Sample Constrained Treatment Effect Estimation

Raghavendra Addanki RADDANKI@ADOBE.COM
Adobe Research

David Arbour DARBOUR26@GMAIL.COM
Adobe Research

Tung Mai TUMAI@ADOBE.COM
Adobe Research

Cameron Musco CMUSCO@CS.UMASS.EDU
University of Massachusetts Amherst

Anup Rao ANUPRAO@ADOBE.COM
Adobe Research

Abstract

Treatment effect estimation is a fundamental problem in causal inference. We focus on designing efficient randomized controlled trials, to accurately estimate the effect of some treatment on a population of n individuals. In particular, we study *sample-constrained treatment effect estimation*, where we must select a subset of $s \ll n$ individuals to experiment on. This subset must be further partitioned into treatment and control groups. Algorithms for partitioning the full population into treatment and control groups, or for choosing a single representative subset, have been well-studied. The key challenge in our setting is jointing choosing a representative subset and a partition for that set.

We focus on both individual and average treatment effect estimation, under a linear effects model. We give provably efficient experimental designs and corresponding estimators, by identifying connections to discrepancy minimization and leverage-score-based sampling used in randomized numerical linear algebra. Our theoretical results obtain a smooth transition to known guarantees when s equals the population size. We also empirically demonstrate the performance of our algorithms.

Keywords: Treatment effect estimation, experimental design, causal inference.

1. Introduction

Experimentation has long been held as a gold standard for inferring causal effects since one can explicitly enforce independence between treatment assignment and other variables which influence the outcome of interest. We consider the the finite population setting of the potential outcomes framework [36, 39], where each individual is associated with a control and treatment value, also called the *potential outcomes*, and based on the treatment assignment, we can observe only one of these values. In the absence of assumptions on the functional form of the potential outcomes, the minimax optimal approach for conducting an experiment is to assign individuals to treatment or control completely at random, without consideration of baseline covariates (features) [24]. However, by considering covariates for each individual, and using additional assumptions of smoothness, substantial gains can be made in terms of the variance of the treatment effect estimate via alternative assignment procedures. The most common approach attempts to minimize imbalance, i.e., the difference between the baseline covariates in the treatment and control groups [6, 24, 34].

While experimental designs that minimize imbalance increase the power of an experiment for a given pool of subjects, there are many practical applications where the experimenter wishes to minimize the total number of subjects who are placed into the experiment. For example, in medicine, clinical trials may carry nontrivial risk to patients. Within industrial applications, experiments may carry substantial costs in terms of testing changes, which decrease the quality of the user experience, or have direct monetary costs.

In this paper, we examine the problem of selecting a subset of s individuals from a larger population and assigning treatments such that the estimated treatment effect has a small error. We consider two different estimands: individual treatment effect (ITE) and average treatment effect (ATE).

We represent the d-covariates of a population of n individuals using $\mathbf{X} \in \mathbb{R}^{n \times d}$. We assume that the treatment and control values, denoted by $\mathbf{y}^1, \mathbf{y}^0 \in \mathbb{R}^n$, are linear functions of the covariates. Under the linearity assumption, the treatment and control values are a linear function of the covariates. Formally, for some $\boldsymbol{\beta}^0, \boldsymbol{\beta}^1 \in \mathbb{R}^d$,

$$\mathbf{y}^1 = \mathbf{X}\boldsymbol{\beta}^1 + \boldsymbol{\zeta}^1 \text{ and } \mathbf{y}^0 = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\zeta}^0,$$

where $\boldsymbol{\zeta}^1, \boldsymbol{\zeta}^0 \in \mathbb{R}^n$ are noise vectors, with each coordinate drawn independently from the Gaussian distribution with zero mean and variance σ^2, i.e., $N(0, \sigma^2)$.

The ITE for the i^{th} individual is $\mathbf{y}_i^1 - \mathbf{y}_i^0$ and ATE is the average of all the ITE values. The goal is to pick a subset of s individuals and partition this subset into control and treatment groups. For an individual i in the treatment group, we measure $\mathbf{y}_i^1$, and for an individual j in the control, we measure $\mathbf{y}_j^0$. From this small set of measurements, we seek to estimate the ITE or ATE over the full population. In Appendix G, we provide additional related work and discuss the main differences to the active learning setting.

Our Contributions. We make the SUTVA assumption, i.e., the treatment outcome value of any individual is independent of treatment assignments of others in the population [46]. For ITE estimation, we propose an algorithm using *leverage score sampling* [49], which is a popular approach to subset selection for fast linear algebraic computation. For ATE estimation, we employ a recursive application of a covariate balancing design [19].

For ITE estimation, we give a randomized algorithm that selects $\Theta(d \log d)$ individuals in expectation, using leverage scores, which measure the importance of an individual based on their covariates. Our algorithm obtains, with high probability, root mean squared error $O\Big(\sqrt{\log d/n} \cdot (\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|) + \sigma\Big)$ (see Corollary 3.2). We argue that this is optimal up to constants and a $\sqrt{\log d}$ factor, *even for approaches that experiment on the full population.*

The key challenge in achieving this bound is to extend leverage scores to our simultaneous linear regression setting, ensuring that we do not share samples across the treatment and control effect estimation problems. To do this, we introduce a *smoothed* covariate matrix, whose leverage scores are bounded. This ensures that, when applying independent leverage score sampling, with high probability few individuals are randomly assigned to both control and treatment, and thus removing such individuals from one of the groups does not introduce too much error.

For ATE estimation we give a randomized algorithm that selects at most s individuals for treatment/control assignment and obtains an error of $\widetilde{O}\left(\sigma/\sqrt{s} + (\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|)/s\right)$, where

$\widetilde{O}(\cdot)$ hides logarithmic factors (see Theorem 4.1). The error decreases with increasing values of s and when $s = n$, it matches state-of-the-art guarantees due to Harshaw et al. [19].

Our algorithm for ATE estimation is based on *covariate balancing*. This is a popular approach where one attempts to assign similar individuals to the treatment and control groups, to ensure that the observed effect is attributed to the administered treatment alone. Harshaw et al. [19] designed an algorithm by minimizing the discrepancy of an augmented covariate matrix, which achieves low ATE estimation error. To extend their approach to our setting, first, we need to select a subset of s individuals that are representative of the entire population, and then balance the covariates. Uniform sampling or importance sampling techniques give high error here. Instead, we employ a recursive strategy, which repeatedly partitions the individuals into two subsets by balancing covariates, and selects the smaller subset to recurse on, until we have selected at most s individuals.

We observe that our techniques for ITE and ATE estimation should extend to the setting when the outcomes are non-linear functions of the covariates, which are linear in some higher-dimensional kernel space. This is immediate for our discrepancy minimization design for ATE, which only requires knowing the pairwise inner products of the covariate vectors. For ITE estimation, leverage score sampling for kernel ridge regression [3] is most likely applicable. Extensions to broader classes of non-linear models are beyond the scope of this work, but they are an interesting future direction.

Finally, in Section 5, we provide an empirical evaluation of the performance of our ITE and ATE estimation methods, comparing against uniform sampling and other baselines on several datasets. Our results suggest that our techniques can help reduce the costs associated with running randomized controlled trials using only a small fraction of the population.

2. Preliminaries

Notation. For a population of n individuals, we represent each with an integer in $[n]$ where we denote $[n] \stackrel{\text{def}}{=} \{1, 2, \cdots, n\}$. We use bold capital letters, e.g., $\mathbf{X}$ to denote matrices and bold lowercase letters, e.g., $\mathbf{y}$ to denote vectors. We use $\mathbf{X}[i, :]$ and $\mathbf{X}[:, j]$ to denote the i^{th} row and j^{th} column of $\mathbf{X}$ respectively, which we always view as column vectors. We assume that $\mathbf{X}$ is row-normalized, i.e., $\|\mathbf{X}[i, :]\| \leq 1\ \forall i \in [n]$. The i^{th} largest singular value of $\mathbf{X}$ is denoted by $\sigma_i(\mathbf{X})$. For any vector $\mathbf{x}$, the Euclidean norm or the ℓ_2-norm is denoted by $\|\mathbf{x}\|$. The leverage score of j^{th} row $\mathbf{X}[j, :]$, denoted by $\ell_j(\mathbf{X})$, is defined as: $\ell_j(\mathbf{X}) \stackrel{\text{def}}{=} \mathbf{X}[j, :]^\top (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}[j, :]$, where $^+$ denotes the Moore–Penrose pseudo-inverse.

Definition 2.1 (Root Mean Squared Error)**.** *For a set of estimated individual treatment effects,* $\widehat{\text{ITE}}(j)$ *for* $j \in [n]$*, the root mean squared error (*RMSE*) is defined as:*

$$\text{RMSE} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \cdot \left\| \widehat{\text{ITE}}(j) - \text{ITE}(j) \right\|.$$

3. Individual Treatment Effect Estimation

We now describe our algorithm for ITE estimation. The algorithm identifies a subset of the population to experiment on, using *importance based sampling* techniques, that are well-studied in randomized numerical linear algebra [49]. Missing details are in Appendix H.

Overview of our approach. Under the linearity assumption, we can reformulate the problem of estimating the ITE for every individual as simultaneously solving two linear

regression instances: one for control and one for treatment, i.e., we regress $\mathbf{y}^0, \mathbf{y}^1$ on $\mathbf{X}$. Intuitively, we want to select s individuals (or equivalently rows) that capture the entire row space of $\mathbf{X}$ and use them to estimate the ITE of all other individuals. Leverage scores capture the importance of a row in making up the row space. E.g., if a row is orthogonal to all the other rows, it's leverage score will be the maximum value of 1.

Unfortunately, if we apply leverage score sampling independently to the regression problems for $\mathbf{y}^0$ and $\mathbf{y}^1$, rows with high leverage leverage scores may be sampled for both instances. This presents a problem, since we can only read at most one of $\mathbf{y}^0_j$ or $\mathbf{y}^1_j$. To mitigate this issue, we construct a *smoothed* matrix $\mathbf{X}^*$, which consists of $\mathbf{X}$ projected onto its singular vectors with high singular values. Intuitively, this dampens the effects of high leverage score 'outlier' rows that don't contribute significantly to the spectrum of $\mathbf{X}$. Formally, we prove that the maximum leverage score of $\mathbf{X}^*$ is bounded, which let's us solve our two regression problems via independent sampling. There will be few repeated samples across our subsets, which introduce minimal error.

