High-Dimensional Feature Selection for Sample Efficient Treatment Effect Estimation

Kristjan Greenewald
MIT-IBM Watson AI Lab

Dmitriy Katz-Rogozhnikov
IBM Research

Karthikeyan Shanmugam
IBM Research

Abstract

The estimation of causal treatment effects from observational data is a fundamental problem in causal inference. To avoid bias, the effect estimator must control for all confounders. Hence practitioners often collect data for as many covariates as possible to raise the chances of including the relevant confounders. While this addresses the bias, this has the side effect of significantly increasing the number of data samples required to accurately estimate the effect due to the increased dimensionality. In this work, we consider the setting where out of a large number of covariates X that satisfy strong ignorability, an unknown sparse subset S is sufficient to include to achieve zero bias, i.e. c-equivalent to X. We propose a common objective function involving outcomes across treatment cohorts with nonconvex joint sparsity regularization that is guaranteed to recover S with high probability under a linear outcome model for Y and subgaussian covariates for each of the treatment cohort. This improves the effect estimation sample complexity so that it scales with the cardinality of the sparse subset S and $\log |X|$, as opposed to the cardinality of the full set X. We validate our approach with experiments on treatment effect estimation.

1 Introduction

Consider the problem of estimating the treatment effect of T on a univariate outcome Y in the presence of (possibly confounding) covariates X, where the treat-

ment variable can take q possible treatment configurations. We assume only observational data is available. The causal graph for this setup is shown in Figure 1.

One of the central issues of causal effect estimation is identifying features that are confounders and controlling for them. Let $Y_t(X)$ denote the counterfactual outcome associated when treatment t is applied as an intervention given X. We consider the simpler case when the observed set of covariates X is admissible or eligible to be used for adjustment. In other words, for any treatment t, $Y_t \perp T|X$, i.e. the counterfactual outcome associated with any treatment t is independent of the treatment choice in the observational data given X. We are interested in the problem of estimating the average treatment effect between the pair of treatments given by $\mathbb{E}_X [Y_t - Y_{t'}]$. This is denoted by ATE. If X is admissible this can be estimated from observational data. Inverse propensity weighing, standardization and doubly robust estimation are standard techniques used [Guo et al., 2020, Imbens and Wooldridge, 2009].

If $X \in \mathbb{R}^p$ is high dimensional (large p), however, the number of samples required to estimate the treatment effects accurately becomes too large to be practical in many applications. In practice, features in X are designed to include as many factors as possible to capture all relevant confounders that are needed to satisfy the admissibility criterion. [Shpitser and Pearl, 2012] showed that if we know the semi-Markovian causal model behind the observational data, then one can algorithmically identify if a given subset of X is admissible or not (even if X is not admissible).

In our work, we focus on the case when X is admissible but no detailed causal model is available. We study sufficient conditions for identifying if a subset $S \subset X$ is admissible given that X is known to be admissible. A subset S_1 is *c-equivalent* to another subset S_2 if S_2 being admissible implies S_1 being admissible and vice versa. In other words, both subsets can be used for adjustment and will yield the same ATE estimate. We rely on sufficient conditions for c-equivalence in [Pearl, 2009] as our main technical tool.
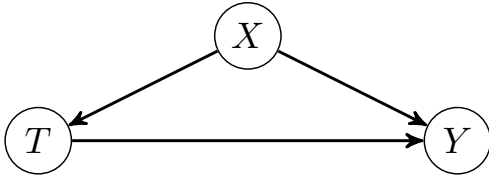
Figure 1: Causal graph. T is a discrete treatment, taking up to q values, and Y is a scalar outcome. X is an observed set of p covariates.

We consider a coarser causal model given in Figure 2, where X has been decomposed into the sets X_1, X_2 (confounders), and X_3 (predictors) based on their connections to Y and T. Applying sufficient conditions for c-equivalence, we show that it is sufficient to use either of two possible sets to form unbiased treatment effect estimates: $S = X_2 \cup X_3$ and $X_1 \cup X_2$.[1]

Prior work on sparse feature selection for treatment effect estimation has focused on the case where $X_1 \cup X_2$ is sparse [Shortreed and Ertefaie, 2017, Cheng et al., 2020], i.e. the number of confounding variables plus the number of variables biasing the treatment is small. In this work, we complete the picture by considering the companion setting where instead $S = X_2 \cup X_3$ is sparse, i.e. the number of confounding variables plus the number of predictors is small. In practice, we suggest running both our method and a method that identifies $X_1 \cup X_2$ and choosing the one that yields the lowest variance unbiased estimate. An added benefit of using S over $X_1 \cup X_2$ is that S includes the set of predictors, which serve to reduce the variance of the treatment effect estimate [Shortreed and Ertefaie, 2017].

Contributions: Given X is admissible and given q treatment cohorts, under a linear outcome model for Y given T and X, we show that maximizing least squares likelihood with a joint sparse non convex regularization recovers the subset S of interest. The number of samples required is $O(kq \log p)$ and the error in the parameter estimates scales as $O(\sqrt{\frac{\log p}{n}})$ where $|X| = p$ and n is the number of samples (Theorem 1 below). We demonstrate the effectiveness of our subset identification step in synthetic experiments as well in combination with doubly robust ATE estimation procedures on real datasets.

Prior Work: [Guo et al., 2020, Imbens and Wooldridge, 2009] provide surveys of methods that address causal effect estimation with observational data both from machine learning and

econometrics perspectives. These surveys review classic approaches to ATE estimation including propensity weighing, doubly robust estimation and matching techniques. We only briefly review a small subset of these works in what follows.

Perhaps the most relevant to our work is [Shortreed and Ertefaie, 2017] which also considered variable selection for causal inference in the regime stated in Figure 1. In contrast to our approach that regresses Y on S given a fixed T, they use an outcome adaptive lasso[2] sparse regression on the logistic transformation of $P(T = 1|X)$, in order to find the $X_1 \cup X_2$ set. This choice limits the approach to binary treatments, and the associated theory is limited to asymptotic consistency, with no indication of sample complexity relative to the sparsity or dimensionality. [Cheng et al., 2020], instead of finding a sparse subset $X_1 \cup X_2$, transform the covariates into a low dimensional space that satisfies conditional independence criteria. [Shah and Bühlmann, 2019] considers variable selection for continuous-valued variable of interest T, which leads them to seek (generalized) linear regressions of Y on X and T simultaneously. Our approach instead uses the discreteness of T to significantly relax the sparse linearity assumption to only apply to the regression of Y on only X, given T. We believe this is a significant advantage in practice, as very many actions of interest are discrete, and moreover it allows us to consider cases where the coefficients of the regression of Y with respect to X differ with treatment T. [Bradic et al., 2019] considers variable selection in the setting of discrete T as we do, but our joint-sparse nonconvex approach allows us to handle the case when sparsity k is linear in samples n, while their individual-sparsity based approach requires k to be "ultra-sparse" $o(\sqrt{n}/\log p)$. Additionally, their approach is limited to ATE, while ours applies to both ATE and ITE.

