from Lemma 2 provide one feasible solution to the constraint in (13), we conclude that, again with high probability, the actual weights we use for approximate residual balancing must satisfy $\|\gamma\|_2^2 \leq \|\gamma^*\|_2^2 \approx n_c^{-1}\xi^\top \Sigma_c^{-1}\xi$. 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., $\|\beta_c\|_0 \leq k$, and that $k\,\log(p)/\sqrt{n} \to 0$.[7]

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 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.[8]

Next, our analysis builds on well-known bounds on the estimation error of the lasso (Bickel et al., 2009; Hastie et al., 2015) that require $\mathbf{X}_c$ to satisfy a form of the restricted eigenvalue condition. 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

$$\mathcal{C}_k(L) = \left\{ \beta \in \mathbb{R}^p : \|\beta\|_1 \leq L \sum_{j=1}^{k} \left|\beta_{i_j}\right| \quad \text{for some} \quad 1 \leq i_1 < ... < i_j \leq p \right\}. \tag{14}$$

Then, $\Sigma_c^{1/2}$ satisfies the $\{k, \omega, L\}$-restricted eigenvalue condition if $\beta^\top \Sigma_c \beta \geq \omega \|\beta\|_2^2$ for all $\beta \in \mathcal{C}_k(L)$.

[6] 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.

[7] 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 $\|\hat{\beta}_c - \beta_c\|_1$ 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.

[8] There are, of course, some 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; Wager et al. (2016) show that lasso regression adjustments allow for efficient average treatment effect estimation in randomized trials even when $k \ll n\,/\,\log(p)$; while the method of Zhu and Bradic (2016) for estimating dense contrasts $\xi \cdot \beta_c$ does not rely on sparsity of β_c, and instead places assumptions on the joint distribution of $\xi \cdot X_i$ and the individual regressors. The point in common between these results is that they let us weaken the sparsity requirements at the expense of strengthening our assumptions about the X-distribution.

Theorem 3. *Under the conditions of Lemma 2, 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[9] $\|\xi\|_\infty \geq \kappa > 0$. Suppose, moreover, that we have homoskedastic noise: $\text{Var}[\varepsilon_i(0) \,|\, X_i] = \sigma^2$ for all $i = 1, ..., n$, and also that the response noise $\varepsilon_i(0) := Y_i(0) - \mathbb{E}[Y_i(0) \,|\, X_i]$ is uniformly sub-Gaussian with parameter $v^2 S > 0$. Finally, suppose that we estimate $\hat{\theta}$ using (12), 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,*

$$\left(\hat{\theta} - \theta\right) / \|\gamma\|_2 \Rightarrow \mathcal{N}\left(0, \sigma^2\right), \quad n_c \|\gamma\|_2^2 / \xi^\top \Sigma_c^{-1} \xi \leq 1 + o_p(1). \tag{15}$$

The statement of Theorem 3 highlights a connection between our debiased estimator (12), 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

$$\sqrt{n_c}\left(\hat{\theta}^{(OLS)} - \theta\right) / \sqrt{\xi^\top \Sigma_c^{-1} \xi} \Rightarrow \mathcal{N}\left(0, \sigma^2\right), \quad \text{and } \sqrt{n_c}\left(\hat{\theta}^{(OLS)} - \theta - \sum_{\{i:W_i=0\}} \gamma_i^* \varepsilon_i(0)\right) \to_p 0, \tag{16}$$

where γ_i^* is as defined in Lemma 2. 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 (13) were explicitly chosen to minimize the variance of $\hat{\theta}$ subject to the estimator being nearly unbiased.

A downside of the above result is that our main goal is to estimate $\mu_c = \overline{X}_t \cdot \beta_c$, and this contrast-defining vector $\overline{X}_t$ fails to satisfy the bound on $\overline{X}_t^\top \Sigma^{-1} \overline{X}_t$ assumed in Theorem 3. In fact, because $\overline{X}_t$ is random, this quantity will in general be on the order of p/n. In the result below, we show how to get around this problem under the weaker assumption that $m_t^\top \Sigma_c^{-1} m_t$ is bounded; at a high level, the proof shows that the the stochasticity $\overline{X}_t$ does not invalidate our previous result. We note that, because $\overline{Y}_t$ is uncorrelated with $\hat{\mu}_c$ conditionally on $\overline{X}_t$, the following result also immediately implies a central limit theorem for $\hat{\tau} = \overline{Y}_t - \hat{\mu}_c$ where $\overline{Y}_t$ is the average of the treated outcomes.

Corollary 4. *Under the conditions of Theorem 3, suppose that we want to estimate $\mu_c = \overline{X}_t \cdot \beta_c$ by replacing ξ with $\overline{X}_t$ in (12), 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 $(X_i - m_t)_j \,|\, W_i = 1$ is sub-Gaussian with parameter $\nu^2 > 0$, and that the overall odds of receiving treatment $\mathbb{P}[W=1]/\mathbb{P}[W=0]$ tend to a limit ρ bounded away from 0 and infinity. Then, setting the tuning parameter in (13) as $K = C\varsigma^2\sqrt{VS} + \nu\sqrt{2.1\,\rho}$, we get*

$$(\hat{\mu}_c - \mu_c) / \|\gamma\|_2 \Rightarrow \mathcal{N}\left(0, \sigma^2\right), \quad n_c \|\gamma\|_2^2 / m_t^\top \Sigma_c^{-1} m_t \leq 1 + o_p(1). \tag{17}$$

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 comparable sparsity assumptions on the Y-model as used by the rest of the high-dimensional inference literature, including Belloni et al. (2014, 2017), Chernozhukov et al. (2017) 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.

