Strategic Instrumental Variable Regression: Recovering Causal Relationships From Strategic Responses

Anonymous Authors[1]

Abstract

In social domains, Machine Learning algorithms often prompt individuals to strategically modify their observable attributes to receive more favorable predictions. As a result, the distribution the predictive model is trained on may differ from the one it operates on in deployment. While such distribution shifts, in general, hinder accurate predictions, our work identifies a unique opportunity associated with shifts due to strategic responses: We show that we can use strategic responses effectively to recover causal relationships between the observable features and outcomes we wish to predict. More specifically, we study a game-theoretic model in which a principal deploys a sequence of models to predict an outcome of interest (e.g., college GPA) for a sequence of T strategic agents (e.g., college applicants). In response, strategic agents invest efforts and modify their features for better predictions. In such settings, unobserved confounding variables (e.g., family educational background) can influence both an agent's observable features (e.g., high school records) and outcomes (e.g., college GPA). Therefore, standard regression methods (such as OLS) generally produce biased estimators. In order to address this issue, our work establishes a novel connection between strategic responses to machine learning models and instrumental variable (IV) regression, by observing that the sequence of deployed models can be viewed as an *instrument* that affects agents' observable features but does not *directly* influence their outcomes. Therefore, two-stage least squares (2SLS) regression can recover the causal relationships between observable features and outcomes.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Machine learning (ML) predictions increasingly inform high-stakes decisions for people in areas such as college admissions (15, 23), credit scoring (17, 19), employment (20), and beyond. One of the major criticisms against the use of ML in socially consequential domains is the failure of these technologies to identify *causal* relationships among relevant attributes and the outcome of interest (9). The single-minded focus of ML on predictive accuracy has given rise to brittle predictive models that learn to rely on spurious correlations—and at times, and harmful stereotypes—to achieve seemingly accurate predictions on held-out test data (24, 8). The resulting models frequently underperform in deployment, and their predictions can negatively impact decision subjects through several distinct pathways. As an example, ML-based decision-making systems often prompt individuals to modify their observable attributes strategically to receive more favorable predictions—and subsequently, decisions (13). (These strategic responses are among the primary causes of distribution shifts leading to the unsatisfactory performance of ML in social domains.) Moreover, recent work has established the potential of these tools to amplify existing social disparities by incentivizing different effort investments across distinct groups of subjects (10, 6, 14).

These issues have led to renewed calls on the ML community to strengthen the bond between ML and causality (16, 21). Knowledge of causal relationships among predictive attributes and outcomes of interest promotes several desirable aims: First, ML practitioners can use this knowledge to debug their models and ensure robustness even if the underlying population shifts over time. Second, policymakers can utilize the causal understanding of a domain in their policy choices and examine a decision-making system's compliance with their goals and values (e.g., they can audit the system for unfairness against particular populations (11).) Finally, predictions rooted in causal associations block pathways of gaming and manipulation and, instead, encourage decision subjects to make meaningful interventions that improve their actual outcomes (as opposed to their assessments alone).

Our work responds to the above calls by offering a new approach to recover causal relationships between observable features and the outcome of interest under strategic responses—without substantially hampering predictive accuracy. We consider settings where a decision-maker deploys a sequence of models to predict the outcome for a sequence of strategic decision subjects. Often in such settings, there are unobserved confounding variables that influence subjects' attributes and outcomes simultaneously. Our key observation is that we can correct for the effect of such confounders by viewing the sequence of **assessment rules as valid *instruments*** which affect subjects' observable features but do not *directly* influence their outcomes. Our main contribution is a general framework that recovers the causal relationships between observed attributes and the outcome of interest by treating assessment rules as instruments.

1.1. Our setting

Consider a stylized setting in which a university decides whether to admit or reject applicants on a rolling basis (for example, (18)) based (in part) on how well they are predicted to perform if admitted to the university (See Figure 1). We model such interactions as a game between a *principal* (here, the university) and a population of *agents* (here, university applicants) who arrive sequentially over T rounds, indexed by $t = 1, 2, \cdots, T$. In each round t, the principal deploys an assessment rule $\boldsymbol{\theta}_t \in \mathbb{R}^m$, which is used to assign agent t a predicted outcome $\widehat{y}_t \in \mathbb{R}$. In our running example, $\widehat{y}$ could correspond to the applicant's predicted college GPA if admitted. The predicted outcome is calculated based on certain observable/measured attributes of the agent, denoted by $\mathbf{x}_t \in \mathbb{R}^m$. For example, in case of a university applicant, these attributes may include the applicants' standardized test scores, high school math GPA, science GPA, humanities GPA, and their extracurricular activities. For simplicity, we assume all assessment rules are *linear*, that is, $\hat{y}_t = \mathbf{x}_t^\top \boldsymbol{\theta}_t$ for all t.