A growing body of recent work has been applying machine learning techniques to the problem of ITE estimation. [Athey and Imbens, 2015, Kuenzel, 2019] introduce meta frameworks for applying supervised learning for ITE estimation. [Hill, 2011] applies Bayesian techniques to ITE estimation. Inspired by the rise of deep learning, [Kallus, 2018] used adversarial training to find covariate representations that match across treatment cohorts. Various recent works apply domain adaptation techniques to learn deep representations that match the treatment cohorts [Yao et al., 2018, Shalit et al., 2017, Yoon et al., 2018]. [Louizos et al., 2017] uses variational autoencoders to find noisy proxies for latent confounders, and uses the result for ITE estimation.

---

[1]Nodes in X that do not have edges to either T or Y should not be included in either of the two sets. We omit these from the figure for simplicity.
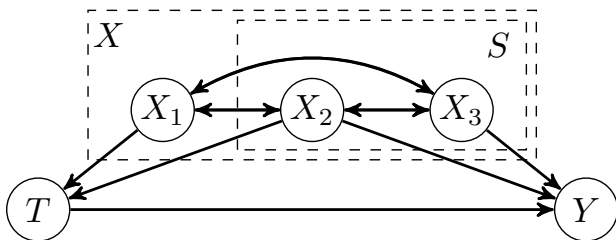
[2]Lasso weighted by the unregularized coefficients.

Figure 2: Partition of X by connections to T and Y. X is composed of X_1 (arrows into T not Y), X_2 (arrows into both T and S, i.e. confounders), and X_3 (arrows into Y not T, i.e. predictors). The identities of these sets are not known a priori and must be discovered from data. S is composed of X_2 and X_3.

[Wager and Athey, 2018] leverage latest advances in learning using forests for ITE estimation problems. When ITE/ATE is not identifiable from data, interval estimates on treatment effects have been obtained in [Kallus et al., 2019, Yadlowsky et al., 2018].

Notation: For a matrix $A \in \mathbb{R}^{p \times q}$, we define $A_{i:}$ to be the ith row of A and $A_{\cdot j}$ to be the jth column of A. We also define the norm $\|A\|_{a,b}$ for $a, b \in \mathbb{R}^+ \cup \infty$ as $\|A\|_{a,b}^a = \sum_{i=1}^p \|A_{i:}\|_b^a$. We denote $|||A|||_a$ as the ath order matrix norm, and $\|A\|_F = \|A\|_{2,2}$ as the Frobenius norm.

2 Treatment Effects and Admissible Sets

In the case of binary treatments, the average treatment effect is given by

$$E[Y|\mathrm{do}(T=1)] - E[Y|\mathrm{do}(T=0)],$$

and the individual treatment effect by

$$E[Y|X, \mathrm{do}(T=1)] - E[Y|X, \mathrm{do}(T=0)].$$

For higher cardinality T, similar pairwise differences involving the $E[Y|X, \mathrm{do}(T=t)]$ are in order.

Since we only have observational data, we make use the following property from [Pearl, 2009]:

Definition 1 (Admissibility). *A set X is called admissible if*

$$p(y|\mathrm{do}(T=t)) = \int p(y|t, x)p(x)dx, \qquad (1)$$

i.e. we can compute the causal effect using observational probabilities controlling for X.

X will be admissible if there are no hidden confounders. Note that the dimensionality of the integral (1) is large since X is high dimensional. Can

we simplify this expression to involve only the sparse subset S? We make use of the following definition [Pearl, 2009].

Definition 2 (c-equivalence). *Two subsets S_1 and S_2 are c-equivalent if*

$$\int p(y|t, s_1)p(s_1)ds_1 = \int p(y|t, s_2)p(s_2)ds_2,$$

i.e. the causal effect distributions controlling for S_1 and S_2 are equal.

Definition 2 implies that if S_2 is c-equivalent to S_1 and S_1 is admissible, then so is S_2.

We now show S and $X_1 \cup X_2$ are c-equivalent to X.

Lemma 1. *Given the causal graph in Figure 2, both the subset S and the subset $X_1 \cup X_2$ are c-equivalent to the set X, hence either subset is sufficient as control to compute an unbiased estimate of the treatment effect.*

This result is proved in supplement Section 9, and relies on two sufficient conditions for strong ignorability and c-equivalence given in Chapter 11 of [Pearl, 2009].

Lemma 1 establishes that there is no additional bias resulting from controlling for either S only or $X_1 \cup X_2$ only instead of the full X. Assuming that X is an admissible set, i.e. there are no hidden confounders, there will be no bias and S will also be an admissible set. The question then is which of these two sets to control for.

One might imagine that when X is high dimensional, the sparser of the two admissible subsets S and $X_1 \cup X_2$ should be used. It has been shown, however [Witte et al., 2020] that effect estimates with S will always be lower variance than estimates using $X_1 \cup X_2$ since S includes all predictors X_3 of Y. Thus finding S has a strict advantage (for effect estimation) over previous works such as [Shortreed and Ertefaie, 2017] that focused on finding and controlling for $X_1 \cup X_2$. In this work we thus propose an estimator for the admissible set S and theoretically prove bounds on its sample complexity when T is discrete and Y continuous.

Given observational samples of X, Y, T, our goal is thus to find the smallest subset S containing all nodes in X that have an edge pointing towards Y in the graph.[3] Since we have assumed that the outcome Y does not have any edge pointing to X or T, it is sufficient to use observational data to condition on $T = t$ and find the set of nodes S_t in X that have edges connecting

---

[3]Note that if there are nodes in X_3 that do not have any direct connections to X_1, they are not needed for admissibility. We choose to include them in S since they reduce the variance of the treatment effect estimator.

to Y in the undirected graph, and then take the union over t as $S = \bigcup_{t=1}^{q} S_t$.