Algorithm Sampling-ITE. We perform row sampling twice, with probabilities, given by $\boldsymbol{\pi} \in \mathbb{R}^n$, proportional to the leverage scores of $\mathbf{X}^*$, to construct two sets S^0, S^1. These two sets are used to estimate the vectors $\mathbf{y}^0$ and $\mathbf{y}^1$, respectively. Missing details about the exact values of probabilities are included in Appendix I.

It is possible that a row gets included in both S^0 and S^1. In that case, we simple remove the row from S^1. As a result, j^{th} row is included in S^1 with probability $\boldsymbol{\pi}_j \cdot (1 - \boldsymbol{\pi}_j)$ for every $j \in [n]$. We construct sampling matrices $\mathbf{W}^0$ and $\mathbf{W}^1$ using probabilities $\boldsymbol{\pi}$ and $\boldsymbol{\pi}(1-\boldsymbol{\pi})$ respectively. Finally, we solve the following linear regressions, for $i = 0, 1$ separately: $\widetilde{\boldsymbol{\beta}}^i = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\| \mathbf{W}^i \mathbf{X}^* \boldsymbol{\beta} - \mathbf{W}^i \mathbf{y}^i \right\|^2$.

Our estimate for each $\mathrm{ITE}(j)$, denoted by $\widehat{\mathrm{ITE}}(j)$ is set to j^{th} entry of $\mathbf{X}^* \widetilde{\boldsymbol{\beta}}^1 - \mathbf{X}^* \widetilde{\boldsymbol{\beta}}^0$. Observe that $S^0 \cap S^1$ is empty. This ensures that we have access to only one of $\mathbf{y}^0_j$ or $\mathbf{y}^1_j$ for any individual j in solving the above two subsampled regressions. Formally, we obtain:

Theorem 3.1. *Suppose $s \geq 120 c_0 d \log d$. There is a randomized algorithm that selects a subset $S \subseteq [n]$ of the population with $\mathbf{E}[|S|] \leq s$, and, with probability at least $9/10$, returns ITE estimates $\widehat{\mathrm{ITE}}(j)$ for all $j \in [n]$ with error:*

$$\mathrm{RMSE} = O\Big(\sqrt{\frac{1}{n} \max\Big\{\frac{s}{d}, \log d\Big\}} \cdot (\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|) + \sigma\Big).$$

The corollary below follows immediately from Theorem 3.1.

Corollary 3.2 (Main ITE Error Bound)**.** *The root mean squared error obtained by Algorithm 1 is minimized when $s = \Theta(d \log d)$ and is given by:*

$$\mathrm{RMSE} = O\Big(\sqrt{\frac{\log d}{n}} \cdot (\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|) + \sigma\Big).$$

Our upper bound on RMSE increases with s, if s grows strictly faster than $d \log d$ asymptotically, i.e., $s = \omega(d \log d)$. Therefore, to obtain low error, we set $s = c \cdot d \log d$ for some constant c, even if the sample constraint allows for larger values. We believe this is an artifact of our analysis; in Section 5, we observe empirically that the error decreases with s.

4. Average Treatment Effect Estimation

In this section, we describe our approach for estimating ATE by building upon a recent work on efficient experimental design [19]. Missing details are in Appendix I.

Horvitz-Thompson Estimator. Suppose $\mathbf{S}^+ \subseteq [n]$ is the population assigned to the treatment group and $\mathbf{S}^- = [n] \setminus \mathbf{S}^+$ is the remaining population, i.e., the control group. A well-studied estimator for estimating the average treatment effect is the Horvitz-Thompson estimator [21], denoted by $\widehat{\tau}$. If every individual is assigned to $\mathbf{S}^+$ (or $\mathbf{S}^-$) with probability 0.5, then, $\widehat{\tau}$ is defined as follows: $\widehat{\tau} = \frac{2}{n}\left(\sum_{i\in\mathbf{S}^+} \mathbf{y}_i^1 - \sum_{i\in\mathbf{S}^-} \mathbf{y}_i^0\right)$.

Overview of Recursive-Covariate-Balancing. Our main idea is to partition the population using the Gram-Schmidt-Walk design (GSW) recursively until the total size of population that we can experiment on reduces to s. The Gram-Schmidt-Walk design produces a random partition of the population with a good balance of covariates in every dimension. In each recursive call, we start by partitioning the available individuals $\mathbf{Z}_t$ into treatment and control groups, denoted by $\mathbf{Z}_t^+, \mathbf{Z}_t^-$ using GSW. Next, we identify the smaller of these two subsets, say $\mathbf{Z}_t^+$ and recurse on $\mathbf{Z}_t^+$. We stop after k recursive calls when there are only s individuals to experiment on, i.e., $|\mathbf{Z}_k^+ \cup \mathbf{Z}_k^-| \leq s$. Finally, we construct our estimator $\widehat{\tau}_s$, similar to the Horvitz-Thompson estimator, by scaling the treatment and control contributions due to $\mathbf{Z}_k^+$ and $\mathbf{Z}_k^-$ as: $\widehat{\tau}_s = 2^t/n \cdot \left(\sum_{j\in\mathbf{Z}_t^+} \mathbf{y}_j^1 - \sum_{j\in\mathbf{Z}_t^-} \mathbf{y}_j^0\right)$.

Theoretical Guarantees. Our analysis approach, inspired by the coreset construction for discrepancy minimization [25], is based on the observation that if we can obtain good estimates for the contributions $\sum_{i\in[n]} \mathbf{y}_i^1$ and $\sum_{i\in[n]} \mathbf{y}_i^0$, we obtain a good estimate for ATE.

Theorem 4.1 (Main ATE Error Bound)**.** *The estimator $\widehat{\tau}_s$ in Algorithm* Recursive-Covariate-Balancing *obtains the following guarantee, with probability at least 2/3:*

$$|\widehat{\tau}_s - \tau| = O\left(\sqrt{\log\log(n/s)} \cdot \left(\frac{\sigma}{\sqrt{s}} + \frac{\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|}{s}\right)\right).$$

Remark. When $s = n$, the above theorem matches the guarantees obtained by GSW design [19]. Moreover, we obtain a better dependence compared to sampling s rows uniformly at random and using the $\mathbf{y}^1, \mathbf{y}^0$ values of the sampled rows to estimate the population mean of treatment and control groups in ATE. An application of standard concentration inequalities or the central limit theorem, will yield a multiplicative factor increase in one of the error terms, with a dependence of $\widetilde{O}\left(1/s \cdot \|\mathbf{X}\|_2 \left(\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|\right)\right)$, instead of the $\widetilde{O}\left(1/s \cdot \left(\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|\right)\right)$ obtained by our algorithm, where $\|\mathbf{X}\|_2$ denotes the spectral norm of $\mathbf{X}$ and $\widetilde{O}(\cdot)$ hides the logarithmic factors.

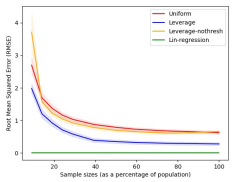
5. Experimental Evaluation

In this section, we provide an evaluation of our algorithms on five datasets: IHDP [20, 13], Twins [5], Lalonde [28], Boston [18], and Synthetic [32]. Missing details are in Appendix J.

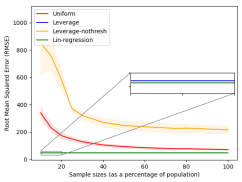
Baselines. (A) ITE. We compare the performance of our Algorithm Sampling-ITE (referred to as *'Leverage'*) with respect to three baselines: *(i) Uniform* – where we use uniform sampling of $\mathbf{X}$, *(ii) Leverage-nothresh* – where we use our Algorithm Sampling-ITE on $\mathbf{X}$, instead of $\mathbf{X}^*$, *(iii) Lin-regression* – which captures the best linear fit regression error, i.e., assuming we have access to both $\mathbf{y}^1, \mathbf{y}^0$. **(B) ATE.** We compare the performance

of our Algorithm RECURSIVE-COVARIATE-BALANCING (referred to as *'Recursive-GSW'*) to three baselines: *(i) Uniform* – in which we sample s rows uniformly and assign them to treatment or control with equal probability, *(ii) GSW-pop* – we partition the entire population using GSW, *(iii) Complete Randomization* – we partition the population with equal probability. The last two baselines are over all the n individuals.

Evaluation. To evaluate the performance of average treatment effect estimation (τ) on the datasets, we compare the deviation error of the estimator $\widehat{\tau}_s$, given by $|\widehat{\tau}_s - \tau|$ for different sample sizes. To evaluate the performance of individual treatment effect estimates, we compare the root mean squared error RMSE (see Defn. 2.1) for different sample sizes.

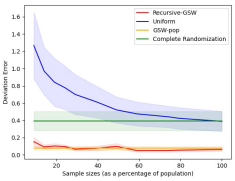


(a) IHDP

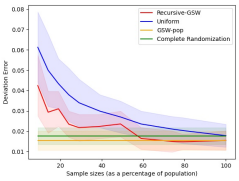


(b) Synthetic

Figure 1: We compare the performance of various methods for estimating ITE, measured using RMSE on y-axis, against different sample sizes (as proportion of dataset size) on x-axis.



(a) IHDP



(b) Twins

Figure 2: We compare the performance of various methods for estimating ATE, measured using deviation error on y-axis, against different sample sizes (as proportion of dataset size) on x-axis.