Measured vs. latent variables. We assume that the agent best-responds to the assessment rule $\boldsymbol{\theta}_t$ by strategically modifying their observable attributes $\mathbf{x}_t$ to receive a favorable predicted outcome. Often agents cannot modify the value of their measured attributes (e.g., SAT score) directly, but only through investing effort in certain activities that are difficult to measure. For example, a student might take standardized test preparation courses to improve their SAT scores, or they may spend time studying the respective subjects to improve their math and humanities GPA.

Latent variable: effort investments. We formalize the above with a vector $\mathbf{a}_t \in \mathbb{R}^d$, which denotes the unobservable *efforts* agent t invests in d activities in response to the assessment rule $\boldsymbol{\theta}_t$. We assume there exists a matrix W_t which maps effort vectors to changes in observable attributes. The (k, j)-th entry of this effort conversion matrix defines the change in the k-th observable attribute $x_{t,k}$ for a one unit increase in the jth coordinate of the effort vector $\mathbf{a}_t$.

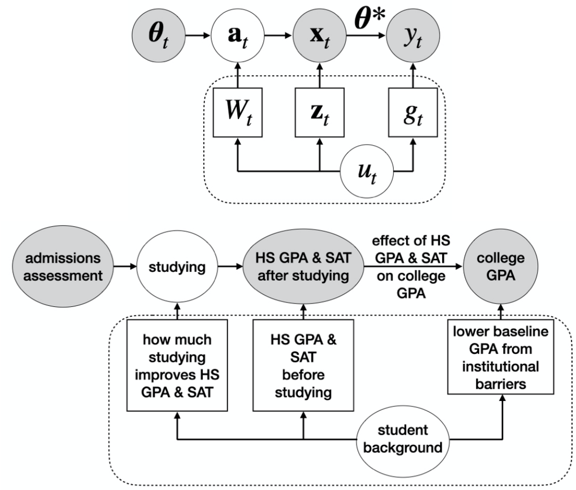


Figure 1: Graphical model for our setting (top) along with the way it corresponds to the admissions running example (bottom). Grey nodes are observed, white unobserved. Observable features $\mathbf{x}_t$ (e.g. high school GPA, SAT scores, etc.) depend on both the agent's private type u_t (e.g. a student's background —whether they have family who went to college, their gender, race, ethnicity, socioeconomic status, etc.) via initial features $\mathbf{z}_t$ (e.g. the SAT score or HS GPA student t would get without studying) and effort conversion matrix W_t (e.g. how much studying translates to an increase in SAT score for student t) and assessment rule $\boldsymbol{\theta}_t$ via action $\mathbf{a}_t$, which could correspond to studying, taking an SAT prep course, etc). An agent's outcome y_t (e.g. college GPA) is determined by their observable features $\mathbf{x}_t$ (via causal relationship $\boldsymbol{\theta}^*$) and type u_t (via baseline outcome error term g_t, which could be lower for students from underserved groups due to institutional barriers, discrimination, etc).

Latent variables: agent types. Each agent t has an unobserved private type u_t that can impact both their observed attributes $\mathbf{x}_t$ and true outcomes y_t. (The type is the confounder we would like to correct for.) In our running example, the type may broadly refer to the student's relevant background factors that cannot be directly observed or measured. For example, the student's background can specify the socioeconomic background of the student (including whether they are the first generation in their family to go to college), as well as their innate talent and abilities.

Importantly, we assume the type u_t has *nested* in it several relevant latent characteristics of the agent, which we refer to using the tuple $(\mathbf{z}_t, W_t, g_t)$:

- Vector $\mathbf{z}_t \in \mathbb{R}^m$ specifies agent t's baseline measurement values. For example, it can specify the baseline values of high school grades and SAT score the student would have received without any effort spent studying or preparing for standardized tests.