3 Oracle ATE/ITE

In this section, we describe treatment effect estimation in the oracle setting where the support S is known. Suppose that an oracle gives us the identity of the optimal S subset. By Lemma 1, we have that

$$P(Y|T) = \int P(Y|S,T)P(S)dS.$$

This can be estimated directly from empirical probabilities, although with continuous S the sample complexity is still significant without additional assumptions. In this work, we make use of the following assumption of linearity with respect to S (to be relaxed in future work).

Assumption 1 (Linearity). *Assume that Y follows the following generative model depending on T and S:*

$$Y = \theta_{:t}^T S + \epsilon,$$

where $\theta \in \mathbb{R}^{k \times q}$ is a matrix of linear coefficients and ϵ is i.i.d. noise.

Suppose a regression estimate $\hat{\theta} \in \mathbb{R}^{k \times q}$ of the coefficient matrix θ is available. For binary treatments, the individual treatment effect (ITE) can then be estimated by

$$\widehat{ITE}(X) = (\hat{\theta}_{:1} - \hat{\theta}_{:0})^T S,$$

where the $\hat{\theta}_i$ are regression coefficient estimates.

Similarly, we can estimate the average treatment effect (ATE) as

$$\widehat{ATE} = (\hat{\theta}_{:1} - \hat{\theta}_{:0})^T \mu_S,$$

where μ_S is the specified mean of S.

We have the following lemma relating treatment effect estimation error to coefficient estimation error. The proof is immediate from norm inequalities.

Lemma 2 (Oracle Effect Estimation Error). *Given Assumption 1, we have for binary treatments: $|\widehat{ITE}(S) - ITE(S)| \leq \|S\|_1 \cdot \sum_{t=0}^{1} \|\hat{\theta}_{:t} - \theta_{:t}\|_{\infty}$, $|\widehat{ATE} - ATE| \leq \|\mu_S\|_1 \cdot \sum_{t=0}^{1} \|\hat{\theta}_{:t} - \theta_{:t}\|_{\infty}$. More generally, for q treatments define $\tau(t) = E[Y|S, \mathrm{do}(T = t)] = \theta_{:t}^T S$. We have for all t that $|\hat{\tau}(t) - \tau(t)| \leq \|S\|_1 \|\hat{\theta} - \theta\|_{\infty,\infty}$.*

Note that the error bounds in Lemma 2 grow linearly with $\|S\|_1$, which tends to grow linearly with the cardinality $|S| = k$. This confirms our motivation for finding sparse solutions to reduce sample complexity. In this section, we assumed that the sparse admissible set S was given to us by an oracle. In the next section, we consider the real world setting where we must recover S from the data itself.

4 Jointly Sparse Variable Selection

Our goal is to estimate the matrix of linear coefficients θ in Assumption 1 using sparse regression of Y versus X given fixed T. The classic approach to sparse regression is the lasso objective, which in our setting is

$$\hat{\theta}_{:,j} = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{2} \theta^T \frac{X_j^T X_j}{n} \theta - \frac{y_j^T X_j}{n} \theta + \lambda \|\theta\|_1,$$

where X_j, y_j are samples from the $T = j$ conditional.[4]

Since we care about the union of the nonzero supports of the $\theta_{:j}$, it is wasteful to force an entry to zero in the $t = 0$ graph if we know it is nonzero in $t = 1$, etc. Hence, we instead use *group sparsity*, which couples the sparsity of the q vectors together.

Traditionally, group sparsity is encouraged via the L-1,2 norm [Huang et al., 2010, Lounici et al., 2011], which is an L1 norm of the L2 norms of the rows of θ. Copying the above, we can write the group-sparsity based objective as

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{p \times q}} \sum_{j=1}^{q} \left[\frac{1}{2} \theta_{:j}^T \frac{X_j^T X_j}{n} \theta_{:j} - \frac{y_j^T X_j}{n} \theta_{:j} \right] + \lambda \|\theta\|_{1,2}$$

This can be solved iteratively as it is a convex problem, and 2-norm error bounds exist [Huang et al., 2010, Lounici et al., 2011].

Unfortunately, it is known that L1 based regression, while successful in estimating coefficients in terms of L2 norm error, does not perform well for variable selection without complex incoherence assumptions [Loh and Wainwright, 2017]. To avoid these difficult-to-interpret assumptions, instead of L1 we will rely on the following class of nonconvex regularizers that retain the sparsity-promoting properties of the "cusp" at $t = 0$, while using a nonconvex shape to not excessively penalize large coefficients.

Definition 3 $((\mu, \gamma)$-amenability). *A regularization function ρ_λ with parameter λ is μ-amenable for some $\mu > 0$ if the following hold:*

- ρ_λ *is symmetric around 0 and $\rho_\lambda(0) = 0$.*

- ρ_λ *is nondecreasing on $\mathbb{R}^+$.*

- *the function $\frac{\rho_\lambda(t)}{t}$ is nonincreasing on $\mathbb{R}^+$.*

- $\rho_\lambda(t)$ *is differentiable at all $t \neq 0$.*

- $\rho_\lambda + \frac{\mu}{2} t^2$ *is convex.*

- $\lim_{t \to 0^+} \rho_\lambda'(t) = \lambda$.

---

[4]For simplicity, throughout the paper we assume n samples are available from each conditional. The results can be easily adjusted to the case of imbalanced sample sets.

Kristjan Greenewald, Dmitriy Katz-Rogozhnikov, Karthikeyan Shanmugam

If in addition there is some scalar $\gamma \in (0, \infty)$ such that $\rho'_\lambda = 0$ for all $t \geq \gamma\lambda$, then ρ_λ is (μ, γ)-amenable.

Two example (μ, γ) amenable regularizers are the SCAD [Fan and Li, 2001] and MCP [Zhang et al., 2010] penalties. For convenience, define $q_\lambda(t) = \lambda|t| - \rho_\lambda(t)$. If ρ_λ is (μ, γ) amenable, then q_λ is everywhere differentiable.