[9]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 $j = 1, ..., p$ with $p \gg n_c$. Then, with high probability, the optimization problem (13) will yield $\gamma = 0$. This leaves us with a simple lasso estimator $\hat{\theta} = \xi \cdot \hat{\beta}_c$ whose risk scales as $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathcal{O}(k^2 \log(p)/(pn_c)) \ll 1/n_c$. The problem with this superefficient estimator is that it is not necessarily asymptotically Gaussian.

3.3 A Robust 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—in particular the transformed independence design assumption—were stronger than ones we may feel comfortable making in applications.

In this section, we propose an alternative analysis of approximate residual balancing based on 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. If we are willing to assume overlap, we can relax the transposed independence design assumption into much more routine regularity conditions on the design, as in Assumption 7.

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

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 $\mathbf{X}_c$ satisfies the $\{k, \omega, 4\}$-restricted eigenvalue condition as defined in Assumption 5, with probability tending to 1.

Following Lemma 2, our analysis again proceeds by guessing a feasible solution to our optimization problem, and then using it to bound the variance of our estimator. Here, however, we use 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 the optimization problem are at least as good as these inverse-propensity weights, and thus our method will be at most as variable as one that uses augmented inverse-propensity weighting (9) with these oracle propensity weights.

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 (4) with weights (5), except we replace the Lagrange-form penalty on the imbalance with a hard constraint $\|\overline{X}_t - \mathbf{X}_c^\top \tilde{\gamma}\|_\infty \leq K\sqrt{\log(p)/n_c}$, with $K = \nu\sqrt{2.1(\rho + (\eta^{-1}-1)^2}$. Moreover, we fit the outcome model using a lasso with penalty parameter set to $\lambda_n = 5\nu\upsilon\sqrt{\log(p)/n_c}$. Then,*

$$\frac{\hat{\mu}_c - \mu_c}{\|\gamma\|_2} \Rightarrow \mathcal{N}(0, \sigma^2) \quad \text{and} \quad \frac{\hat{\tau} - \tau}{\sqrt{n_t^{-1} + \|\gamma\|_2^2}} \Rightarrow \mathcal{N}(0, \sigma^2), \tag{18}$$

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

$$\limsup_{n\to\infty} \; n_c \, \|\gamma\|_2^2 \;\leq\; \rho^{-2}\, \mathbb{E}\left[\left(\frac{e(X_i)}{1-e(X_i)}\right)^2 \middle| W_i = 0\right]. \tag{19}$$

The rate of convergence guaranteed by (19) is the same as what we would get if we actually knew the true propensities and could use them for weighting (Robins et al., 1994, 1995). 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 (17) is stronger than (19); however, there exist designs where the bounds match (Wang and Zubizarreta, 2017).

Finally, in applications, it is often of interest to have confidence intervals for μ_c and τ rather than just point estimates; below, we propose such a construction. Much like the sandwich variance estimates for ordinary least squares regression, our proposed confidence intervals are heteroskedasticity robust even though the underlying point estimates were motivated using an argument written in terms of a homoskedastic sampling distribution.

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

$$(\hat{\mu}_c - \mu_c) \Big/ \sqrt{\widehat{V}_c} \Rightarrow \mathcal{N}(0, 1), \quad \widehat{V}_c = \sum_{\{i:W_i=0\}} \gamma_i^2 \left(Y_i - X_i \cdot \hat{\beta}_c\right)^2. \tag{20}$$

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

$$(\hat{\mu}_t - \mu_t) \Big/ \sqrt{\widehat{V}_t} \Rightarrow (0, 1), \quad \widehat{V}_t = \frac{1}{n_t^2} \sum_{\{i:W_i=1\}} \left(Y_i - X_i \hat{\beta}_t\right)^2, \tag{21}$$

where $\hat{\beta}_t$ is obtained using the lasso with $\lambda_n = 5\nu v \sqrt{\log(p)/n_c}$. Moreover, $\hat{\mu}_c$ and $\hat{\mu}_t$ are independent conditionally on X and W, thus implying that $(\hat{\tau} - \tau) / (\widehat{V}_c + \widehat{V}_t)^{1/2} \Rightarrow \mathcal{N}(0, 1)$. 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 participants 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.[10]

[10] 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 gold 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 $\mathbb{E}[\tau]$, 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.

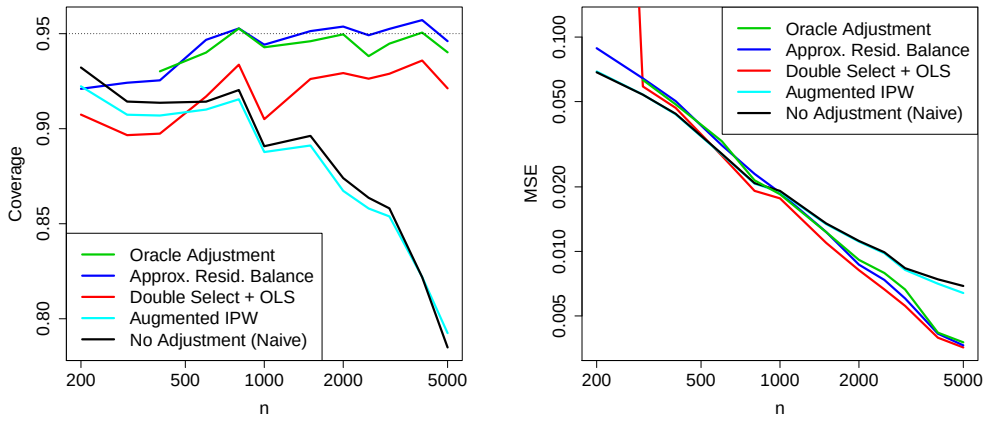


Figure 1: Finite sample performance of the average treatment effect on the treated for different estimators, aggregated over 1,000 replications. The target coverage rate, 0.95, is denoted with a dotted line.