Results. For every dataset, we run each experiment for 1000 trials and plot the mean using a colored line and shade the region between 30 and 70 percentile around the mean to signify the *confidence interval* in Figs. 2, 1 representing ATE and ITE results.
(i) **ITE**. For all sample sizes, we observe that the RMSE obtained by our algorithm labeled as *Leverage* in Figure 1, is significantly smaller than that of all the other baselines, including *Uniform* and *Leverage-nothresh*. E.g., we observe that when the sample size is 20% of the population in IHDP dataset, the error obtained by *Leverage* is at least 50% times smaller than that of *Uniform* and *Leverage-nothresh*. For the Synthetic dataset, the error obtained by *Leverage* is extremely close to that of the error due to the best linear fit, *Lin-regression* (see the zoomed in part of the figure). Our algorithms result in a reduction of experimental costs for ITE estimation using only a fraction of the dataset. (ii) **ATE**. For all datasets, we observe that the deviation error obtained by our algorithm labeled as *Recursive-GSW* in Figure 2, is significantly smaller than that of *Uniform* baseline. Surprisingly, for the IHDP dataset, our approach is significantly better than *Complete-randomization*, for all sample sizes, including using just 10% of data. For all the remaining datasets using a sample of size 30%, we achieve the same error (up to the confidence interval) as that of *Complete-randomization*. Complete randomization is one of the most commonly used methods for experimental design and our results indicate a substantial reduction in experimental costs. For IHDP dataset, a sample size of about 10% of the population is sufficient to achieve a similar error as that of *GSW-pop*. For the remaining datasets, we observe that for sample sizes of about 30% of the population, the deviation error obtained by our algorithm is within the shaded confidence interval of the error obtained by *GSW-pop*. Therefore, for a specified error tolerance level for ATE, we can reduce the associated experimental costs using just a small subset of the dataset using our algorithm.

Acknowledgements. Most of this work was done while R. Addanki was a student at UMass Amherst. Part of this work was done while R. Addanki was a visiting student at the Simons Institute for the Theory of Computing. This work was supported by a Dissertation Writing Fellowship awarded by the Manning College of Information and Computer Sciences, UMass Amherst to R. Addanki. In addition, this work was supported by NSF grants CCF-2046235, IIS-1763618, as well as Adobe and Google Research Grants, awarded to C. Musco; and NSF grants CCF-1934846, CCF-1908849, and CCF-1637536, awarded to Andrew McGregor.

References

[1] Raghavendra Addanki, Shiva Kasiviswanathan, Andrew McGregor, and Cameron Musco. Efficient intervention design for causal discovery with latents. In *International Conference on Machine Learning*, pages 63–73. PMLR, 2020.

[2] Raghavendra Addanki, Andrew McGregor, and Cameron Musco. Intervention efficient algorithms for approximate learning of causal graphs. In *Algorithmic Learning Theory*, pages 151–184. PMLR, 2021.

[3] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in neural information processing systems*, 28, 2015.

[4] Ayya Alieva, Ashok Cutkosky, and Abhimanyu Das. Robust pure exploration in linear bandits with limited budget. In *International Conference on Machine Learning*, pages 187–195. PMLR, 2021.

[5] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[6] David Arbour, Drew Dimmery, and Anup Rao. Efficient balanced treatment assignments for experimentation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3070–3078. PMLR, 2021.

[7] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[8] Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–597, 2018.

[9] Steffen Bondorf, Binbin Chen, Jonathan Scarlett, Haifeng Yu, and Yuda Zhao. Sublinear-time non-adaptive group testing with o (k log n) tests via bit-mixing coding. *IEEE Transactions on Information Theory*, 67(3):1559–1570, 2020.

[10] Chun Lam Chan, Pak Hou Che, Sidharth Jaggi, and Venkatesh Saligrama. Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1832–1839. IEEE, 2011.

[11] Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Conference on Learning Theory*, pages 663–695. PMLR, 2019.

[12] Albert Cohen, Mark A Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834, 2013.

[13] Vincent Dorie. Npci: Non-parametrics for causal inference. *URL: https://github.com/vdorie/npci*, 2016.

[14] Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 2000.

[15] Frederick Eberhardt. Causation and intervention. *PhD Thesis, Carnegie Mellon University*, 2007.

[16] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733. PMLR, 2018.

[17] Robert Greevy, Bo Lu, Jeffrey H Silber, and Paul Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004.

[18] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

[19] Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments using the gram-schmidt walk. *arXiv preprint arXiv:1911.03071*, 2019.

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

[21] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[22] Kosuke Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in medicine*, 27(24):4857–4873, 2008.

[23] Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 07 2017.

[25] Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993. PMLR, 2019.

[26] Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

[27] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1875–1884. JMLR. org, 2017.

[28] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[29] Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

[30] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

[31] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[32] Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning*, pages 91–99. PMLR, 2014.

[33] Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[34] Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

[35] Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2021.

[36] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

[37] Alexander G Nikolaev, Sheldon H Jacobson, Wendy K Tam Cho, Jason J Sauppe, and Edward C Sewell. Balance optimization subset selection (boss): An alternative approach for causal inference with observational data. *Operations Research*, 61(2):398–412, 2013.

[38] Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Budgeted heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pages 8693–8702. PMLR, 2021.

[39] Donald B Rubin. Comment: Randomization analysis of experimental data. *Journal of the American Statistical Association*, 75(371):591, 1980.

[40] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

[41] Burr Settles. Active learning literature survey. 2009.

[42] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[43] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015.

[44] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[45] Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. Active learning for decision-making from imbalanced observational data. In *International Conference on Machine Learning*, pages 6046–6055. PMLR, 2019.

[46] Stefan Wager. Stats 361: Causal inference. 2020.

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

[48] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

[49] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

F. Preliminaries

Assumption F.1 (Linearity Assumption)**.** *Under the linearity assumption, the treatment and control values are a linear function of the covariates. Formally, for some* $\boldsymbol{\beta}^0, \boldsymbol{\beta}^1 \in \mathbb{R}^d$*,*

$$\mathbf{y}^1 = \mathbf{X}\boldsymbol{\beta}^1 + \boldsymbol{\zeta}^1 \text{ and } \mathbf{y}^0 = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\zeta}^0,$$

where $\boldsymbol{\zeta}^1, \boldsymbol{\zeta}^0 \in \mathbb{R}^n$ *are noise vectors, with each coordinate drawn independently from the Gaussian distribution with zero mean and variance* σ^2*, i.e.,* $N(0, \sigma^2)$*. We further assume that* $\mathbf{X}$ *is row-normalized, i.e.,* $\|\mathbf{X}[i,:]\| \leq 1 \ \forall i \in [n]$*.*

Definition F.2 (Individual Treatment Effect)**.** *Given a population of* n *individuals, the individual treatment effect (ITE) of* $j \in [n]$ *is the difference between the treatment and control values:*

$$\mathrm{ITE}(j) \stackrel{\text{def}}{=} \mathbf{y}^1_j - \mathbf{y}^0_j.$$

Definition F.3 (Average Treatment Effect)**.** *Given a population of* n *individuals, the average treatment effect (ATE), denoted by* τ*, is the average individual treatment effect:*

$$\tau \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j \in [n]} \mathrm{ITE}(j) = \frac{1}{n} \sum_{j \in [n]} \mathbf{y}^1_j - \mathbf{y}^0_j.$$

The following claim about leverage scores is well known.

Claim F.4 ([49])**.** $\sum_{j \in [n]} \ell_j(\mathbf{X}^*) = \mathit{rank}(\mathbf{X}^*)$*.*

We will employ the following well-known result on the ℓ_2-norm of a Gaussian vector.

Fact F.5. *Suppose* $\boldsymbol{\zeta} \in \mathbb{R}^n$ *be a vector such that each co-ordinate* $\boldsymbol{\zeta}_i$ *is drawn independently from the normal distribution* $N(0, \sigma^2)$*. Then, with probability* $\geq 1 - 1/n$*:*

$$\|\boldsymbol{\zeta}\| \leq 2\sigma \cdot \sqrt{n}.$$

G. Related Work

Parametric Assumptions. Without parametric assumptions, ITE estimation is not feasible [42]. We focus on linear models in particular, since they are important in developing theory. E.g., in the literature on optimal designs in active learning, much of the foundational theory is built around linear models. Identifying estimators based on linearity assumptions is an active area of study in the causal inference literature [19, 48].

Active Learning. Our setup is similar to active learning [41], where the goal is to minimize the number of individual labels that we access for solving linear regression or other downstream tasks. They key difference is that we must both select a subset of individuals, and for each i, can measure only one of two labels: $\mathbf{y}^1_i$ or $\mathbf{y}^0_i$. In particular, ITE estimation can be thought of as solving two *simultaneous* active linear regression problems – one for the treatment outcomes and one for the control outcomes. Thus, standard active learning-based approaches, such as [11, 12, 33], fall short. Even when s equals the population size n, i.e., when active learning becomes trivial, our problem does not. We must still pick a

partition of the full population into treatment and control groups. Overall, sample constrained treatment effect estimation by designing efficient randomized controlled trials has received little attention, compared to various approaches that use observational data, such as [23, 38, 45].

Other Related Work. For ATE estimation, the most well-studied approaches to experiment design are covariate balancing and randomization. A variety of design techniques have been studied based on these approaches, such as blocking [17], matching [22, 44], rerandomization [29, 34], and optimization [24]. Using observational data, treatment effect estimation using covariate regression adjustment [30] and various active learning-based sampling techniques have gained recent attention [23, 37, 45]. Compared to ATE, estimating ITE is significantly harder and has received attention only recently using machine learning methods [7, 42, 47]. There has been a lot of recent work on efficient experimental designs to minimize experimental costs, in various domains, such as causal discovery [1, 2, 15, 16, 27, 43], multi-arm bandits [4, 26, 35], and group testing [9, 10, 14].

H. Individual Treatment Effect Estimation

In this section, we present the missing details from section 3.