Applying a (μ, γ) regularizer ρ_λ on the row 2-norms we have the following objective function:

$$\hat{\theta} = \arg \min_{\|\theta\|_{1,2} \leq R} \left\{ \sum_{j=1}^{q} \left[\frac{1}{2}\theta_{:j}^T \frac{X_j^T X_j}{n} \theta_{:j} - \frac{y_j^T X_j}{n} \theta_{:j} \right] \right.$$
$$\left. + \sum_{i=1}^{p} \rho_\lambda \left(\|\theta_{i:}\|_2 \right) \right\}, \qquad (2)$$

For convenience, define the unregularized loss function

$$\mathcal{L}_n(\theta) = \sum_{j=1}^{q} \left[\frac{1}{2}\theta_{:j}^T \frac{X_j^T X_j}{n} \theta_{:j} - \frac{y_j^T X_j}{n} \theta_{:j} \right]. \qquad (3)$$

We show below that the objective (2) is a convex problem when R and μ are chosen appropriately. We thus optimize the objective 2 using proximal gradient descent. Since the function $q_\lambda(t) = \lambda|t| - \rho_\lambda(t)$ is everywhere differentiable, it can be included in the gradient step computation, leaving the proximal step to be the proximal operator for the L-1,2 norm. This proximal operator is simply a soft thresholding on the norms of the rows of θ, i.e. setting the rows of θ as $\max(0, \|\theta_{i:}\|_2 - \lambda) \frac{\theta_{i:}}{\|\theta_{i:}\|_2}$. If an optimization step would go outside the constraint set, we reduce the step size until the constraint is satisfied. A summary of the optimization algorithm is in supplement Section 7.

4.1 Theoretical Analysis

For the analysis, we make the additional assumption[5]

Assumption 2 (Subgaussianity)**.** *Assume that conditioned on $T = t$, X is subgaussian with parameter bounded from above by σ_x for all t, and the noise term ϵ given in Asspt. 1 is subgaussian with parameter σ_ϵ.*

We can then state the following bound.

Theorem 1. *Suppose $p > q$ and for all $j = 1, \ldots, q$, $y_j = (\theta^*_{:j})^T X_j + \epsilon$ where X and ϵ are subgaussian, and all $\theta^*_{:j}$ have support contained in some unknown set S with $|S| = k$ where k is unknown. Furthermore, choose (λ, R) such that $\|\theta^*\|_{1,2} < \frac{R}{2}$ and $c_\ell\sqrt{\frac{q\log p}{n}} \leq \lambda \leq \frac{c_u\sqrt{q}}{R}$, and assume $n \geq C \max\{R^2, k\}q\log p$ for*

---

some constants c_ℓ, c_u, C described in the proof. Suppose ρ_λ is a (μ, γ)-amenable regularizer with $\mu < \frac{1}{2}\min_j \lambda_{\min}(\Sigma_x^{(j)})$ where $\Sigma_x^{(j)} = \mathbb{E}_{X_j} \frac{1}{n} X_j^T X_j$. Finally, suppose that

$$\theta^*_{\min} := \min_{i \in S} \|\theta^*_{i:}\|_2 \geq \lambda\gamma + c_3\sqrt{\frac{\log p}{n}}. \qquad (4)$$

Then with probability at least $1 - c_1\exp(-c_2\min[k, \log p])$ the objective (2) has a unique stationary point $\hat{\theta}$ with support equal to S and satisfying

$$\|\hat{\theta} - \theta^*\|_{\infty,\infty} \leq c_3\sqrt{\frac{\log p}{n}}.$$

Remark 1. *Note that the key result is the exact recovery of the support S. The infinity norm bound controlling the error on the individual coefficients in the support S can be combined with the support recovery result to bound other norms of the error, for instance knowing that there are only $|S|q$ coefficients with nonzero error implies $\|\hat{\theta} - \theta^*\|_F \leq \sqrt{|S|q}\|\hat{\theta} - \theta^*\|_{\infty,\infty}$, yielding a tight bound on the Frobenius norm of the error.*

Remark 2. *This proof technique can also yield consistency for the L-1,2 norm regularizer with an appropriate incoherence assumption, see [Loh and Wainwright, 2017] Proposition 3. Recall however that the incoherence assumption is not interpretable and difficult to check in practice.*

Remark 3. *The infinity norm error rate in Theorem 1 is optimal (since it coincides with the estimation error of the optimal oracle estimator).*

Proof of Theorem 1. Various steps in the proof are outlined below:

0. Define and verify a joint Restricted Strong Convexity condition.

1. Optimize the *oracle* program where the supports of $\hat{\theta}_{:j}$ are restricted to the true S:

$$\hat{\theta} = \arg \min_{\theta \in S, \|\theta\|_{1,2} \leq R} \sum_{j=1}^{q} \left[\frac{1}{2}\theta_{:j}^T \frac{X_j^T X_j}{n} \theta_{:j} - \frac{y_j^T X_j}{n} \theta_{:j} \right]$$
$$+ \sum_{i \in S} \rho_\lambda \left(\|\theta_{i:}\|_2 \right), \qquad (5)$$

and show the solution is in the interior of the constraint set. By the restricted strong convexity of $\mathcal{L}_n$, this implies that the solution is a zero subgradient point.

2. Define the dual variable $\hat{z}$ where $\hat{z}_S \in \nabla\|\hat{\theta}_S\|_{1,2}$ and $\hat{z}_{S^c}$ satisfying the zero subgradient condition, and establish *strict dual feasibility* of $\hat{z}_{S^c}$ by showing that $\|\hat{z}_{S^c}\|_{\infty,2} \leq 1$. This implies $\hat{\theta}$ is a stationary point of the full objective (2).

3. Show that $\hat{\theta}$ is the unique global minimum of the full objective (2).

Step 0: First, we verify a restricted strong convexity condition. Adapted from the $q = 1$ case in [Loh and Wainwright, 2017], we require the following property of the loss function:

Definition 4 (Joint Restricted Strong Convexity (Joint RSC)). *We say a loss $\mathcal{L}_n(\theta)$, $\theta \in \mathbb{R}^{p \times q}$ satisfies an (α, τ) joint RSC condition if for all $\Delta \in \mathbb{R}^{p \times q}$*

$$\langle \nabla \mathcal{L}_n(\theta + \Delta) - \nabla \mathcal{L}_n(\theta), \Delta \rangle \tag{6}$$
$$\geq \begin{cases} \alpha_1 \|\Delta\|_F^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_{1,2}^2 & \|\Delta\|_F \leq 1 \\ \alpha_2 \|\Delta\|_F - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_{1,2} & \|\Delta\|_F \geq 1. \end{cases}$$

The following is proven in supplement Section 11.