- Matrix W_t specifies agent t's *effort conversion matrix*—that is, how various effort investments translate to changes in observable features.
- g_t summarized all other environmental factors that can impact the agent's true outcome when we control for observable attributes. For example, it may reflect the effect of the institutional barriers the student faces on their actual college GPA.

We assume agent t's observable features take the form $\mathbf{x}_t = \mathbf{z}_t + W_t\mathbf{a}_t$.

Agent best responses. We assume the agent selects their effort profile $\mathbf{a}_t$ in order to maximize their predicted outcome $\hat{y}_t$, subject to some *effort* cost $c(\cdot)$ associated with modifying their observable attributes. In particular, we assume the cost function is quadratic, that $c(\mathbf{a}_t) = \frac{1}{2}\|\mathbf{a}_t\|_2^2$. This is a common assumption in the strategic classification literature (e.g., (22, 12, 3)). The agents select their effort $\mathbf{a}_t$ by solving the following optimization problem:

$$\mathbf{a}_t = \operatorname*{argmax}_{\mathbf{a}} \left\{ \hat{y}_t - \frac{1}{2}\|\mathbf{a}\|_2^2 \right\}$$

Given any deployed assessment rule $\boldsymbol{\theta}_t$, the agent's best-response effort is $\mathbf{a}_t = W_t^\top \boldsymbol{\theta}_t$.

True outcome model. *After* each round, the principal gets to observe the agent's true outcome $y_t \in \mathbb{R}$, which takes the form

$$y_t = \mathbf{x}_t^\top \boldsymbol{\theta}^* + g_t.$$

Here $\boldsymbol{\theta}^*$ is the *true* relationship between an agent's observable features and outcome. (Recall that $g_t \in \mathbb{R}$ captures the dependence of agent t's outcome y_t on unobservable or unmeasured factors.) Note that since $\mathbf{z}_t$, W_t, and g_t may be correlated with one another, ordinary least squares generally will *not* produce consistent estimator for $\boldsymbol{\theta}^*$ (see Appendix A.1 for more details).

1.2. Overview of results

We provide a general method to infer the causal relationship parameter $\boldsymbol{\theta}^*$. We make the novel observation that the principal's assessment rule $\boldsymbol{\theta}_t$ is a valid *instrument*, and leverage this observation to recover $\boldsymbol{\theta}^*$ via two-stage least squares regression (2SLS). Our method applies to both *off-policy* and *on-policy* settings: one can directly apply 2SLS on historical data $\{(\boldsymbol{\theta}_t, \mathbf{x}_t, y_t)\}_{t=1}^T$, or the principal can intentionally deploy a sequence of varying assessment rules (e.g., by making small perturbations on a fixed rule) and then apply 2SLS on the collected data.

2. IV regression in the strategic learning setting

Instrumental variable (IV) regression allows for consistent estimation of the relationship between an outcome and observable features in the presence of confounding terms. We focus on two-stage least-squares regression (2SLS), a kind of IV estimator. 2SLS independently estimates the relationship between an *instrumental variable* $\boldsymbol{\theta}_t$ and the observable features $\mathbf{x}_t$, as well as the relationship between $\boldsymbol{\theta}_t$ and the outcome y_t via simple least squares regression. In this setting, we view the assessment rules $\{\boldsymbol{\theta}\}_{t=1}^T$ as algorithmic instruments and perform IV regression to estimate the true causal parameter $\boldsymbol{\theta}^*$. There are two criteria for $\boldsymbol{\theta}_t$ to be a valid instrument: (1) $\boldsymbol{\theta}_t$ influences the observable features $\mathbf{x}_t$, and (2) $\boldsymbol{\theta}_t$ is independent from the private type u_t. By design, criterion (1) is satisfied. We aim to design a mechanism that satisfies criterion (2) by choosing assessment rule $\boldsymbol{\theta}_t$ randomly, independent of the private type u_t. As can be seen by Figure 1, the principal's assessment rule $\boldsymbol{\theta}_t$ satisfies these criteria.

Formally, given a set of observations $\{\boldsymbol{\theta}_t, \mathbf{x}_t, y_t\}_{t=1}^T$, we compute the estimate $\widehat{\boldsymbol{\theta}}$ of the true casual parameters $\boldsymbol{\theta}$ from the following process of two-stage least squares regression (2SLS). We use $\widetilde{\boldsymbol{\theta}}_t$ to denote the vector $\begin{bmatrix}\boldsymbol{\theta}_t & 1\end{bmatrix}^\top$.