For some $\gamma \geq 0$, to be fixed later, we define a *smoothed* matrix for $\mathbf{X}$, the projection onto singular vectors with high singular values, as follows:

Definition H.1 (Smoothed matrix)**.** *Given $\mathbf{X} \in \mathbb{R}^{n \times d}$ with singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, let Γ^* be the set of indices corresponding to singular values greater than $\sqrt{\gamma}$, i.e., $\Gamma^* \stackrel{\text{def}}{=} \{i \mid \sigma_i(\mathbf{X}) \geq \sqrt{\gamma}\}$; we denote $d' \stackrel{\text{def}}{=} |\Gamma^*|$. Let $\mathbf{\Sigma}^* = \mathbf{\Sigma}(\Gamma^*, \Gamma^*)$ denote the principal sub-matrix of $\mathbf{\Sigma}$ associated with these large singular values. Similarly, let $\mathbf{U}^* \in \mathbb{R}^{n \times d'}, \mathbf{V}^* \in \mathbb{R}^{d \times d'}$ be the associated column sub-matrices of $\mathbf{U}$ and $\mathbf{V}$. Then, we define: $\mathbf{X}^* \stackrel{\text{def}}{=} \mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*T}$.*

H.1 Leverage Score Sampling

Sampling Matrix. Our algorithm Sampling-ITE, will sample individuals, corresponding to rows of the smoothed matrix of $\mathbf{X}$, i.e., $\mathbf{X}^*$, independently – the i^{th} row is included in the sample with some probability $\boldsymbol{\pi}_i$. Let the set of rows sampled be denoted by S.

We can associate a sampling matrix $\mathbf{W}$ with S. The j^{th} row of $\mathbf{W}$ is associated with the j^{th} element in the set S (under some fixed order). If the j^{th} element in S is the row for individual i for some $i \in [n]$, then, $\mathbf{W}[j,:]$ is equal to $\mathbf{e}_i/\sqrt{\boldsymbol{\pi}_i}$. Here, $\mathbf{e}_i \in \mathbb{R}^n$ denotes the i^{th} standard basis vector. In this way, $\mathbf{W}\mathbf{X}^*$ consists of the subset of rows sampled in S, reweighted by the inverse squareroot of their sampling probabilities, which is necessary to keep expectations correct in solving the linear regression.

H.2 Theoretical Guarantees

First, we bound the error due to sampling. Critically, we show that the leverage scores of $\mathbf{X}^*$, and in turn the probabilities $\boldsymbol{\pi}$, are bounded by $1/\gamma$. Thus, the sampling probabilities for S^1, $\boldsymbol{\pi}(1-\boldsymbol{\pi})$ are not too far from $\boldsymbol{\pi}$ itself.

Algorithm 1 SAMPLING-ITE

Input: Smoothed covariates $\mathbf{X}^* \in \mathbb{R}^{n\times d}$, sampling probabilities $\boldsymbol{\pi} \in [0,1]^n$.
Output: Estimates for ITE(j) for each individual $j \in [n]$.

1: Add each $j \in [n]$ to set S^0 independently, with prob. $\boldsymbol{\pi}_j$.
2: Add each $j \in [n]$ to set S^1 independently, with prob. $\boldsymbol{\pi}_j$.
3: Construct sampling matrix $\mathbf{W}^0$ from S^0 using probabilities $\boldsymbol{\pi}$.
4: Construct sampling matrix $\mathbf{W}^1$ from $S^1 \setminus S^0$ using probabilities $\boldsymbol{\pi}(1-\boldsymbol{\pi})$.
5: Let $\widetilde{\boldsymbol{\beta}}^i = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\|\mathbf{W}^i\mathbf{X}^*\boldsymbol{\beta} - \mathbf{W}^i\mathbf{y}^i\right\|^2$ for $i = 0, 1$.
6: For each $j \in [n]$, let $\widehat{\mathrm{ITE}}(j)$ be the j^{th} entry of the vector $\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^1 - \mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0$.
7: **return** $\widehat{\mathrm{ITE}}(j)$ $\forall j \in [n]$.

We argue that the error introduced by ignoring small singular values and using $\mathbf{X}^*$ in place of $\mathbf{X}$ is small. Using $\mathbf{X}^*$ instead of $\mathbf{X}$ introduces error that depends on the threshold γ used in the construction of $\mathbf{X}^*$ (Def H.1).

Claim H.2. *For every $\boldsymbol{\beta} \in \mathbb{R}^d$, $\|\mathbf{X}^*\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}\| \le \sqrt{\gamma}\cdot\|\boldsymbol{\beta}\|$.*

Proof. Using the singular value decomposition of $\mathbf{X}, \mathbf{X}^*$:

$$\begin{aligned}\|\mathbf{X}^*\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}\| &= \|\mathbf{U}^*\boldsymbol{\Sigma}^*\mathbf{V}^*\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\boldsymbol{\beta}\| \\ &\le \|\mathbf{U}^*\boldsymbol{\Sigma}^*\mathbf{V}^* - \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\|_2 \cdot \|\boldsymbol{\beta}\|,\end{aligned}$$

where $\|\cdot\|_2$ denotes the spectral norm (the largest singular value) of the matrix. By construction, $\|\mathbf{U}^*\boldsymbol{\Sigma}^*\mathbf{V}^* - \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}\|_2 \le \sqrt{\gamma}$, giving the claim. □

We next argue that the leverage scores of the smoothed matrix $\mathbf{X}^*$ are bounded by $1/\gamma$. As we assume the row norms of $\mathbf{X}$ are bounded by 1, the row norms of $\mathbf{X}^*$ are also bounded. Thus, there can be no rows in $\mathbf{X}^*$ that are nearly orthogonal to all other rows – i.e., there can be no rows with very high leverage scores. Such rows would lead to small singular values. However, we know that the smallest singular value of $\mathbf{X}^*$ is at least $\sqrt{\gamma}$. In particular, we prove:

Claim H.3. *$\ell_j(\mathbf{X}^*) \le 1/\gamma$, for all $j \in [n]$.*

Proof. It is well known that $\ell_j(\mathbf{X}^*) = \mathbf{X}^*[j,:]^T(\mathbf{X}^{*T}\mathbf{X}^*)^+\mathbf{X}[j,:] = \|\mathbf{U}^*[j,:]\|^2$. This can be checked by writing $\mathbf{X}^*$ in its SVD. Further, $\|\mathbf{X}^*[j,:]\|^2 \le \|\mathbf{X}[j,:]\|^2$ and so, by assumption, $\|\mathbf{X}^*[j,:]\|^2 = \|\boldsymbol{\Sigma}^*\mathbf{U}^*[j,:]\|^2 \le 1$. Since all diagonal entries of $\boldsymbol{\Sigma}^*$ are at least $\sqrt{\gamma}$, this gives, $\ell_j(\mathbf{X}^*) = \|\mathbf{U}^*[j,:]\|^2 \le 1/\gamma$, completing the claim. □

Setting $\boldsymbol{\pi}$. It is well known that if we sample rows of $\mathbf{X}^*$ with probabilities $\boldsymbol{\pi}$ proportional to the leverage scores, we will obtain a $(1\pm\epsilon)$ relative error approximation for linear regression [40]. The result of Sarlos [40] applies to sampling s rows *with replacement*, each equal to j with probability $\boldsymbol{\pi}_j/\|\boldsymbol{\pi}\|$. It is not hard to observe that it extends to the variant where each row is included in the sample independently with similar probability. Therefore, we have:

Lemma H.4 (Follows from [40])**.** *For $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, let $S \subseteq [n]$ include each $j \in [n]$ independently with probability $\boldsymbol{\pi}_j$ satisfying $\boldsymbol{\pi}_j \ge \min\left\{1, \ell_j(\mathbf{X})\cdot c\cdot[\log(\mathrm{rank}(\mathbf{X})) + \frac{1}{\delta\epsilon}]\right\}$ for some large enough constant c. Let $\mathbf{W} \in \mathbb{R}^{|S|\times n}$ be a sampling matrix that includes row*

$\mathbf{e}_j/\sqrt{\boldsymbol{\pi}_j}$ *if* $j \in S$, *where* $\mathbf{e}_j \in \mathbb{R}^n$ *is the* j^{th} *standard basis vector. Let* $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\mathbf{W}\mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{y}\|^2$. *Then,* $\mathbb{E}[|S|] = \sum_{j=1}^n \boldsymbol{\pi}_j$ *and with probability* $\geq 1-\delta$*:*

$$\left\|\mathbf{X}\widetilde{\boldsymbol{\beta}} - \mathbf{y}\right\| \leq (1+\epsilon)\cdot \min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|.$$

If the $\boldsymbol{\pi}_j$*'s are within constants of the required bound,* $\mathbb{E}[|S|] = O\left(d\log d + \frac{d}{\epsilon\delta}\right)$.

Note that the bound on $\mathbb{E}[|S|]$ follows from the well known fact that the sum of leverage scores, is equal to the rank, i.e., $\sum_{j=1}^n \ell_j(\mathbf{X}) = \text{rank}(\mathbf{X}) \leq d$ [49].

The sampling probabilities are set to $\boldsymbol{\pi}_j = \min\{1, \ell_j(\mathbf{X}^*)\cdot c_0 \cdot [\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon]\}$ for some constant $c_0 \geq 2c$, where c is the constant in Lemma H.4. Thus, by the lemma, we will have, with probability $\geq 29/30$, $\left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0 - \mathbf{y}^0\right\| \leq (1+\epsilon)\left\|\mathbf{X}^*\boldsymbol{\beta}^0 - \mathbf{y}^0\right\|$.

It remains to show that we will have a similar guarantee for the control group. The rows in S^1 are included independently with probability $\boldsymbol{\pi}_j\cdot(1-\boldsymbol{\pi}_j)$. If we can prove that $\boldsymbol{\pi}_j\cdot(1-\boldsymbol{\pi}_j) \geq \frac{\boldsymbol{\pi}_j}{2}$, then Lemma H.4 will still apply, since we have set $c_0 = 2c$. To do so, it suffices to argue that $\boldsymbol{\pi}_j \leq 1/2$ by setting the parameters appropriately.

Claim H.5. *If* $\gamma = 4c_0\max\{\log(\text{rank}(\mathbf{X}^*)), 30/\epsilon\}$ *and* $\boldsymbol{\pi}_j = \min\{1, \ell_j(\mathbf{X}^*)\cdot c_0\cdot[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon]\}$*, we have* $\boldsymbol{\pi}_j \leq 1/2$ *for every* $j \in [n]$.