Lemma 3 (Joint RSC for least squares loss). *Assume that $n \geq O(k \log p)$ and $n \geq 4R^2 q \log p$. With high probability (at least $1 - qc_1 \exp(-cn)$), $\mathcal{L}_n$ is (α, τ)-joint RSC for $\alpha_1 = \alpha_2 = \frac{1}{2} \min_j(\lambda_{\min}(\Sigma_x^{(j)}))$ and $\tau_1 = q$, $\tau_2 = \sqrt{q}$. Furthermore, the objective (5) is strongly convex on $\mathbb{R}^S$.*

We also have that with high probability

$$\|\nabla \mathcal{L}_n(\theta^*)\|_{\infty,2} \leq c' \sqrt{\frac{q \log p}{n}}, \tag{7}$$

by applying a norm inequality (2-norm is $\leq \sqrt{q}$ times infinity norm) to the union bounded bound in the proof of Corollary 1 in [Loh and Wainwright, 2015] (the $q = 1$ case) and using $q < p$.

Step 1: We recall $\|\theta^*\|_{1,2} \leq R/2$ and use the joint RSC conditions to bound $\|\tilde{\nu}\|_{1,2}$, where we set $\tilde{\nu} := \hat{\theta} - \theta^*$. We state the result as a lemma, proven in supplement Section 10.

Lemma 4. *Suppose $\hat{\theta}$ is a zero subgradient point of the objective (5) supported on S, i.e.*

$$\nabla \mathcal{L}_n(\hat{\theta}_S) + \nabla \rho_\lambda(\hat{\theta}_S) = 0. \tag{8}$$

Then $\|\tilde{\nu}\|_{1,2} < \frac{R}{2}$, yielding $\|\hat{\theta}\|_{1,2} < R$.

Since $\|\hat{\theta}\|_{1,2}$ is strictly less than R, $\hat{\theta}$ is in the interior of the constraint set, and thus has zero subgradient.

Step 2: Denote $\hat{\Gamma}^{(j)} = \frac{X_j^T X_j}{n}$, $\hat{\gamma}^{(j)} = \frac{X_j^T y_j}{n}$. Then taking the gradients of (5) yields for all j

$$\nabla \mathcal{L}_n(\theta_{:j}) = \hat{\Gamma}^{(j)} \theta_{:j} - \hat{\gamma}^{(j)}, \quad \nabla^2 \mathcal{L}_n(\theta_{:j}) = \hat{\Gamma}^{(j)}. \tag{9}$$

Consider the estimator $\hat{\theta}^{\mathcal{O}}$ formed by solving (5) with $\lambda = 0$. We then can write

$$\hat{\Gamma}^{(j)}(\hat{\theta}_{:j}^{\mathcal{O}} - \theta_{:j}^*) = \nabla \mathcal{L}_n(\hat{\theta}_{:j}^{\mathcal{O}}) - \nabla \mathcal{L}_n(\theta_{:j}^*), \forall j,$$

yielding (since $\hat{\Gamma}_{SS}^{(j)}$ is invertible since $n \geq k$ by assumption)

$$\hat{\theta}_{Sj}^{\mathcal{O}} - \theta_{Sj}^* = (\hat{\Gamma}_{SS}^{(j)})^{-1}(-(\hat{\Gamma}_{SS}^{(j)} \theta_{Sj}^* - \hat{\gamma}_S^{(j)})). \tag{10}$$

Appendix D.1.1 of [Loh and Wainwright, 2017] showed that

$$\left\| (\hat{\Gamma}_{SS}^{(j)})^{-1}(\hat{\Gamma}_{SS}^{(j)} \theta_{Sj}^* - \hat{\gamma}_S^{(j)}) \right\|_\infty \leq \lambda_{\max}^{1/2}(\Sigma_x^{(j)}) \sigma_\epsilon \sqrt{\frac{2 \log p}{n}}, \tag{11}$$

with probability at least $1 - c_1'' \exp(-c_2'' \min(k, \log p))$.

Hence we obtain via the union bound that

$$\|\hat{\theta}^{\mathcal{O}} - \theta^*\|_{\infty,\infty} \leq c_3 \sqrt{\frac{\log p}{n}}, \|\hat{\theta}^{\mathcal{O}} - \theta^*\|_{\infty,2} \leq c_3 \sqrt{\frac{q \log p}{n}} \tag{12}$$

with probability at least $1 - c_1 \exp(-c_2 \min(k, \log p))$ (since $k > \log q$ and $p > q$) where c_1, c_2, c_3 are constants.

Now we have the following result, proved in supplement Section 13.

Lemma 5. *Suppose ρ_λ is (μ, γ) amenable and*

$$\theta_{\min}^* = \min_{i \in S} \|\theta_{i:}^*\|_2 \geq \lambda \gamma + c_3 \sqrt{\frac{\log p}{n}}.$$

Then with probability at least $1 - c_1 \exp(-c_2 \min(k, \log p))$

$$\lambda \hat{z}_{i:} - \nabla q_\lambda(\|\hat{\theta}_{i:}\|_2) = 0 \quad \forall i \in S.$$

Lemma 5 implies that if $\theta_{\min}^*$ satisfies the given condition, then $\nabla_{\theta_S} \rho_\lambda(\hat{\theta}_{S:}) = 0$, implying that $\hat{\theta}^{\mathcal{O}}$ is a zero subgradient point of (5) and hence $\hat{\theta} = \hat{\theta}^{\mathcal{O}}$. Hence the bound (12) also applies to $\hat{\theta}$ as in the theorem statement.

Now, define the shifted objective function as

$$\bar{\mathcal{L}}_n(\theta) = \mathcal{L}_n(\theta) - \sum_{i=1}^p q_\lambda(\|\theta_{i:}\|_2). \tag{13}$$

Making $\hat{\theta} = (\hat{\theta}_S, 0)$, the zero subgradient condition becomes

$$\nabla \bar{\mathcal{L}}_n(\hat{\theta}) + \lambda \hat{z} = 0, \tag{14}$$

where $\hat{z} \in \partial \|\hat{\theta}\|_{1,2}$. Note that where rows of $\hat{\theta}$ are zero, the corresponding rows of $\hat{z}$ can be any vector in the unit 2-sphere. Where the rows are nonzero, it is a unit vector parallel to the row. Hence we have the strict dual feasibility condition $\|\hat{z}_c\|_{\infty,2} \leq 1 - \delta$ for some delta we choose later.