1. Estimate $\Omega = \mathbb{E}[W_t W_t^\top]$, $\mathbb{E}[\mathbf{z}_t^\top]$ using $\begin{bmatrix}\widehat{\Omega} \\ \bar{\mathbf{z}}^\top\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t \mathbf{x}_t^\top$
2. Estimate $\boldsymbol{\lambda} = \Omega\boldsymbol{\theta}^*$, $(\mathbb{E}[g_t] + \mathbb{E}[\mathbf{z}_t^\top]\boldsymbol{\theta}^*)$ using $\begin{bmatrix}\widehat{\boldsymbol{\lambda}} \\ \bar{g} + \bar{\mathbf{z}}^\top\boldsymbol{\theta}^*\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t y_t$
3. Estimate $\boldsymbol{\theta}^*$ as $\widehat{\boldsymbol{\theta}} = \widehat{\Omega}^{-1}\widehat{\boldsymbol{\lambda}}$,

We assume that $\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top$ is invertible, as is standard in the 2SLS literature. For proof that IV regression produces a consistent estimator of $\boldsymbol{\theta}^*$, see Appendix A.3.

Theorem 2.1. *Given a sequence of bounded assessment rules $\{\boldsymbol{\theta}_t\}_{t=1}^T$ and the (observable feature, outcome) pairs $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ they induce, the distance between the true causal parameters $\boldsymbol{\theta}^*$ and the estimate $\widehat{\boldsymbol{\theta}}$ obtained via IV regression is bounded as*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}\left(\frac{\sqrt{T}}{\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)}\right)$$

with high probability, if g_t is a bounded random variable.

Proof Sketch. See Appendix B.1 for the full proof. The bound follows by substituting our expressions for $\mathbf{x}_t$, y_t into the IV regression estimator, applying the Cauchy-Schwarz inequality to split the bound into two terms (one dependent on $\{(\boldsymbol{\theta}_t, \mathbf{x}_t)\}_{t=1}^T$ and one dependent on $\{(\boldsymbol{\theta}_t, g_t)\}_{t=1}^T$), and using a Chernoff bound to bound the term dependent on $\{(\boldsymbol{\theta}_t, g_t)\}_{t=1}^T$ with high probability.

While in some settings, the principal may only have access to *observational* data, in other settings, the principal may be able to actively deploy assessment rules on the agent population. We show that in scenarios in which this is possible, the principal can play random assessment rules centered around some "reasonable" assessment rule to achieve an $\mathcal{O}\left(\frac{1}{\sigma_\theta^2\sqrt{T}}\right)$ error bound on the estimated causal relationship $\widehat{\boldsymbol{\theta}}$, where σ_θ^2 is the variance in each coordinate of $\boldsymbol{\theta}_t$. Note that while playing random assessment rules may be seen as unfair in some settings, the principal is free to set the variance parameter σ_θ^2 to an "acceptable" amount for the domain they are working in. We formalize this notion in the following corollary.

Corollary 2.2. *If each $\theta_{t,j}$, $j \in 1, \ldots, m$, is drawn independently from some distribution $\mathcal{P}_j$ with variance σ_θ^2, $\mathbf{z}_t$ and W_t are bounded random variables, $W_t W_t^\top$ is full-rank, and $\sigma_{min}(\mathbb{E}[W_t W_t^\top]) > 0$, then, with high probability,*

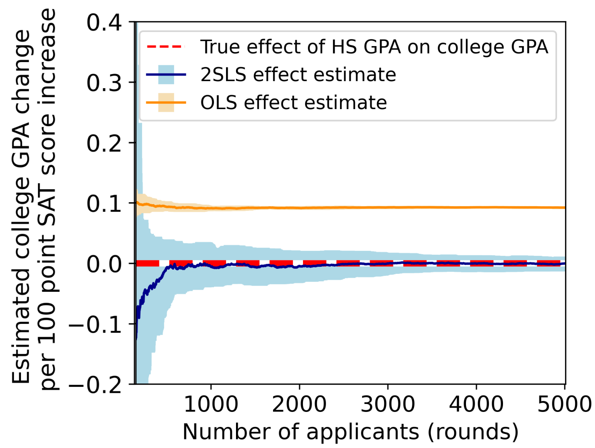
$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = \mathcal{O}\left(\frac{1}{\sigma_\theta^2\sqrt{T}}\right).$$