Proof.

$$\begin{aligned}\boldsymbol{\pi}_j \leq \ell_j(\mathbf{X}^*)\cdot c_0\cdot[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon] &\leq 1/\gamma\cdot c_0\cdot[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon] \text{ (Claim H.3)}\\ &\leq \frac{c_0[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon]}{4c_0\max\{\log(\text{rank}(\mathbf{X}^*)), 30/\epsilon\}} \leq \frac{1}{2}.\end{aligned}$$

□

Combining Lemma H.4 and Claim H.5, we get:

Lemma H.6. *Suppose* $\gamma = 4c_0\max\{\log(\text{rank}(\mathbf{X}^*)), 30/\epsilon\}$ *and* $\boldsymbol{\pi}_j = \min\left\{1, \ell_j(\mathbf{X}^*)\cdot c_0\cdot[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon]\right\}$*, for some sufficiently large constant* c_0*. Then, Algorithm* SAMPLING-ITE *satisfies, for* $i = 0, 1$*, with probability at least* $14/15$*:*

$$\left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^i - \mathbf{y}^i\right\| \leq (1+\epsilon)\cdot\left\|\mathbf{X}^*\boldsymbol{\beta}^i - \mathbf{y}^i\right\|.$$

Further, $\mathbb{E}[|S^0 \cup S^1|] \leq 2\sum_{j=1}^n \boldsymbol{\pi}_j = O(d\log d + d/\epsilon)$.

Proof. From Lemma H.4 and Claim H.5, we have:

$$\left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^i - \mathbf{y}^i\right\| \leq (1+\epsilon)\cdot\left\|\mathbf{X}^*\boldsymbol{\beta}^i - \mathbf{y}^i\right\| \text{ for every } i = 0, 1.$$

Using union bound, the total failure probability is $\leq \frac{1}{30} + \frac{1}{30} \leq \frac{1}{15}$.

From Algorithm 1, let S^0, S^1 denote the set of people assigned to treatment and control respectively. From Lemma H.4, we have:

$$\begin{aligned}\mathbf{E}[|S^0 \cup S^1|] \leq 2\sum_{j=1}^n \boldsymbol{\pi}_j &\leq 2\sum_{j\in[n]} \ell_j(\mathbf{X}^*)\cdot c_0\cdot[\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon]\\ &\leq 2c_0 d\cdot[\log d + 30/\epsilon] = O(d\log d + d/\epsilon) \text{ (using Claim F.4 and rank}(\mathbf{X}^*) \leq d).\end{aligned}$$

□

Corollary H.7. *Suppose* $\gamma = 4c_0 \max\{\log(\text{rank}(\mathbf{X}^*)), 30/\epsilon\}$ *and* $\boldsymbol{\pi}_j = \min\left\{1, \ell_j(\mathbf{X}^*) \cdot c_0 \cdot [\log(\text{rank}(\mathbf{X}^*))+30/\epsilon]\right\}$*, for some sufficiently large constant* c_0*. Then, Algorithm* SAMPLING-ITE *satisfies, for* $i = 0, 1$*, with probability at least* $14/15$*:*

$$\left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^i - \mathbf{y}^i\right\| \leq (1+\epsilon) \cdot \left(\sqrt{\gamma}\left\|\boldsymbol{\beta}^i\right\| + \left\|\boldsymbol{\zeta}^i\right\|\right) \text{ for } i = 0, 1.$$

Proof.

$$\begin{aligned}
\left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^i - \mathbf{y}^i\right\| &\leq (1+\epsilon) \cdot \left\|\mathbf{X}^*\boldsymbol{\beta}^i - \mathbf{y}^i\right\| \quad \text{(from Lemma H.6)} \\
&\leq (1+\epsilon) \cdot \left(\left\|\mathbf{X}^*\boldsymbol{\beta}^i - \mathbf{X}\boldsymbol{\beta}^i\right\| + \left\|\mathbf{X}\boldsymbol{\beta}^i - \mathbf{y}^i\right\|\right) \quad \text{(using triangle inequality)} \\
&\leq (1+\epsilon) \cdot \left(\sqrt{\gamma}\left\|\boldsymbol{\beta}^i\right\| + \left\|\boldsymbol{\zeta}^i\right\|\right) \quad \text{(from Claim H.2)}
\end{aligned}$$

□

RMSE Guarantees. The root mean squared error (Defn. 2.1) for the ITE estimates is given by:

$$\text{RMSE} = \frac{1}{\sqrt{n}}\left\|(\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^1 - \mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0) - (\mathbf{y}^1 - \mathbf{y}^0)\right\|.$$

By setting $\epsilon = 120c_0 d \log d/s$ in Corollary H.7, we get the following theorem for our Algorithm 1:

Theorem H.8 (Theorem 3.1 restated)**.** *Suppose* $s \geq 120c_0 d \log d$*. There is a randomized algorithm that selects a subset* $S \subseteq [n]$ *of the population with* $\mathbf{E}[|S|] \leq s$*, and, with probability at least* $9/10$*, returns ITE estimates* $\widehat{\text{ITE}}(j)$ *for all* $j \in [n]$ *with error:*

$$\text{RMSE} = O\Big(\sqrt{\frac{1}{n}\max\left\{\frac{s}{d}, \log d\right\}} \cdot (\left\|\boldsymbol{\beta}^1\right\| + \left\|\boldsymbol{\beta}^0\right\|) + \sigma\Big).$$

Proof. Let $\epsilon = \frac{120c_0 d \log d}{s}$ and $\gamma = 4c_0 \max\{\log d, 30/\epsilon\}$. Using triangle inequality, we have:

$$\begin{aligned}
\left\|(\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^1 - \mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0) - (\mathbf{y}^1 - \mathbf{y}^0)\right\|_2 &\leq \left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^1 - \mathbf{y}^1\right\|_2 + \left\|\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0 - \mathbf{y}^0\right\|_2 \\
&\leq (1+\epsilon) \cdot \left(\sqrt{\gamma}\left\|\boldsymbol{\beta}^1\right\| + \sqrt{\gamma}\left\|\boldsymbol{\beta}^0\right\| + \left\|\boldsymbol{\zeta}^0\right\| + \left\|\boldsymbol{\zeta}^1\right\|\right) \quad \text{(from Corollary H.7)}
\end{aligned}$$

From the definition of RMSE, we get:

$$\begin{aligned}
\text{RMSE} &= \left(\frac{1}{n}\left\|(\mathbf{X}^*\widetilde{\boldsymbol{\beta}}^1 - \mathbf{X}^*\widetilde{\boldsymbol{\beta}}^0) - (\mathbf{y}^1 - \mathbf{y}^0)\right\|^2\right)^{1/2} \\
&\leq \frac{1}{\sqrt{n}}\left[2\sqrt{\gamma} \cdot (\left\|\boldsymbol{\beta}^1\right\| + \left\|\boldsymbol{\beta}^0\right\|) + 2 \cdot (\left\|\boldsymbol{\zeta}^0\right\| + \left\|\boldsymbol{\zeta}^1\right\|)\right] \\
&\leq \frac{1}{\sqrt{n}}\left[2\sqrt{\gamma} \cdot (\left\|\boldsymbol{\beta}^1\right\| + \left\|\boldsymbol{\beta}^0\right\|) + 8\sigma\sqrt{n}\right] \quad \text{(from Lemma F.5)} \\
&\leq 2\sqrt{\frac{4c_0}{n}\max\left\{\log d, \frac{s}{c_0 d}\right\}} \cdot (\left\|\boldsymbol{\beta}^1\right\| + \left\|\boldsymbol{\beta}^0\right\|) + 8\sigma
\end{aligned}$$

Using union bound, the probability of failure is upper bounded by $\frac{1}{15} + \frac{2}{n} \leq \frac{1}{10}$, for large n.

From Algorithm 1, let S^0, S^1 denote the set of people assigned to treatment and control respectively. From Lemma H.6, we have:

$$\begin{aligned}\mathbf{E}[|S^0 \cup S^1|] \leq 2\sum_{j=1}^{n} \boldsymbol{\pi}_j &\leq 2 \sum_{j\in[n]} \ell_j(\mathbf{X}^*) \cdot c_0 \cdot [\log(\text{rank}(\mathbf{X}^*)) + 30/\epsilon] \\ &\leq 2c_0 d \cdot [\log d + 30/\epsilon] \text{ (using Claim F.4 and } \text{rank}(\mathbf{X}^*) \leq d) \\ &\leq 4c_0 d \max\left\{\log d, \frac{s}{4c_0 d \log d}\right\} = \max\{4c_0 d \log d, s\} \leq s.\end{aligned}$$

Hence, the theorem. □

The corollary below follows immediately from Theorem 3.1.

Corollary H.9 (Corollary 3.2 restated)**.** *The root mean squared error obtained by Algorithm 1 is minimized when $s = \Theta(d \log d)$ and is given by:*

$$\text{RMSE} = O\Big(\sqrt{\frac{\log d}{n}} \cdot (\|\boldsymbol{\beta}^1\| + \|\boldsymbol{\beta}^0\|) + \sigma\Big).$$

Our upper bound on RMSE increases with s, if s grows strictly faster than $d \log d$ asymptotically, i.e., $s = \omega(d \log d)$. Therefore, to obtain low error, we set $s = c \cdot d \log d$ for some constant c, even if the sample constraint allows for larger values. We believe this is an artifact of our analysis; in Section 5, we observe empirically that the error decreases with s.

Remark. We observe that the RMSE bound in Corollary 3.2 is nearly optimal, even for algorithms that *experiment on the full population*. The $O(\sigma)$ term cannot be improved by more than constants, as a consequence of our noise model (Linearity Assumption). Even if we knew the true $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^0$, our RMSE would be $O(\sigma)$.

The term $(\|\boldsymbol{\beta}^0\| + \|\boldsymbol{\beta}^1\|)/\sqrt{n}$ is also necessary. Suppose the matrix $\mathbf{X}$ is such that all rows, except row j, are zero vectors. Row j is a standard basis vector, i.e., its i^{th} entry is 1 for some i. Suppose also that $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^0$ are both independently set to that same standard basis vector with probability 1/2, and set to zero otherwise. Then, with probability 1/2, $\text{ITE}(j) = 0$ and with probability 1/2, $\text{ITE}(j) = \pm 1$. No algorithm which observes just one of $\mathbf{y}_j^1$ or $\mathbf{y}_j^0$ can obtain expected error o(1) in estimating $\text{ITE}(j)$. That is, no algorithm can obtain RMSE $o(1/\sqrt{n}) = o\left((\|\boldsymbol{\beta}^0\| + \|\boldsymbol{\beta}^1\|)/\sqrt{n}\right)$.