We expand the zero subgradient condition (14) as

$$\left(\nabla \mathcal{L}_n(\hat{\theta}_{i:}) - \nabla \mathcal{L}_n(\theta_{i:}^*) \right) \tag{15}$$
$$+ \left(\nabla \mathcal{L}_n(\theta_{i:}^*) - \nabla q_\lambda(\|\hat{\theta}_{i:}\|_2) \right) + \lambda \hat{z}_{i:} = 0, \quad \forall i.$$

Note that by the selection property, for all $i \notin S$, $\nabla q_\lambda(\|\hat{\theta}_{i:}\|_2) = \nabla q_\lambda(0) = 0$. Additionally, by Lemma 5 combined with (12) and the assumption (4) we know that $\lambda \hat{z}_{i:} - \nabla q_\lambda(\|\hat{\theta}_{i:}\|_2) = 0$ for all $i \in S$.

Using (9) we can then simplify the condition (15) as

$$\hat{\Gamma}^{(j)}(\hat{\theta}_{:j} - \theta^*_{:j}) + \hat{\Gamma}^{(j)}\theta^*_{:j} - \hat{\gamma}^{(j)} + \begin{bmatrix} 0 \\ (\hat{z}_{S^c})_{:j} \end{bmatrix} = 0, \forall j. \quad (16)$$

Since furthermore we have $\hat{\theta}_{S^c} = \theta^*_{S^c} = 0$, this allows us to solve for each $[\hat{z}_{S^c}]_{:j}$ separately:

$$[\hat{z}_{S^c}]_{:j} = \frac{1}{\lambda}\left[\hat{\gamma}^{(j)}_{S^c} - \hat{\Gamma}^{(j)}_{S^c S}[\hat{\Gamma}^{(j)}_S S]^{-1}\hat{\gamma}^{(j)}_S\right]$$

where we have partitioned $\hat{\Gamma}^{(j)} = \begin{bmatrix} \hat{\Gamma}^{(j)}_{SS} & \hat{\Gamma}^{(j)}_{SS^c} \\ \hat{\Gamma}^{(j)}_{S^c S} & \hat{\Gamma}^{(j)}_{S^c S^c} \end{bmatrix}$.

This quantity was analyzed by [Loh and Wainwright, 2017] Appendix D.1.1. With probability at least $1 - c\exp(-c'\log p)$,

$$\|\hat{\gamma}^{(j)}_{S^c} - \hat{\Gamma}^{(j)}_{S^c S}[\hat{\Gamma}^{(j)}_{SS}]^{-1}\hat{\gamma}^{(j)}_S\|_\infty \leq C\sqrt{\frac{\log p}{n}},$$

assuming $n \geq O(k\log p)$. Using the union bound and definition of $\infty, 2$ norm, we then have that with probability at least $1 - c\exp(\log q - c'\log p)$

$$\|\hat{z}_{S^c}\|_{\infty,2} \leq C\sqrt{\frac{q\log p}{n}}. \quad (17)$$

Strict dual feasibility follows whenever $\lambda > C\sqrt{\frac{q\log p}{n}}$.

Step 3: Since by Step 2 $\hat{\theta}$ is a zero subgradient point of the full objective (2), it is also a local optima of the full objective (2). Furthermore, Lemma 6, proven in supplement Section 14 shows all local optima of (2) must be supported on S.

Lemma 6. *Suppose that $\tilde{\theta}$ is a stationary point of (2) with $\|\hat{z}_{S^c}\|_{\infty,2} \leq 1/2$ and the conditions of Theorem 1 hold with $c_u = \frac{\alpha_2}{8}$ and $c_\ell = \sqrt{q^{-1}\tau_1\alpha_2}$, and $n \geq \max\{\frac{16}{\alpha_2^2}R^2\tau_2^2, \frac{200\tau_1}{\alpha_1-\mu}k\}\log p$. Then for all j, $\mathrm{supp}(\tilde{\theta}_{:j}) \subseteq S$.*

Recall that in Step 2 (17) we showed that the condition of Lemma 6 is satisfied when $C\sqrt{\frac{q\log p}{n}} \leq 1/2$, i.e. whenever $n \geq 4C^2 q\log p$. Hence by strict convexity on the $\mathbb{R}^S$ space (Lemma 3), $\hat{\theta}_S$ is the unique global optimum of the full objective (2). $\square$

4.2 Implication for Effect Estimation

Since we recover the support S with high probability (Theorem 1), we can plug in the bound for the oracle estimate and obtain the following bounds for the linear estimates.

Lemma 7 (Effect Estimation Error). *Given the assumptions of Theorem 1, with high probability the following hold for our estimator. For binary treatments (C is a constant): $|\widehat{ITE}(S) - ITE(S)| \leq 2C\|S\|_1\sqrt{\frac{\log p}{n}}$, $|\widehat{ATE} - ATE| \leq 2C\|\mu_S\|_1\sqrt{\frac{\log p}{n}}$. More generally, for q possible treatments define $\tau(t) = E[Y|S, \mathrm{do}(T = t)] = \theta^T_{:t}S$. We have for all t $|\hat{\tau}(t) - \tau(t)| \leq C'\|S\|_1\sqrt{\frac{q\log p}{n}}$.*

Remark 4 (Comparisons). *Note we only lose a log factor relative to the oracle estimator. The comparison to a nonsparse estimator (one that sets $S = X$) depends on $\|S\|_1$, but for diffuse X such that subsets S typically have $\|S\|_1 = O_p(|S|)$, our estimator improves on the nonsparse estimator by a factor of $\frac{|S|\sqrt{\log p}}{p}$ which is significant for sparse S.*

Remark 5 (Application to nonlinear settings). *We note that our S recovery algorithm is not limited to being used in conjunction with linear effect estimation. Our approach can be used to find a sparse S, and then any desired effect estimator can be applied to the data, controlling only for the set S.*