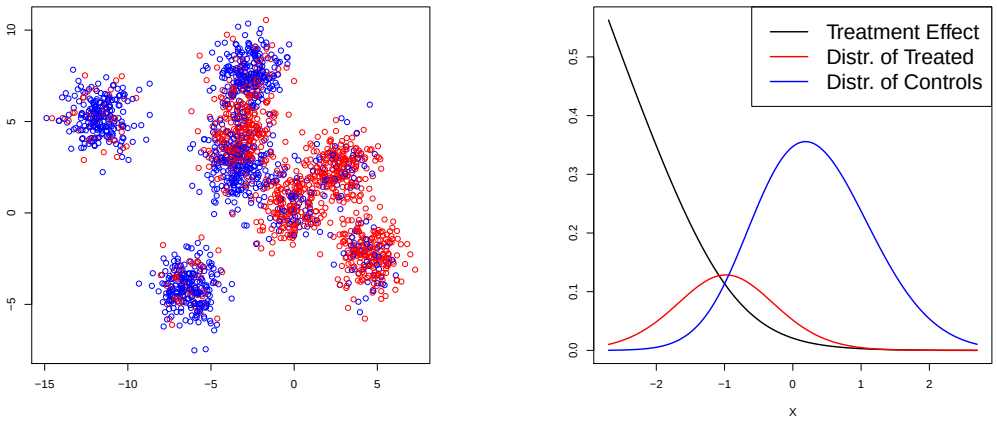
We compare the behavior of different methods for estimating the average treatment effect on the treated using randomly drawn subsamples of the original data (the full dataset has $n = 19,170$). In addition to approximate residual balancing, we consider augmented inverse-propensity weighting (9) and double selection following Belloni et al. (2014) as our baselines. 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 is not always well defined because some samples may result in counties where either everyone or no one is treated.

Figure 1 compares the performance of the different methods. We see that approximate residual balancing and double selection both do well in terms of mean-squared error. Moreover, confidence intervals built via approximate residual balancing achieve effectively nominal coverage; double selection also gets reasonable coverage and improves with n. In contrast, augmented inverse-propensity weighting does not perform well here. The problem appears to be that estimating treatment propensities is quite difficult, and a cross-validated logistic elastic net often learns an effectively constant propensity model.

5 Simulation Experiments

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** difference-in-means estimation $\hat{\tau} = \overline{Y}_{\text{t}} - \overline{Y}_{\text{c}}$ that ignores the covariate information X; the **elastic net** (Zou and Hastie, 2005), or equivalently, Procedure 1 with trivial weights $\gamma_i = 1/n_{\text{c}}$; **approximate balancing**, or equivalently, Procedure 1 with trivial parameter estimates $\hat{\beta}_{\text{c}} = 0$ (Zubizarreta, 2015); **inverse-propensity weighting**, as discussed in Section 2.2.1, with propensity estimates $\hat{e}(X_i)$ obtained by elastic net logistic regression, with the propensity scores trimmed at 0.05 and 0.95; **augmented inverse-propensity weighting**, which pairs elastic net regression adjustments with the above inverse-propensity weights (9); the **weighted elastic net**, motivated by Kang and Schafer (2007), that uses inverse-propensity weights as sample weights for the elastic net regression; **targeted maximum likelihood estimation** (TMLE), which fine-tunes the elastic net regression estimates along the direction specified by the inverse-propensity weights (Van Der Laan and Rubin, 2006), 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.



(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.

Unless otherwise specified, all outcome and propensity models were fit using a (linear or logistic) elastic net. 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 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 five different simulation settings. Our first setting is a **two-cluster** layout, with data drawn as $Y_i = (C_i + Z_i) \cdot \beta + W_i + \varepsilon_i$. Here, $W_i = \text{Bernoulli}(0.5)$, $Z_i \sim \mathcal{N}(0,\, I_{p\times p})$, $\varepsilon_i \sim \mathcal{N}(0,\, 1)$, and $C_i \in \mathbb{R}^p$ is a cluster center that is one of $C_i \in \{0,\, \delta\}$, such that $\mathbb{P}\left[C_i = 0 \,\middle|\, W_i = 0\right] = 0.8$ and $\mathbb{P}\left[C_i = 0 \,\middle|\, W_i = 1\right] = 0.2$. We consider two settings for the between-cluster vector δ: a "dense" setting where $\delta = 4/\sqrt{n}\ \mathbf{1}$, and a "sparse" setting where $\delta_j = 40/\sqrt{n}\ 1\left(\{j = 1 \text{ modulo } 10\}\right)$. 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 $W_i = 1$ wit 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 illustrate this simulation concept in Figure 2a. In both cases, we chose β as one of

$$\begin{aligned}&\text{dense} : \beta \propto (1,\, 1/\sqrt{2},\, ...,\, 1/\sqrt{p}), \quad \text{harmonic} : \beta \propto (1/10, 1/11, ..., 1/\left(p+9\right)),\\ &\text{moderately sparse} : \beta \propto (\underbrace{10,\, ...,\, 10}_{10},\, \underbrace{1,\, ...,\, 1}_{90},\, \underbrace{0,\, ...,\, 0}_{p-100}), \text{ and very sparse} : \beta \propto (\underbrace{1,\, ...,\, 1}_{10},\, \underbrace{0,\, ...,\, 0}_{p-10}).\end{aligned} \tag{22}$$

The signal strength was scaled such that $\|\beta\|_2 = 2$ in the two-cluster layout and $\|\beta\|_2 = 3$ in the many-cluster layout.