I. Average Treatment Effect Estimation

Harshaw et al. [19] present an experimental design based on the Gram-Schmidt-Walk algorithm for discrepancy minimization [8]. Their Gram-Schmidt-Walk design produces a random partition of the population with a good balance in every dimension, i.e., control and treatment groups have similar covariates. For the Horvitz-Thompson estimator, they give a tradeoff between covariate balancing and robustness (estimation error). Formally, they obtain:

Algorithm 2 RECURSIVE-COVARIATE-BALANCING

Input: Covariate matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, number of experiments to be run s.
Output: Estimate for ATE.

Set $t = 1, \mathbf{Z}_t := \mathbf{X}$, $n_t = n$.
while True do
  $\mathbf{Z}_t^+, \mathbf{Z}_t^- \leftarrow$ GRAM-SCHMIDT-WALK$(\mathbf{Z}_t, \delta')$ where $\delta' = \log(16 \log(n/s))$.
  if $n_t \leq s$ **then**
    break
  else if size$(\mathbf{Z}_t^+) \geq$ size$(\mathbf{Z}_t^-)$ **then**
    Set $\mathbf{Z}_{t+1} \leftarrow \mathbf{Z}_t^-$ and $n_{t+1} \leftarrow$ size$(\mathbf{Z}_t^-)$.
  else
    Set $\mathbf{Z}_{t+1} \leftarrow \mathbf{Z}_t^+$ and $n_{t+1} \leftarrow$ size$(\mathbf{Z}_t^+)$.
  end if
  $t \leftarrow t + 1$
end while
Use $\mathbf{Z}_t^+, \mathbf{Z}_t^-$ to construct the ATE estimator as:

$$\widehat{\tau}_s = 2^t/n \cdot \left(\sum_{j \in \mathbf{Z}_t^+} \mathbf{y}_j^1 - \sum_{j \in \mathbf{Z}_t^-} \mathbf{y}_j^0 \right).$$

return $\widehat{\tau}_s$.

Lemma I.1 (Proposition 3 in [19])**.** *For all $\Delta > 0$, with probability at least $1 - 2\exp\left(-\frac{\Delta^2 n}{8L}\right)$, the Gram-Schmidt-Walk design satisfies: $|\widehat{\tau} - \tau| \leq \Delta$, where*

$$L = \frac{2}{n} \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left(\left\| \frac{\mathbf{y}^1 + \mathbf{y}^0}{2} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \|\boldsymbol{\beta}\|^2 \right).$$

Theoretical Guarantees. Our analysis approach, inspired by the coreset construction for discrepancy minimization [25], is based on the observation that if we can obtain good estimates for the contributions $\sum_{i \in [n]} \mathbf{y}_i^1$ and $\sum_{i \in [n]} \mathbf{y}_i^0$, we obtain a good estimate for ATE (τ). Using the next lemma, we argue that after a call to GSW algorithm that partitions $[n]$ into the sets $\mathbf{S}^+$ and $\mathbf{S}^-$, we can obtain additive approximations of $\sum_{i \in [n]} \mathbf{y}_i^1$ and $\sum_{i \in [n]} \mathbf{y}_i^0$. Our approximations are the contributions of treatment and control values in $\mathbf{S}^+$ and $\mathbf{S}^-$ scaled appropriately, i.e., $\sum_{i \in \mathbf{S}^+} 2 \cdot \mathbf{y}_i^1$ and $\sum_{i \in \mathbf{S}^-} 2 \cdot \mathbf{y}_i^0$.

Lemma I.2. *Suppose the Gram-Schmidt-Walk design [19] partitions the population $[n]$ into two disjoint groups $\mathbf{S}^+$ and $\mathbf{S}^-$. Under the linearity assumption, with probability $1 - 1/3\log(n/s)$, for both the control and treatment groups, the following holds:*

$$\left| \sum_{j \in \mathbf{S}^+} 2\mathbf{y}_j^i - \sum_{j \in [n]} \mathbf{y}_j^i \right| \leq 4\sqrt{\log(16 \log(n/s))} \cdot \left(2\sigma\sqrt{n} + \left\|\boldsymbol{\beta}^i\right\|\right) \quad \text{for } i = 0, 1.$$

Proof. The Gram-Schmidt-Walk design uses the covariate matrix $\mathbf{X}$ but not the treatment and control values $\mathbf{y}^1, \mathbf{y}^0$, for constructing the partition of the population $\mathbf{S}^+, \mathbf{S}^-$. For the sake of analysis, consider the setting where $\mathbf{y}_i^1 = \mathbf{y}_i^0$ for all $i \in [n]$. Therefore, the average treatment effect, $\tau = 0$, and the estimator $\widehat{\tau}$ satisfies:

$$\widehat{\tau} - \tau = \frac{2}{n} \left(\sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \sum_{i \in \mathbf{S}^-} \mathbf{y}_i^0 \right) = \frac{2}{n} \left(\sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \sum_{i \in \mathbf{S}^-} \mathbf{y}_i^1 \right)$$

From Lemma I.1, we have:

$$\begin{aligned} L &= \frac{2}{n} \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left(\left\| \frac{\mathbf{y}^1 + \mathbf{y}^0}{2} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \|\boldsymbol{\beta}\|^2 \right) \\ &= \frac{2}{n} \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left(\left\| \mathbf{y}^1 - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \|\boldsymbol{\beta}\|^2 \right) \\ &\leq \frac{2}{n} \left(\left\| \mathbf{y}^1 - \mathbf{X}\boldsymbol{\beta}^1 \right\|^2 + \left\| \boldsymbol{\beta}^1 \right\|^2 \right) = \frac{2}{n} \left(\left\| \boldsymbol{\zeta}^1 \right\|^2 + \left\| \boldsymbol{\beta}^1 \right\|^2 \right). \end{aligned}$$

From Lemma I.1, with probability at least $1 - 2/\log(16\log(n/s))$, we have:

$$\begin{aligned} |\widehat{\tau} - \tau| &= \left| \frac{2}{n} \left(\sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \sum_{i \in \mathbf{S}^-} \mathbf{y}_i^1 \right) \right| \\ &\leq \sqrt{\frac{16\log(16\log(n/s))}{n^2} \left(\left\| \boldsymbol{\zeta}^1 \right\|^2 + \left\| \boldsymbol{\beta}^1 \right\|^2 \right)} \\ &\leq \frac{4\sqrt{\log(16\log(n/s))}}{n} \cdot \left(\left\| \boldsymbol{\zeta}^1 \right\| + \left\| \boldsymbol{\beta}^1 \right\| \right) \end{aligned}$$

For simplicity, let :

$$\begin{aligned} \Delta^1 &= 4\sqrt{\log(16\log(n/s))} \cdot \left(\left\| \boldsymbol{\zeta}^1 \right\| + \left\| \boldsymbol{\beta}^1 \right\| \right) \\ &\leq 4\sqrt{\log(16\log(n/s))} \cdot \left(2\sigma\sqrt{n} + \left\| \boldsymbol{\beta}^1 \right\| \right), \end{aligned}$$

where the last inequality follows from Fact F.5, with probability at least $1 - 1/n$.

$$\begin{aligned} \sum_{i \in \mathbf{S}^-} \mathbf{y}_i^1 &\geq \sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \Delta^1 \\ \sum_{i \in \mathbf{S}^-} \mathbf{y}_i^1 + \sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 &\geq \sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 + \sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \Delta^1 \\ \sum_{i \in [n]} \mathbf{y}_i^1 &\geq \sum_{i \in \mathbf{S}^+} 2\mathbf{y}_i^1 - \Delta^1 \\ \Rightarrow 2 \sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 - \sum_{i \in [n]} \mathbf{y}_i^1 &\leq \Delta^1. \end{aligned}$$

Similarly, we can argue that $\sum_{i \in [n]} \mathbf{y}_i^1 - 2\sum_{i \in \mathbf{S}^+} \mathbf{y}_i^1 \leq \Delta^1$. Using union bound, the inequality holds with probability at least $1 - \frac{2}{16\log(n/s)} - \frac{1}{n} \geq 1 - \frac{1}{6\log(n/s)}$. Following the exact proof, we can obtain a similar bound for $\mathbf{y}^0$ using the set $\mathbf{S}^-$. Hence, the lemma. □

Building upon the previous lemma, we argue in Theorem I.3 that the additive approximation errors obtained from repeated use of GSW in our algorithm RECURSIVE-COVARIATE-BALANCING result in a low estimation error.