5 Experimental Results

5.1 Synthetic Data

We use synthetic data generated as follows. X is generated from an isotropic Gaussian distribution. T is generated by sampling from a multinomial distribution with probabilities given by $\mathrm{softmax}(\Phi^T X)$, where $\Phi \in \mathbb{R}^{p \times q}$ has i.i.d. Gaussian elements. The output Y is then generated according to the linear model in Assumption 1, where the k nonzero rows of θ have been sampled from an i.i.d Gaussian distribution. We choose $k = 10$, and use the MCP penalty [Zhang et al., 2010] as our nonconvex regularizer ρ. For binary treatments, i.e. $q = 2$, Figure 8 shows the probability of correctly recovering the set S (with cardinality 10) as the total size p of X and the number of samples n are varied. Note that the number of samples required for consistent recovery of S depends approximately logarithmically on p, as predicted. We next verify the benefits of using joint sparsity over a simple taking of the union of sparse subsets recovered independently for each value of T. Figure 4 compares our approach with the independent sparsity approach (also using nonconvex regularization) for $q = 10$. Note that our algorithm significantly outperforms the independent sparsity approach. Figure 7 in the supplement shows results for $q = 40$, showing that as q increases, the number of samples required in fact decreases slightly (since in our experimental setup $\|\theta\|_{1,2}$ grows in expectation as $\sqrt{q}$). Additionally, Figure 8 in the supplement compares results when the the co-
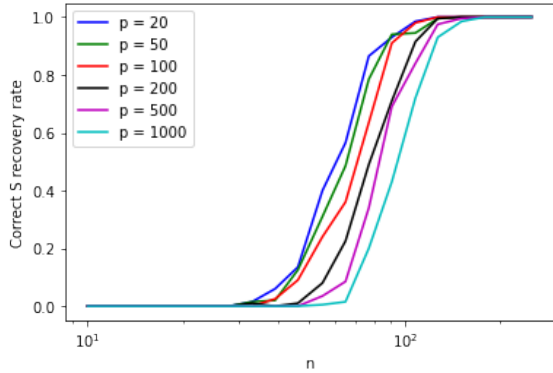
Figure 3: Empirical probability of our algorithm correctly recovering the sparse set S as a function of n and p, for binary actions ($q = 2$).
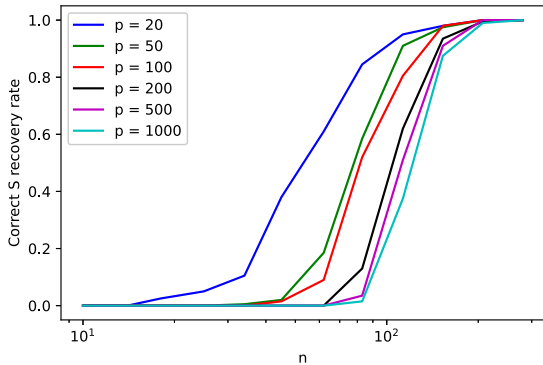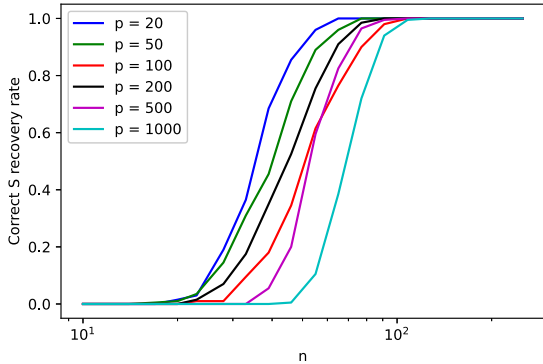

Figure 4: Empirical probability of our joint sparse algorithm (upper) and independent sparsity approach (lower) correctly recovering the sparse set S as a function of n and p, for $q = 10$.

efficients in S are i.i.d. Gaussian for all t values, and when only $t = 0$ has nonzero coefficients in S. This latter setting would be the most amenable to the individual sparsity approach, yet our method still matches the performance of individual sparsity in this regime.
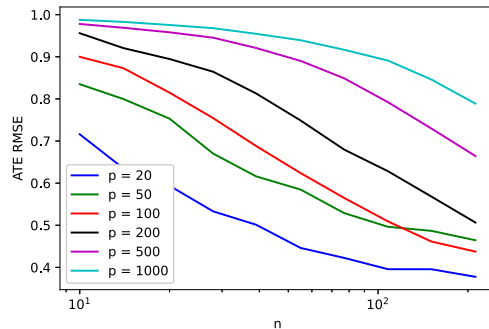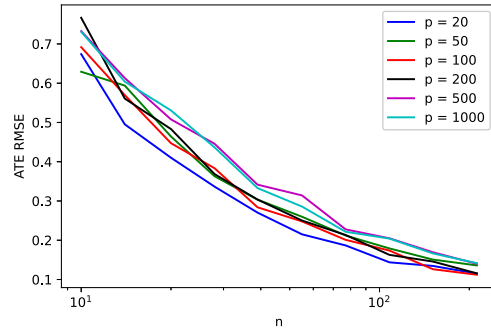

Figure 5: Empirical RMSE of estimating the causal ATE of the second action over the first action (true effect 1.0) for our joint sparse algorithm (upper) and the full-cardinality estimate with no variable selection (lower), for $q = 10$. The sparse approach uses the same regularization as Figure 4. Note that variable selection results in a estimate that converges faster to the true value, particularly for higher p.

Figures 5 and 6 show the RMSE for estimating the average treatment effect (ATE) and the average error of estimating the individual treatment effect (ITE) respectively, averaged over 400 random trials. The ITE error is the error averaged over predicting the ITE for each value of X, T in the synthetic dataset. Notice that as expected, our variable selectin approach results in faster convergence of the estimate, and the approach not using variable selection is more unstable and converges more slowly.

5.2 Real Datasets

Cattaneo2: effect of smoking on birth weight. This dataset[6] [Abadie and Imbens, 2006] studies the effect of maternal smoking on babies' birth weight in grams, and consists of 4642 singleton births in Pennsylvania, US. Actions are 0: no smoking (3778 samples), 1: 1-5 cigarettes daily (200 samples), 2: 6-10 cigarettes daily (337 samples), and 3: 11 or more cigarettes daily (327 samples). 20 covariates are in-