Proof Sketch. We begin by breaking up $\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)$ into two terms, $\|A\|_2$ and $\sigma_{min}(B)$, where A and B are functions of $\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top$. We use the Chernoff and matrix Chernoff inequality to bound $\|A\|_2$ and $\sigma_{min}(B)$ with high probability respectively. For the full proof, see Appendix B.3.

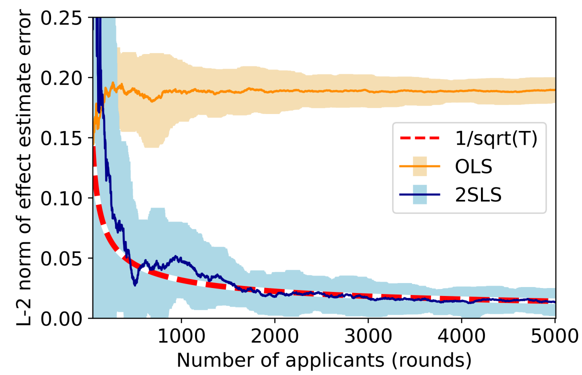
3. Experiments

We experimentally validate our methods on a semi-synthetic dataset based on real university admissions data. See Appendix C for the full description.

Results. In Figure 2a, we compare the true effect of SAT on college GPA ($\boldsymbol{\theta}^*$) with the estimate of this quantity given by our method of 2SLS from Section 2 ($\hat{\boldsymbol{\theta}}_{\text{2SLS}}$) and with the estimate given by OLS ($\hat{\boldsymbol{\theta}}_{\text{OLS}}$). In Figure 2b, we compare the estimation errors of OLS and 2SLS, i.e. $\|\widehat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\|_2$ and $\|\widehat{\boldsymbol{\theta}}_{\text{2SLS}} - \boldsymbol{\theta}^*\|_2$. We find that our 2SLS method converges to the true effect parameters (at a rate of about $\frac{1}{\sqrt{T}}$), whereas OLS has a constant bias. Although our setting assumes that SAT score has no causal effect on college GPA, OLS mistakenly predicts that, on average, a 100 point increase in SAT score leads to about a 0.1 point increase in college GPA. If SAT were not causally related to college performance in real life, these biased estimates could lead universities to erroneously use SAT scores in admissions decisions. This highlights the advantage of our method since it recovers causal relationships to avoid using arbitrary assessments, especially in the presence of confounding.



(a) True effect of SAT on college GPA vs. OLS and 2SLS. OLS versus 2SLS estimates for SAT effect on college GPA over 5000 rounds.



(b) Effect estimate error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ for OLS and 2SLS. OLS effect estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\|_2$ (in orange) and 2SLS estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{2SLS}} - \boldsymbol{\theta}^*\|_2$ (in blue) over 5000 rounds.

Figure 2: Evaluation of strategic IV regression on our semi-synthetic university admissions data. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation.

4. Conclusion

We established the possibility of recovering the causal relationship between observable attributes and the outcome of interest in settings where a decision-maker utilizes a series of linear assessment rules to evaluate strategic individuals. Our key observation was that in such settings, assessment rules serve as valid instruments. (Since they causally impact observable attributes but don't directly cause changes in the outcome.) This observation enables us to present a 2SLS method to correct for confounding bias in causal estimates.

References

[1] Allensworth, E. M. and Clark, K. High school gpas and act scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher*, 49(3):198–211, 2020. doi: 10.3102/0013189X20902110.

[2] Bernstein, D. S. *Matrix mathematics: theory, facts, and formulas*. Princeton university press, 2009.

[3] Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

[4] Dua, D. and Graff, C. UCI machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

[5] Goodman, J., Gurantz, O., and Smith, J. Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–158, 2020.

[6] Heidari, H., Nanda, V., and Gummadi, K. P. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.

[7] Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, 2019.

[8] Kusner, M. J. and Loftus, J. R. The long road to fairer algorithms, 2020.