Theorem I.3 (Theorem 4.1 restated)**.** *The estimator $\widehat{\tau}_s$ in Algorithm* Recursive-Covariate-Balancing *obtains the following guarantee, with probability at least 2/3:*

$$|\widehat{\tau}_s - \tau| = O\left(\sqrt{\log\log(n/s)} \cdot \left(\frac{\sigma}{\sqrt{s}} + \frac{\left\|\boldsymbol{\beta}^1\right\| + \left\|\boldsymbol{\beta}^0\right\|}{s}\right)\right).$$

Proof. Suppose the Algorithm Recursive-Covariate-Balancing gets terminated after $k \leq \lceil \log(n/s) \rceil$ recursive calls to Gram-Schmidt-Walk. Therefore, in the estimator $\widehat{\tau}_s$, we scale it using 2^k. For simplicity of notation, we use $\mathbf{S}^+, \mathbf{S}^-$ to denote the sets $\mathbf{Z}_k^+$ and $\mathbf{Z}_k^-$ respectively. Using Lemma I.2 , we show that the scaled contribution of treatment values, i.e., $\sum_{j \in \mathbf{S}^+} 2^k \cdot \mathbf{y}_j^1$ is close to the contribution on the entire population, i.e., $\sum_{j \in [n]} \mathbf{y}_j^1$. As this holds for both the control and treatment groups, our final estimate $\widehat{\tau}_s$ has low error. We have:

$$\widehat{\tau}_s - \tau = \frac{2^k}{n}\left(\sum_{j \in \mathbf{S}^+} \mathbf{y}_j^1 - \sum_{j \in \mathbf{S}^-} \mathbf{y}_j^0\right) - \frac{1}{n}\left(\sum_{i \in [n]} \mathbf{y}_i^1 - \sum_{i \in [n]} \mathbf{y}_i^0\right)$$

$$n\left|\widehat{\tau}_s - \tau\right| \leq \left|\sum_{j \in \mathbf{S}^+} 2^k \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right| + \left|\sum_{j \in \mathbf{S}^-} 2^k \cdot \mathbf{y}_j^0 - \sum_{i \in [n]} \mathbf{y}_i^0\right|$$

Consider the first term to which we add and subtract $\sum_{j \in \mathbf{S}^+ \cup \mathbf{S}^-} \mathbf{y}_j^1$. This gives us:

$$\begin{aligned}\left|\sum_{j \in \mathbf{S}^+} 2^k \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right| &\leq \left|2^{k-1}\left(\sum_{j \in \mathbf{S}^+} 2 \cdot \mathbf{y}_j^1 - \sum_{j \in \mathbf{Z}_k} \mathbf{y}_j^1\right)\right| + \left|\sum_{j \in \mathbf{Z}_k} 2^{k-1} \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right| \\ &\leq \left|2^{k-1} \cdot 4\sqrt{\log(16\log(n/s))}\left(2\sigma\sqrt{|\mathbf{Z}_k|} + \left\|\boldsymbol{\beta}^1\right\|\right)\right| + \left|\sum_{j \in \mathbf{Z}_k} 2^{k-1} \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right|,\end{aligned}$$

where the last step follows from Lemma I.2. Repeating this k times gives us:

$$\begin{aligned}\left|\sum_{j \in \mathbf{S}^+} 2^k \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right| &\leq 4\sqrt{\log(16\log(n/s))} \cdot 2^k \cdot \\ &\quad \left[\left(\frac{\sqrt{|\mathbf{Z}_k|}}{1}\frac{\sqrt{|\mathbf{Z}_{k-1}|}}{2} + \cdots + \frac{\sqrt{|\mathbf{Z}_1|}}{2^k}\right)\sigma + \left(\frac{1}{2} + \cdots + \frac{1}{2^k}\right)\left\|\boldsymbol{\beta}^1\right\|\right] \\ &\leq 4\sqrt{\log(16\log(n/s))} \cdot \frac{n}{s} \cdot \left[\left(\frac{\sqrt{s}}{1} + \frac{\sqrt{2s}}{2} + \cdots + \frac{s\sqrt{n}}{n}\right)\sigma + \left\|\boldsymbol{\beta}^1\right\|\right] \\ \left|\frac{1}{n}\left(\sum_{j \in \mathbf{S}^+} 2^k \cdot \mathbf{y}_j^1 - \sum_{i \in [n]} \mathbf{y}_i^1\right)\right| &\leq 4\sqrt{\log(16\log(n/s))} \cdot \left[\left(\frac{1}{\sqrt{s}} + \frac{1}{\sqrt{2s}} + \frac{1}{\sqrt{4s}} \cdots + \frac{1}{\sqrt{n}}\right)\sigma + \frac{\left\|\boldsymbol{\beta}^1\right\|}{s}\right] \\ &\leq 4\sqrt{\log(16\log(n/s))} \cdot \left(\frac{4\sigma}{\sqrt{s}} + \frac{\left\|\boldsymbol{\beta}^1\right\|}{s}\right).\end{aligned}$$

Similarly, we can show that:

$$\left|\frac{1}{n}\left(\sum_{j\in\mathbf{S}^-} 2^k\cdot\mathbf{y}_j^0-\sum_{i\in[n]}\mathbf{y}_i^0\right)\right|\leq 4\sqrt{\log(16\log(n/s))}\cdot\left(\frac{4\sigma}{\sqrt{s}}+\frac{\left\|\boldsymbol{\beta}^0\right\|}{s}\right).$$

Using union bound, the total failure probability is upper bounded by $\frac{1}{3\log(n/s)}\cdot\log(n/s)\leq\frac{1}{3}$. Hence, the theorem.

□

J. Experimental Evaluation: Additional Details

Data Generation We evaluate our approaches on five datasets:

(i) *IHDP*. This contains data regarding the cognitive development of children, and consists of 747 samples with 25 covariates describing properties of the children and their mothers, and whose outcome values are simulated [20, 13].

(ii) *Twins*. This contains data regarding the mortality rate in twin births in USA between 1989-1991 [5]. Following the work of [31], we select twins belonging to same sex, with weight less than 2kg, resulting in about 11984 twins (pairs), each with 48 covariates. The purpose of the experiment is to evaluate the effect of weight (treatment) on mortality (outcome). We use the binary value corresponding to the mortality value as the treatment outcome. As we have a pair of outcome values for every twin pair, we use them as potential outcomes.

(iii) *LaLonde*. This contains data regarding the effectiveness of a job training program on the real earnings of an individual after completion of the program [28], which is also the outcome value. The corresponding covariate matrix contains 445 rows and 10 covariates per row.

(iv) *Boston*. This is constructed based on the housing prices in the Boston area [18]. The treatment variable was air pollution and the outcome value recorded for each sample is the median house price. The corresponding covariate matrix contains 506 rows and 12 covariates per row.

(v) *Synthetic*. In Figure 3, we observe a high disparity in the leverage score values (and the spectrum) of the covariate matrix in the real datasets. In order to generate fully synthetic dataset that shows a similar pattern, we used an approach due to [32]. In particular, we used their third dataset configuration, i.e., $\mathbf{X}\in\mathbb{R}^d$ is generated from multi-variate t-distribution with 1 degree of freedom and the covariance matrix is $\Sigma\in\mathbb{R}^{d\times d}$ where $\Sigma_{ij}=2\cdot(0.5)^{|i-j|}$. For both the potential outcomes, we use a random linear function and add Gaussian noise. E.g., for the control outcome $\mathbf{y}^0=\mathbf{X}\boldsymbol{\beta}^0+\boldsymbol{\zeta}^0$, we generate $\boldsymbol{\beta}^0\in\mathbb{R}^d$ by drawing each co-ordinate uniformly from $[0,1]$ and normalize it to make it a unit vector. The Gaussian noise is generated from $N(0,c\cdot I_{n\times n})$, where we vary c in the range of $[1/n^{0.5},1/d^{0.5}]$ to ensure that the contribution of the noise term to the ℓ_2-norm is very small.

Setup We used a personal Apple Macbook Pro laptop with 16GB RAM and Intel i5 processor for conducting all our experiments. It took less than an hour to complete each experiment on each dataset. We used publicly available code for the implementation of GSW algorithm [19].

Baselines (i) **ATE.** We compare the performance of our Algorithm RECURSIVE-COVARIATE-BALANCING (referred to as *'Recursive-GSW'*) to three baselines: *(i) Uniform.* We sample s rows uniformly at random and assign them to treatment and control groups with equal probability. By scaling the total sum of treatment values from the sampled set by the inverse sampling probability, we estimate the contribution of treatment values in ATE and follow a similar procedure for the control group. *(ii) GSW-pop.* We use the GSW algorithm to partition the full population and return the estimate obtained using the Horvitz-Thompson estimator for ATE. *(iii) Complete Randomization.* We partition the population into treatment and control using complete randomization, i.e., with equal probability, and return the estimate obtained using the Horvitz-Thompson estimator for ATE. The last two baselines are overall n individuals rather than a subset of size s.

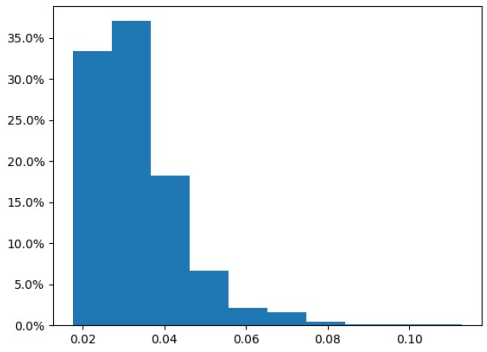
(ii) **ITE.** We compare the performance of our Algorithm SAMPLING-ITE (referred to as *'Leverage'*) with respect to three baselines: *(i) Uniform.* We run Algorithm 1 on $\mathbf{X}$ and uniform sampling distribution given by $\boldsymbol{\pi}_j = s/n \ \forall j$.*(ii) Leverage-nothresh.* We run Algorithm 1 on $\mathbf{X}$, instead of $\mathbf{X}^*$ with the probability distribution $\boldsymbol{\pi}_j \propto \ell_j(\mathbf{X}) \forall j$.*(iii) Linregression.* This captures the best linear fit regression error, i.e., assuming we have access to both $\mathbf{y}^1, \mathbf{y}^0$, we regress these vectors on $\mathbf{X}$ to obtain $\boldsymbol{\beta}^1, \boldsymbol{\beta}^0$, and use the resultant ITE estimates $\mathbf{X}\boldsymbol{\beta}^1 - \mathbf{X}\boldsymbol{\beta}^0$.

Results For every dataset, we run each experiment for 1000 trials and plot the mean using a colored line. Also, we shade the region between 30 and 70 percentile around the mean to signify the *confidence interval* as shown in Figures 4, 5 representing ATE and ITE results respectively.

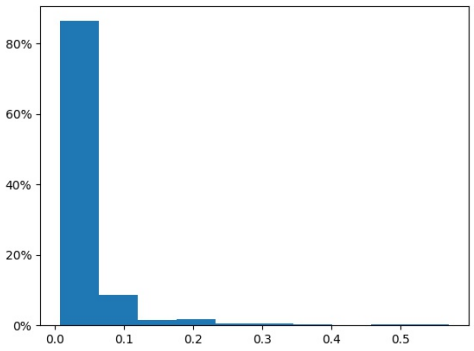
(i) **ATE**. For all datasets, we observe that the deviation error obtained by our algorithm labeled as *Recursive-GSW* in Figure 4, is significantly smaller than that of *Uniform* baseline. Surprisingly, for the IHDP dataset, our approach is significantly better than *Complete-randomization*, for all sample sizes, including using just 10% of data. For almost all the remaining datasets using a sample of size 30%, we achieve the same bias (up to the confidence interval) as that of *Complete-randomization.* For the Boston dataset, our approach is better than *Complete-randomization*, for all sample sizes. Complete randomization is one of the most commonly used methods for experimental design and our results indicate a substantial reduction in experimental costs. For IHDP dataset, a sample size of about 10% of the population is sufficient to achieve a similar bias as that of *GSW-pop.* For the remaining datasets, we observe that for sample sizes of about 30% of the population, the deviation error obtained by our algorithm is within the shaded confidence interval of the bias obtained by *GSW-pop.* Therefore, for a specified error tolerance level for ATE, we can reduce the associated experimental costs using just a small subset of the dataset using our algorithm.