---

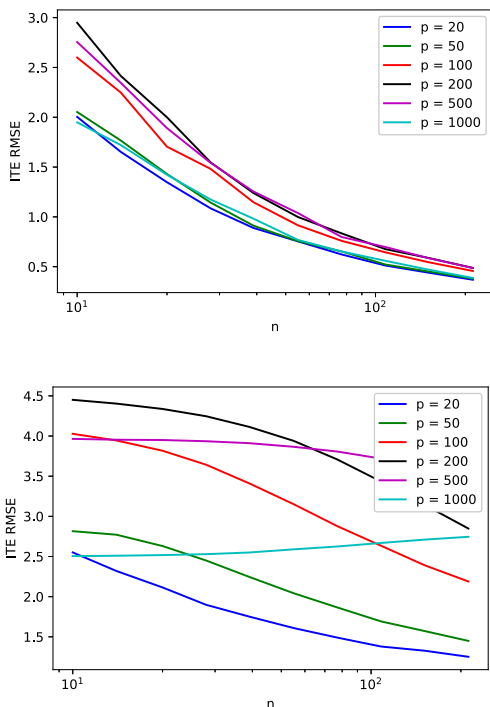[6]http://www.stata-press.com/data/r13/cattaneo2.dta

Figure 6: Average empirical RMSE of estimating the causal ITE of the second action over the first action for our joint sparse algorithm (upper) and the full-cardinality estimate with no variable selection (lower), for $q = 10$. Note that variable selection results in a estimate that converges faster to the true value, particularly for higher p.

cluded. Results comparing (nonsparse) doubly robust effect estimates [Shimoni et al., 2019] and the effect estimates obtained by using the doubly robust estimator on the sparse set S obtained by our method are shown in Table 1. We randomly split the data to have 20% used for the S estimation and 80% used for the effect estimation. Our sparse approach is tuned via cross validation and on average yields a sparse S of cardinality 10.9 (out of $|X| = 20$). Note that the sparse approach, unlike the full approach, yields a binary effect estimate consistent within the known empirical estimated interval [Abadie and Imbens, 2006]. For additional method comparisons for the binary effect, see Figure 5 in [Cheng et al., 2020] - only the dimensionality reduction method of [Cheng et al., 2020] provides an estimate in the empirical interval as we do.

IHDP: Effects of high-intensity care on low birth rate and premature infants. This semi-synthetic dataset[7] [Hill, 2011] consists of data on 25 covariates and an assigned treatment variable indicating whether the child was assigned to high-intensity

---

[7]https://github.com/vdorie/npci

	Nonsparse Doubly Robust Estimate	Sparse DR Estimate (Ours)
Effect of 1 vs. 0	-151.4g(21.3)	-195.0g(28.6)
Effect of 2 vs. 0	-161.9g(16.6)	-264.0g(34.2)
Effect of 3 vs. 0	-189.2g(21.1)	-236.0g(26.6)
Binary effect(> 0 vs 0)	-162.4g(8.5)	-239.3g(10.8)

Table 1: Estimated average treatment effects on Cattaneo2 dataset. Actions – 0: no smoking, 1: 1-5 cigarettes daily, 2: 6-10 daily, and 3: 11 or more. For binary action effect, the empirical estimated interval is known to be (-250g, -200g). Standard deviations over 20 random data splits are given in parentheses.

care. Following the procedure in [Hill, 2011], the treated and non-treated populations are biased and a response variable is generated according to the "A" scheme therein (which is designed to have sparse edges from X to Y). Since the response is generated synthetically, the true ATE is known to be 4.36. Results for the non-sparse doubly robust estimator and the doubly robust estimator applied to the sparse S recovered by our approach are shown in Table 2. Both methods work reasonably well, with our sparse version outperforming (selecting on average $|S| = 6.4$). The sparse performance is on par with the well-performing methods with results shown in Figure 2 of [Cheng et al., 2020].

	Nonsparse Doubly Robust Estimate	Sparse DR Estimate (Ours)
ATE	4.76(.51)	4.49(.57)

Table 2: Estimated average treatment effects on semi-synthetic IHDP dataset, with standard deviations over 20 random trials in parentheses. The true ATE is 4.36.

Additional details and real data experiments are in the supplement.

6 Conclusion

We considered using sparse regression to reduce the sample complexity of estimating causal effects in the presence of large numbers of covariates. We presented an algorithm based on joint-sparsity promoting non-convex regularization, proved that it correctly recovers the sparse support S with high probability, and tested it experimentally. In future work, we plan to use the power of the joint RSC concept to generalize our sparse estimator to more flexible nonlinear settings and to losses for categorical outcomes.

References

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

[Athey and Imbens, 2015] Athey, S. and Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26.

[Bradic et al., 2019] Bradic, J., Wager, S., and Zhu, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*.

[Cheng et al., 2020] Cheng, D., Li, J., Liu, L., and Liu, J. (2020). Sufficient dimension reduction for average causal effect estimation. *arXiv preprint arXiv:2009.06444*.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

[Guo et al., 2020] Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

[Hill, 2011] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[Huang et al., 2010] Huang, J., Zhang, T., et al. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004.

[Imbens and Wooldridge, 2009] Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

[Kallus, 2018] Kallus, N. (2018). Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*.

[Kallus et al., 2019] Kallus, N., Mao, X., and Zhou, A. (2019). Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2281–2290.

[Kuenzel, 2019] Kuenzel, S. R. (2019). *Heterogeneous Treatment Effect Estimation Using Machine Learning*. PhD thesis, UC Berkeley.

[Loh and Wainwright, 2015] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616.

[Loh and Wainwright, 2017] Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482.

[Louizos et al., 2017] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

[Lounici et al., 2011] Lounici, K., Pontil, M., Van De Geer, S., Tsybakov, A. B., et al. (2011). Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204.

[Pearl, 2009] Pearl, J. (2009). *Causality*. Cambridge university press.

[Shah and Bühlmann, 2019] Shah, R. D. and Bühlmann, P. (2019). Double-estimation-friendly inference for high-dimensional misspecified models. *arXiv preprint arXiv:1909.10828*.

[Shalit et al., 2017] Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

[Shimoni et al., 2019] Shimoni, Y., Karavani, E., Ravid, S., Bak, P., Ng, T. H., Alford, S. H., Meade, D., and Goldschmidt, Y. (2019). An evaluation toolkit to guide model selection and cohort definition in causal inference. *arXiv preprint arXiv:1906.00442*.

[Shortreed and Ertefaie, 2017] Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

[Shpitser and Pearl, 2012] Shpitser, I. and Pearl, J. (2012). Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876*.

[Vershynin, 2010] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

[Vershynin, 2012] Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

[Wager and Athey, 2018] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

[Witte et al., 2020] Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020). On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45.

[Yadlowsky et al., 2018] Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

[Yao et al., 2018] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643.

[Yoon et al., 2018] Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

[Zhang et al., 2010] Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.