Our next two simulations are built using more traditional structural models. We first consider a **sparse two-stage** setting closely inspired by an experiment of Belloni et al. (2014). Here $X_i \sim \mathcal{N}(0,\, \Sigma)$ with

$\Sigma_{ij} = \rho^{|i-j|}$, and $\theta_i = X_i \cdot \beta_W + \varepsilon_{i1}$. Then, $W_i \sim \text{Bernoulli}(1/(1+e^{\theta_i}))$, and finally $Y_i = X_i \cdot \beta_Y + 0.5\, W_i + \varepsilon_{i2}$ where ε_{i1} and ε_{i2} are independent standard Gaussian. Following Belloni et al. (2014), we set the structure model as $(\beta_Y)_j \propto 1/j^2$ for $j = 1, ..., p$; for the propensity model, we consider both a "very sparse" propensity model $(\beta_W)_j \propto 1/j^2$, and also a "dense" propensity model $(\beta_W)_j \propto 1/\sqrt{j}$. A potential criticism of this simulation design is that the signal is perhaps unusually sparse (in the 4-th column of Table 3, adjusting for differences in the two most important covariates removes 93% of the bias associated with all the covariates); moreover, we note that all the important coefficients of both β_Y and β_W are close to each other in terms of their indices; thus, the effect of using a correlated design may be mitigated. Thus, we also ran a **moderately sparse two-stage** simulation, just like the above one, except we now used choices for β_Y as in (22), the only difference being that we shifted the indices of the betas, multiplying them by 23 mod p (e.g., the harmonic setup now has $(\beta)_j \propto 1/[10+(23\,(j-1) \bmod p)]$). Here, we drew the treatment assignments from a well-specified logistic model, $W_i \sim \text{Bernoulli}(1/(1+\exp(-\sum_{j=1}^{100} X_{ij}/40)))$.

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.

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 relative terms, this appears to hurt the double-selection method most. Meanwhile, in Table 3, 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 problem somewhat more difficult (Table 4), its performance decays substantially, and double selection lags both approximate residual balancing and propensity-based methods in its performance.

Generally, we find that 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. In these experiments, the weighted elastic net and TMLE also somewhat improve over AIPW.

Encouragingly, approximate residual balancing also does a good job in the misspecified setting from Table 5. It appears that our stipulation that the approximately balancing weights (5) must be non-negative (i.e., $\gamma_i \geq 0$) helps prevent our method from extrapolating 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. Perhaps the reason AIPW and TMLE do not do as well here is that there are very strong linear effects.

We evaluate coverage of confidence intervals in the "many-cluster" setting for different choices of β, n, and p; results are given in Table 6. 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.

6 Discussion

In this paper, we introduced approximate residual balancing as a method for unconfounded average treatment effect estimation in high-dimensional linear models. Under standard assumptions from the high-dimensional inference literature, our method allows for $\sqrt{n}$-consistent inference of the average treatment effect without any structural assumptions on the treatment assignment mechanism beyond overlap.

Widely used doubly robust methods, pioneered by Robins et al. (1994) and studied further by several authors (e.g., Belloni et al., 2017; Farrell, 2015; Kang and Schafer, 2007; Scharfstein et al., 1999; Robins

et al., 2007; Tan, 2010; Van Der Laan and Rubin, 2006), approach this problem by trying to estimate two different nuisance components, the outcome model and the propensity model. These methods then achieve consistency if either nuisance component is itself consistently estimated, and achieve semiparametric efficiency if both components as estimated fast enough. In contrast, our method "bets" on linearity twice, both in fitting the lasso and in attempting to balance away its bias. In well specified linear models, this bet allows us to considerably extend the class of problems for with $\sqrt{n}$-consistent inference of average treatment effects is possible; thus, if a practitioner believes linearity to be a reasonable assumption in a given problem, our estimator may be a promising choice.

We end by mentioning two important questions left open by this paper. First, it would be important to develop a better understanding of how to choose the tuning parameter ζ in (5) that trades off bias and variance in our balancing. Results from Theorems 3 and 5 provide some guidance on choosing ζ (via a constraint-form characterization); however, in our experiments, we achieved good performance by simply setting $\zeta = 1/2$ everywhere. The difficulty in choosing ζ is that we are trying to trade off an observable quantity (sampling variance) against an unobservable one (residual bias), and so cannot rely on simple methods like cross-validation that require unbiased estimates of the loss criterion we are trying to minimize. It would be of considerable interest to either devise a data-adaptive choice for ζ, or understand why a fixed choice $\zeta = 1/2$ appears to achieve systematically good performance.

It would also be interesting to extend our approach to generalized linear models, where there is a non-linear link function ψ for which $\mathbb{E}[Y_i(c) \,\big|\, X_i = x] = \psi(x \cdot \beta_c)$. In causal inference applications, this setting frequently arises when the outcomes Y_i^{obs} are binary, and we are willing to work with a logistic regression model. In this case, the first-order error component from using a pilot estimator $\hat{\beta}_c$ for estimating μ_c with a plug-in estimator $n_c^{-1} \sum_{\{W_i=1\}} \psi(X_i \cdot \hat{\beta}_c)$ would be of the form $n_c^{-1} \sum_{\{W_i=1\}} \psi'(X_i \cdot \hat{\beta}_c) X_i (\beta_c - \hat{\beta}_c)$. An analogue to Proposition 1 then suggests using an estimator