(ii) **ITE**. For all sample sizes, we observe that the RMSE obtained by our algorithm labeled as *Leverage* in Figure 5, is significantly smaller than that of all the other baselines, including *Uniform* and *Leverage-nothresh.* E.g., we observe that when the sample size is 20% of the population in IHDP dataset, the error obtained by *Leverage* is at least 50%

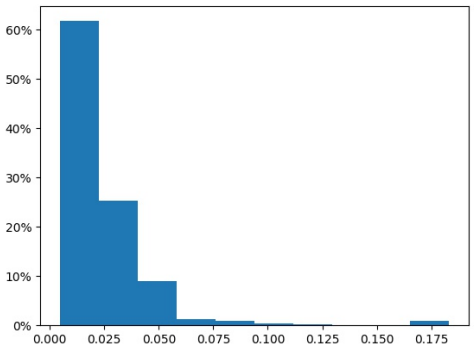
times smaller than that of *Uniform* and *Leverage-nothresh*. For the Synthetic and Twins datasets, the error obtained by *Leverage* is extremely close to that of the error due to the best linear fit, *Lin-regression* (see the zoomed in part of the figure). Similar to ATE results, our algorithms result in a reduction of experimental costs for ITE estimation using only a fraction of the dataset.



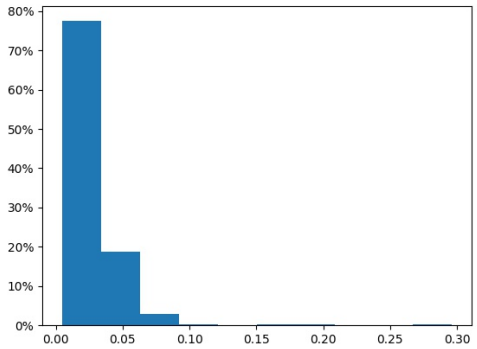
(a) IHDP



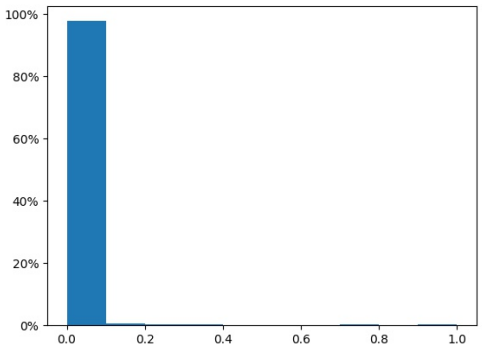
(b) Twins



(c) Lalonde

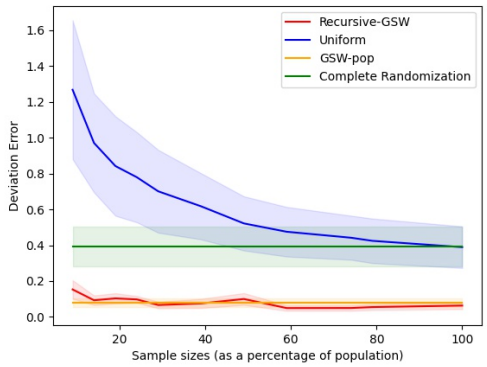


(d) Boston

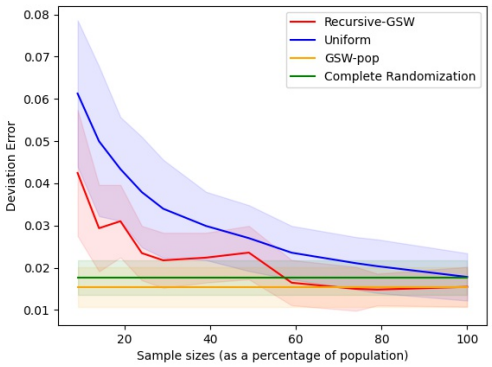


(e) Synthetic

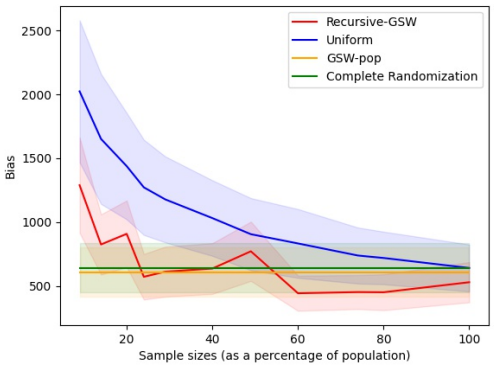
Figure 3: We plot the histogram of leverage scores of the covariate matrices for each of the datasets. On y-axis, we measure the percentage of the dataset corresponding to a particular leverage score (on the x-axis).



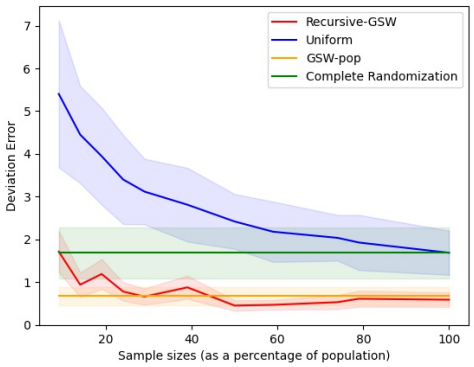
(a) IHDP



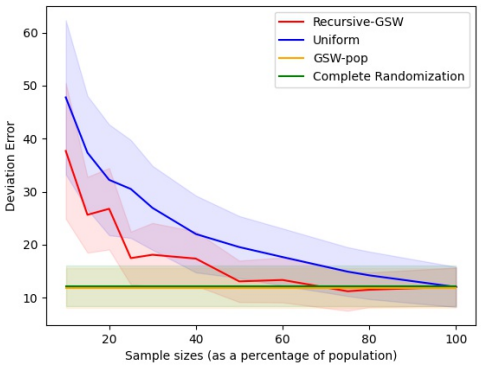
(b) Twins



(c) Lalonde

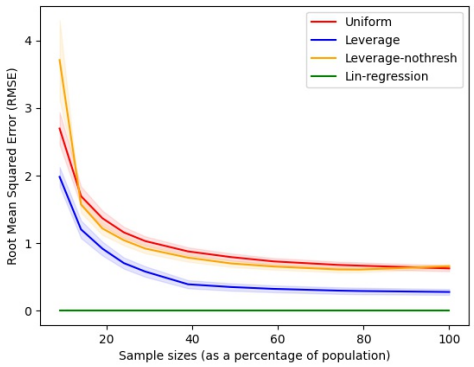


(d) Boston

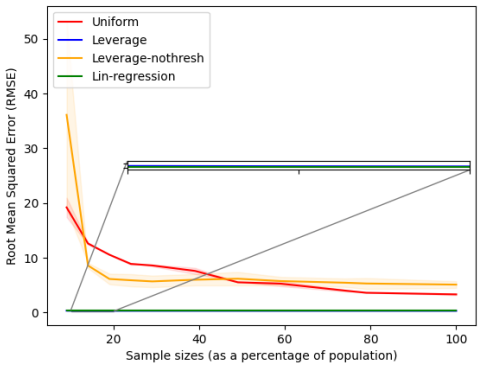


(e) Synthetic

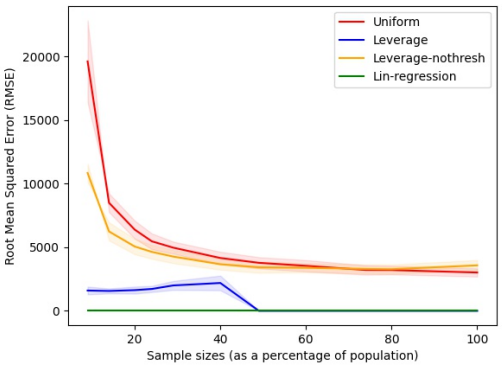
Figure 4: We compare the performance of various methods for estimating ATE, measured using deviation error on y-axis, against different sample sizes (as proportion of dataset size) on x-axis.



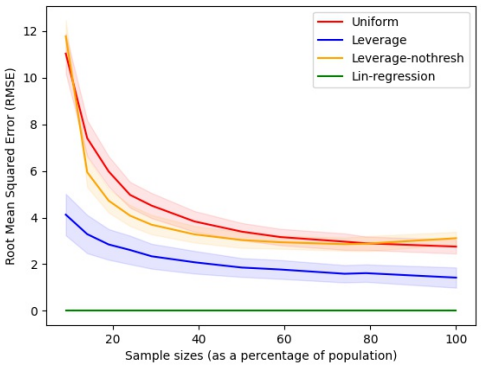
(a) IHDP



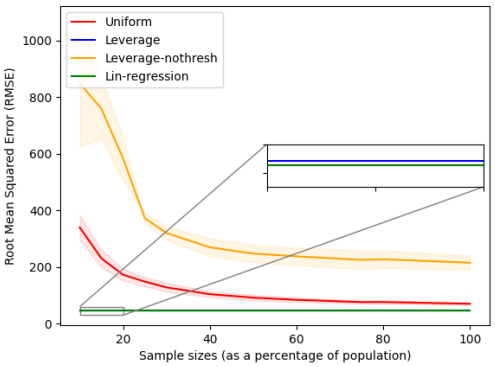
(b) Twins



(c) Lalonde



(d) Boston



(e) Synthetic

Figure 5: We compare the performance of various methods for estimating ITE, measured using RMSE on y-axis, against different sample sizes (as proportion of dataset size) on x-axis.