$$
\begin{aligned}
\hat{\mu}_c &= \frac{1}{n_c} \sum_{\{W_i=1\}} \psi\left(X_i \cdot \hat{\beta}_c\right) + \sum_{\{W_i=0\}} \gamma_i \left(Y_i^{\text{obs}} - \psi\left(X_i \cdot \hat{\beta}_c\right)\right), \\
\gamma &= \text{argmin}_{\tilde{\gamma}} \left\{ \zeta \left\| \frac{1}{n_t} \sum_{\{W_i=1\}} \psi'\left(X_i \cdot \hat{\beta}_c\right) X_i - \sum_{\{W_i=0\}} \tilde{\gamma}_i \, \psi'\left(X_i \cdot \hat{\beta}_c\right) X_i \right\|_\infty^2 \right. \\
&\qquad \left. + (1-\zeta) \sum_{\{W_i=0\}} \tilde{\gamma}_i^2 \, \psi'\left(X_i \cdot \hat{\beta}_c\right) \text{ subject to } \sum_{\{W_i=0\}} \tilde{\gamma}_i = 1, \quad 0 \le \tilde{\gamma}_i \le n_c^{-2/3} \right\}.
\end{aligned}
\tag{23}
$$

However, due to space constraints, we leave a study of this estimator to further work.

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 and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

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

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE: Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015.

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

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, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

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.

D. A. Hirshberg and S. Wager. Balancing out regression error: Efficient treatment effect estimation without smooth propensities. *arXiv preprint arXiv:1712.00038*, 2017.

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.

MOSEK. *MOSEK Rmosek Package*, 2015. URL http://docs.mosek.com/8.0/rmosek.pdf.

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.

J. Robins, L. Li, R. Mukherjee, E. Tchetgen Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *Annals of Statistics, forthcoming*, 2017.

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. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when inverse probability weights are highly variable. 22(4):544–559, 2007.

J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

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.

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

D. B. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840, 2008.

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.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

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

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

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

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science & Business Media, 2007.

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.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, (45):12673–12678, 2016.

Y. Wang and J. R. Zubizarreta. Approximate balancing weights: Characterizations from a shrinkage estimation perspective. *arXiv preprint arXiv:1705.00998*, 2017.

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.

Q. Zhao and D. Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

Y. Zhu and J. Bradic. Linear hypothesis testing in dense high-dimensional linear models. *arXiv preprint arXiv:1610.02987*, 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.

Beta Model	dense		harmonic		moderately sparse		very sparse	
Propensity Model	dense	sparse	dense	sparse	dense	sparse	dense	sparse
Naive	6.625	7.119	3.557	3.924	1.257	1.256	0.711	0.722
Elastic Net	4.328	1.058	2.190	0.665	0.716	0.350	0.237	0.204
Approximate Balance	**3.960**	1.179	2.130	0.686	0.789	0.362	0.464	0.316
Approx. Residual Balance	**3.832**	**0.423**	**1.854**	**0.320**	**0.495**	**0.213**	**0.185**	**0.165**
Inv. Propensity Weight	5.341	3.094	2.866	1.707	1.026	0.596	0.586	0.398
Augmented IPW	4.082	0.618	2.031	0.415	0.607	0.242	0.209	**0.166**
Weighted Elastic Net	4.086	0.562	1.984	0.385	0.575	0.232	0.207	**0.171**
TMLE Elastic Net	**3.811**	0.591	**1.843**	0.399	**0.495**	0.239	**0.192**	**0.165**
Double-Select + OLS	6.625	0.620	3.540	0.430	0.525	0.233	0.254	**0.165**

Table 1: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the two-cluster setting. We used $n = 500$, $p = 2000$, and scaled the signal such that $\|\beta\|_2 = 2$. All numbers are averaged over 400 simulation replications.

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	4.921	3.283	5.100	3.231	4.776	3.270	5.078	3.348
Elastic Net	2.527	1.353	1.618	**0.869**	0.727	**0.385**	0.168	0.108
Approximate Balance	2.575	1.324	2.505	1.379	2.567	1.248	2.434	1.396
Approx. Residual Balance	**2.123**	**1.172**	**1.528**	**0.866**	**0.653**	**0.383**	**0.158**	**0.105**
Inv. Propensity Weight	2.626	1.983	2.625	1.943	2.586	1.896	2.568	2.000
Augmented IPW	**2.102**	1.233	**1.515**	**0.852**	**0.656**	**0.376**	**0.154**	**0.103**
Weighted Elastic Net	**2.061**	**1.176**	1.576	**0.862**	0.727	**0.388**	0.194	**0.108**
TMLE Elastic Net	**2.095**	**1.208**	**1.500**	**0.847**	**0.656**	**0.375**	0.163	**0.103**
Double-Select + OLS	2.726	1.526	3.364	1.840	2.201	1.092	0.211	0.114

Table 2: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the many-cluster setting. We used $n = 800$, $p = 4000$, and scaled the signal such that $\|\beta\|_2 = 3$. All numbers are averaged over 400 simulation replications.

Propensity Model	sparse				dense			
First Stage Sig. Strength	$\|\beta_W\|_2 = 1$		$\|\beta_W\|_2 = 4$		$\|\beta_W\|_2 = 1$		$\|\beta_W\|_2 = 4$	
Structure Sig. Strength ($\|\beta_Y\|_2$)	1	4	1	4	1	4	1	4
Naive	0.963	3.796	1.701	6.804	0.535	2.129	0.784	3.130
Elastic Net	0.277	0.246	0.648	0.619	0.202	0.279	0.307	0.433
Approximate Balance	0.195	0.662	0.585	2.313	0.198	0.731	0.260	0.987
Approx. Residual Balance	0.109	0.102	0.287	0.326	0.107	0.138	**0.134**	**0.192**
Inv. Propensity Weight	0.484	1.876	0.932	3.722	0.301	1.191	0.421	1.651
Augmented IPW	0.164	0.151	0.374	0.384	0.130	0.188	0.181	0.258
Weighted Elastic Net	0.174	0.163	0.686	0.708	0.132	0.193	0.201	0.300
TMLE Elastic Net	0.161	0.149	0.234	0.227	0.122	0.175	0.355	0.389
Double-Select + OLS	**0.081**	**0.077**	**0.115**	**0.123**	**0.092**	**0.093**	0.190	**0.194**

Table 3: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the sparse two-stage setting. We used $n = 1000$, $p = 2000$, and $\rho = 0.5$. All numbers are averaged over 400 simulation replications.

Beta Model	dense		harmonic		moderately sparse		very sparse	
Autocovariance (ρ)	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9
Naive	1.236	2.659	1.088	1.938	0.951	1.096	0.814	0.814
Elastic Net	1.075	1.235	**0.631**	0.597	**0.225**	**0.132**	0.098	**0.096**
Approximate Balance	1.153	**1.125**	1.034	0.994	0.921	0.717	0.827	0.569
Approx. Residual Balance	**0.994**	**1.146**	**0.614**	**0.554**	**0.219**	**0.131**	0.109	**0.100**
Inv. Propensity Weight	1.236	2.645	1.084	1.932	0.950	1.091	0.814	0.813
Augmented IPW	1.082	1.231	**0.629**	0.597	**0.224**	**0.132**	0.098	**0.096**
Weighted Elastic Net	1.089	1.234	**0.624**	0.597	**0.225**	**0.132**	0.099	**0.096**
TMLE Elastic Net	1.065	1.233	**0.629**	0.597	**0.225**	**0.132**	0.099	**0.096**
Double-Select + OLS	1.312	2.659	1.064	1.938	0.629	0.204	**0.092**	**0.097**

Table 4: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the moderately sparse two-stage setting. We used $n = 600$, $p = 2000$, and scaled the signal such that $\|\beta\|_2 = 1$. All numbers are averaged over 400 simulation replications.

n	400					1000				
p	100	200	400	800	1600	100	200	400	800	1600
Naive	1.734	1.738	1.734	1.736	1.747	1.724	1.679	1.706	1.698	1.720
Elastic Net	0.446	0.468	0.492	0.517	0.540	0.376	0.380	0.389	0.401	0.413
Approximate Balance	0.523	0.582	0.609	0.656	0.700	0.297	0.327	0.379	0.395	0.464
Approx. Residual Balance	**0.249**	**0.276**	**0.270**	**0.295**	**0.310**	**0.168**	**0.175**	**0.176**	**0.179**	**0.194**
Inv. Propensity Weight	1.060	1.081	1.111	1.154	1.189	0.831	0.831	0.874	0.875	0.940
Augmented IPW	0.340	0.359	0.377	0.406	0.425	0.249	0.254	0.261	0.266	0.285
Weighted Elastic Net	0.313	0.338	0.355	0.385	0.412	0.204	0.209	0.220	0.221	0.249
TMLE Elastic Net	0.347	0.365	0.381	0.407	0.428	0.273	0.275	0.282	0.286	0.301
Double-Select + OLS	0.285	0.292	0.301	0.320	0.339	0.250	0.250	0.246	0.244	0.246

Table 5: Root-mean-squared error $\sqrt{\mathbb{E}\left[(\hat{\tau}-\tau)^2\right]}$ in the misspecified setting. All numbers are averaged over 400 simulation replications.

		$\beta_j \propto 1\left(\{j \leq 10\}\right)$		$\beta_j \propto 1/j^2$		$\beta_j \propto 1/j$	
n	p	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$	$\eta = 0.25$	$\eta = 0.1$
400	800	0.95	0.87	0.97	0.88	0.87	0.70
400	1600	0.92	0.87	0.94	0.89	0.88	0.72
400	3200	0.90	0.82	0.94	0.86	0.86	0.71
800	800	0.94	0.92	0.97	0.92	0.95	0.85
800	1600	0.95	0.92	0.97	0.92	0.91	0.83
800	3200	0.94	0.90	0.95	0.91	0.91	0.79
1600	800	0.96	0.94	0.96	0.94	0.97	0.93
1600	1600	0.95	0.93	0.98	0.94	0.97	0.91
1600	3200	0.95	0.92	0.96	0.95	0.94	0.90

Table 6: Coverage for approximate residual balancing confidence intervals as constructed in Corollary 6, with data generated as in the many cluster setting; we scaled the signal such that $\|\beta\|_2 = 3$. The target coverage is 0.95. All numbers are averaged over 1000 simulation replications.

A Proofs

Proof of Proposition 1

First, we can write

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

Thus,

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

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

Proof of Lemma 2

For any $j = 1, ..., p$, write

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

where e_j is the j-th basis vector, and Q_i denotes the i-th row of the Q matrix (defined in Assumption 3) as a column vector. Here, A_j is a rank-1 matrix, with Frobenius norm

$$\|A_j\|_F^2 = \mathrm{tr}\left(\Sigma_{\mathrm{c}}^{\frac{1}{2}}e_j\xi^{\top}\Sigma_{\mathrm{c}}^{-1}\xi e_j^{\top}\Sigma_{\mathrm{c}}^{\frac{1}{2}}\right) = (\Sigma_{\mathrm{c}})_{jj}\,\xi^{\top}\Sigma_{\mathrm{c}}^{-1}\xi \le VS.$$

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

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

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:

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

We can turn the above moment bound into a tail bound by applying Markov's inequality. Plugging in $t_n := \sqrt{\log(p/2\delta)}\,/\,(\varsigma^2\sqrt{C_1VS})$ and also applying a symmetric argument to $(-\mathbf{X}_{\mathrm{c}}^{\top}\gamma^* + \xi)_j$, we find that for large enough n and any $\delta > 0$,

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

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

Proof of Theorem 3

We start by mimicking Proposition 1, and write

$$\begin{aligned}\hat{\theta} - \theta &= \xi \cdot \left(\hat{\beta}_c - \beta_c\right) + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i - X_i \cdot \hat{\beta}_c\right) \\ &= \sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i(0) + \left(\xi - \mathbf{X}_c^\top \gamma\right) \cdot \left(\hat{\beta}_c - \beta_c\right) \\ &= \sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i(0) + \mathcal{O}\left(\left\|\xi - \mathbf{X}_c^\top \gamma\right\|_\infty \left\|\hat{\beta}_c - \beta_c\right\|_1\right)\end{aligned} \tag{24}$$

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

$$\left\|\xi - \mathbf{X}_c^\top \gamma\right\|_\infty \left\|\hat{\beta}_c - \beta_c\right\|_1 = \mathcal{O}\left(\frac{k \log(p)}{n_c}\right), \tag{25}$$

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

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 $j \in \{1, ..., p\}$ with $|\xi_j| \geq \kappa$; and thus, any feasible solution to (13) must eventually satisfy $(\mathbf{X}_c^\top \gamma)_j^2 \geq \kappa^2/2$. By Cauchy-Schwarz, this implies

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

as desired. Given this result, a standard application of Lyapunov's central limit theorem (Lemma 8) paired with the bound (25) 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 Lemma 2. The concentration bound from Theorem 2.1 in Rudelson and Vershynin (2013) implies that $n_c \|\gamma^*\|_2 / (\xi^\top \Sigma_c^{-1} \xi) \to_p 1$, and so Lemma 2 implies that, with probability tending to 1, the optimization program for γ is feasible and

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

thus concluding the proof.

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

$$\left\|\hat{\beta}_c - \beta_c\right\|_1 \leq \frac{5\varsigma^2}{4} \frac{24\,\upsilon}{\omega} k \sqrt{\frac{\log p}{n_c}}. \tag{26}$$

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

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

Thus, pairing these two fact about $\mathbf{X}_c$ with sparsity as in Assumption 4 and the sub-Gaussianity of the noise $\varepsilon_i(0)$, 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 $\hat{\beta}_c$ by running the lasso with $\lambda = 5\,\varsigma^2\,\upsilon\,S\,\sqrt{\log(p)/n_c}$, then, with probability tending to 1, (26) 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* $\max_i |\gamma_i| \leq n_c^{-2/3}$ *and* $\|\gamma\|_2^2 = \Omega_p(1/n_c)$. *Then, we obtain a central limit theorem*

$$\frac{1}{\|\gamma\|_2} \sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i(0) \Rightarrow \mathcal{N}\left(0,\, \sigma^2\right). \tag{27}$$

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

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

Next, we can again use unconfoundedness to verify that

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

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 (13). Thus,

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

because, by assumption, $\|\gamma\|_2^{-1} = \mathcal{O}_P\left(\sqrt{n_c}\right)$. Thus Lyapunov's theorem implies the central limit statement (27). □

Proof of Corollary 4

The key idea in establishing this result is that we need to replace the "oracle" weights defined in Lemma 2 with

$$\gamma_i^{**} = \frac{1}{n_c} m_t \Sigma_c^{-1} X_i. \tag{28}$$

Once we have verified that, with high probability, these candidate weights γ^{**} satisfy the constraint from (13), i.e., $\left\|\overline{X}_t - \mathbf{X}_c^\top \gamma^{**}\right\|_\infty \leq K\sqrt{\log(p)/n_c}$, we can establish the result (17) by replicating the proof of Theorem 3 verbatim. Now, by Lemma 2, we know that with probability tending to 1,

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

Meanwhile, a standard Hoeffding bound together with the fact that $n_t/n_c \to_p \rho$ establishes that, with probability tending to 1,

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

Combining these two bounds yields the desired result.

Proof of Theorem 5

For concreteness, we note that we are studying the following estimator,

$$\gamma = \text{argmin}_{\tilde{\gamma}} \left\{ \|\tilde{\gamma}\|_2^2 \ : \ \left\|\overline{X}_\text{t} - \mathbf{X}_\text{c}^\top \tilde{\gamma}\right\|_\infty \leq K \sqrt{\frac{\log(p)}{n_\text{c}}}, \sum_{\{i:W_i=0\}} \tilde{\gamma}_i = 1, \, 0 \leq \tilde{\gamma}_i \leq n_\text{c}^{-2/3} \right\}, \tag{29}$$

$$\hat{\mu}_\text{c} = \overline{X}_\text{t} \cdot \hat{\beta}_\text{c} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^\text{obs} - X_i \cdot \hat{\beta}_\text{c} \right), \tag{30}$$

$K = \nu\sqrt{2.1(\rho + (\eta^{-1}-1)^2}$. Note that, here, we have re-incorporated the positivity and sum constraints on γ. The positivity constraint stops us from extrapolating outside of the support of the data, and appears to improve robustness to model misspecification.

Our proof mirrors the one used for Theorem 3. We again start by proposing a class of candidate weights γ^* that satisfy the constraints (29); except, this time, we motivate our candidate weights using the overlap assumption:

$$\gamma_i^* = \frac{e\left(X_i\right)}{1-e\left(X_i\right)} \Big/ \sum_{\{i:W_i=0\}} \frac{e\left(X_i\right)}{1-e\left(X_i\right)}. \tag{31}$$

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 $\gamma_i^* \leq n_\text{c}^{-2/3}$ once n_c is large enough.

Lemma 9. *Under the conditions of Theorem 5, the weights γ^* defined in (31) satisfy the following bounds with probability at least $1-\delta$, for any $\delta > 0$:*

$$\left\|\overline{X}_\text{t} - \mathbf{X}_\text{c}^\top \gamma^*\right\|_\infty \leq \nu \sqrt{2 \log\left(\frac{10\,p}{\delta}\right)\left(\frac{1}{n_\text{t}} + \frac{(1-\eta)^2}{n_\text{c}\,\eta^2}\right)} + \mathcal{O}\left(\frac{1}{n_\text{c}}\right), \quad \textit{and} \tag{32}$$

$$\begin{aligned} n_\text{t} \ \|\gamma^*\|_2^2 \leq \frac{1}{\rho_n^2}\,\mathbb{E}\left[\left(\frac{e(X_i)}{1-e(X_i)}\right)^2 \middle| W_i = 0\right] \\ + \frac{(1-\eta)^2\,(2-2\eta+\rho_n\eta)}{\rho_n^3\,\eta^3} \sqrt{\frac{1}{2n_\text{c}} \log\left(\frac{10\,p}{\delta}\right)} + \mathcal{O}\left(\frac{1}{n_\text{c}}\right), \end{aligned} \tag{33}$$

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 $\hat{\beta}_\text{c}$ by running the lasso with $\lambda = 5\,\nu\,\upsilon\,\sqrt{\log(p)/n_\text{c}}$, then, with probability tending to 1,

$$\left\|\hat{\beta}_\text{c} - \beta_\text{c}\right\|_1 \leq \frac{5\nu}{4}\,\frac{24\,\upsilon}{\omega}\,k\,\sqrt{\frac{\log p}{n_\text{c}}}. \tag{34}$$

Thus, we again find that

$$\sqrt{n}\,\left\|\overline{X}_\text{t} - \mathbf{X}_\text{c}^\top \gamma\right\|_\infty \left\|\hat{\beta}_\text{c} - \beta_\text{c}\right\|_1 = \mathcal{O}_P\left(\frac{k\,\log(p)}{\sqrt{n}}\right), \tag{35}$$

and so, thanks to our sparsity assumption, we can us Proposition 1 to show that the error in $\hat{\beta}_\text{c}$ does not affect the asymptotic distribution of our estimator at the $\sqrt{n}$-scale; provided the problem (29) is feasible.

Next, thanks to Lemma 9, we know that the problem (29) is feasible with high probability. Moreover, because the weights γ obtained via (29) satisfy $\sum \gamma = 1$, we trivially find that $\|\gamma\|_2^2 \geq 1/n_\text{c}$ and can apply Lemma 8 to get a central limit result for $\hat{\mu}_\text{c} - \hat{\mu}_\text{c}$. Finally, invoking (33) and the fact that $\|\gamma\|_2 \leq \|\gamma^*\|_2$ with probability tending to 1, we obtain the desired rate bound (19).

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 $j = 1, ..., p$:

- $\overline{X}_{\mathrm{t},j} = n_{\mathrm{t}}^{-1} \sum_{\{i:W_i=1\}} X_{ij}$ 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 $\mathbb{E}[A_j] = \mathbb{E}\left[X_j \,\middle|\, W = 1\right]$, and so given our re-centering $\mathbb{E}[A_j] = 0$.
- $D := n_{\mathrm{c}}^{-1} \sum_{\{i:W_i=0\}} e(X_i)/(1-e(X_i)) - \rho_n$ is sub-Gaussian with parameter $(1-\eta)^2/(4n_{\mathrm{c}}\,\eta^2)$, where $\rho_n = \mathbb{P}[W=1]/\mathbb{P}[W=0]$ 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 $(1-\eta)^4/(4n_{\mathrm{c}}\,\eta^4)$ 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 $1-\delta$:

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

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

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

Moreover, noting that

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

we see that on ε_δ,

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

and so γ^* in fact satisfies all desired constraints.

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

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

To establish this claim, note that our bias bound (35) 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 (20), 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

$$\frac{1}{V_c} \sum_{\{i:W_i=0\}} \gamma_i^2 \left(Y_i - X_i \cdot \beta_c\right)^2 \to_p 1, \text{ and} \tag{36}$$

$$\frac{1}{V_c} \sum_{\{i:W_i=0\}} \gamma_i^2 \left(X_i \cdot \left(\beta_c - \hat{\beta}_c\right)\right)^2 \to_p 0. \tag{37}$$

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

$$\text{Var}\left[\sum_{\{i:W_i=0\}} \gamma_i^2 \left(Y_i - X_i \cdot \beta_c\right)^2 \,\middle|\, \gamma\right] \leq C_4 \, v^4 \, \|\gamma\|_4^4 \leq C_4 \, v^4 \, n_c^{-4/3} \, \|\gamma\|_2^2,$$

and so (36) holds by Markov's inequality. Meanwhile, to establish (37), 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

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

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

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

We can use sub-Gaussianity of X_i (Assumption 7) and the bound (34) on the L_1-error of $\hat{\beta}_c$ to find a constant $C(\nu, \omega, v)$ for which

$$\|\mathbf{X}_c\|_\infty^2 \left\|\beta_c - \hat{\beta}_c\right\|_1^2 \leq C(\nu, \omega, v) \, \log(p \, n_c) \;\; k^2 \, \frac{\log(p)}{n_c}$$

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

$$\log(p \, n_c) \;\; k^2 \, \frac{\log(p)}{n_c} \ll \frac{\log(p \, n_c)}{\log(p)},$$

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