[9] Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[10] Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

[11] Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[12] Mendler-Dünner, C., Perdomo, J. C., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. *arXiv preprint arXiv:2006.06887*, 2020.

[13] Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

[14] Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368. ACM, 2019.

[15] Pangburn, D. Schools are using software to help pick who gets in. what could go wrong? https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to\-help-pick-who-gets-in-what-could-go-wrong, 2019.

[16] Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[17] Petrasic, K., Saul, B., Greig, J., and Bornfreund, M. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

[18] (PSB), P. S. B. Penn state beaver admissions faq, 2021. URL https://beaver.psu.edu/admissions/faq.

[19] Rice, L. and Swesnik, D. Discriminatory effects of credit scoring on communities of color. *Suffolk UL Rev.*, 46:935, 2013.

[20] Sánchez-Monedero, J., Dencik, L., and Edwards, L. What does it mean to'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 458–468, 2020.

[21] Schölkopf, B. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[22] Shavit, Y., Edelman, B., and Axelrod, B. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[23] Somvichian-Clausen, A. Experts see new roles for artificial intelligence in college admissions process. *The Hill*, 2021.

[24] Sweeney, L. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[25] Yeh, I.-C. and hui Lien, C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.

Expert Systems with Applications, 36(2, Part 1): 2473–2480, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.12.020. URL https://www.sciencedirect.com/science/article/pii/S0957417407006719.

A. Parameter estimation in the causal setting

A.1. Ordinary least squares is not consistent

The least-squares estimate of $\boldsymbol{\theta}^*$ is given as

$$\widehat{\boldsymbol{\theta}}_{LS} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t y_t.$$

However, $\widehat{\boldsymbol{\theta}}_{LS}$ is not a consistent estimator of $\boldsymbol{\theta}^*$. To see this, let us plug in our expression for y_t into our expression for $\widehat{\boldsymbol{\theta}}_{LS}$. We get

$$\widehat{\boldsymbol{\theta}}_{LS} = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t (\mathbf{x}_t^\top \boldsymbol{\theta}^* + g_t)$$

After distributing terms and simplifying, we get

$$\widehat{\boldsymbol{\theta}}_{LS} = \boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t g_t.$$

$\mathbf{x}_t$ and g_t are not independent due to their shared dependence on the agent's private type u_t. Because of this, $\left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{T} \mathbf{x}_t g_t$ will generally not equal $\mathbf{0}_m$, even as the number of data points (agents) grows large. To see this, recall that $\mathbf{x}_t = \mathbf{z}_t + W_t \mathbf{a}_t$, so $\sum_{t=1}^{T} \mathbf{x}_t g_t = \sum_{t=1}^{T} (\mathbf{z}_t + W_t W_t^\top \boldsymbol{\theta}_t) g_t$. g_t and $\mathbf{z}_t$ are both determined by the agent's private type. Take the example where $\mathbf{z}_t = [g_t, 0, \dots, 0]^\top$. In this setting, $\sum_{t=1}^{T} \mathbf{z}_t g_t = [g_t^2, 0, \dots, 0]^\top$, which will always be greater than 0 unless $g_t = 0, \forall t$.

A.2. 2SLS derivations

Define $\widetilde{\boldsymbol{\theta}}_t = \begin{bmatrix} \boldsymbol{\theta}_t \\ 1 \end{bmatrix}$. $\mathbf{x}_t$ can now be written as $\mathbf{x}_t = \begin{bmatrix} W_t W_t^\top & \mathbf{z}_t \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_t \\ 1 \end{bmatrix}$.

Lemma A.1. *Using OLS, we can estimate* $\begin{bmatrix} \mathbb{E}[W_t W_t^\top] \\ \mathbb{E}[\mathbf{z}_t]^\top \end{bmatrix}$ *as*

$$\begin{bmatrix} \widehat{\Omega} \\ \bar{\mathbf{z}}^\top \end{bmatrix} = \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t \mathbf{x}_t^\top = \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \begin{bmatrix} \sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top \\ \sum_{t=1}^{T} \mathbf{x}_t^\top \end{bmatrix},$$

where $\widehat{\Omega} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top$.

Proof. In order to calculate $\widehat{\Omega}$, we will make use of the following fact:

Fact A.2 (Block Matrix Inversion ((2))). *If a matrix P is partitioned into four blocks, it can be inverted blockwise as* *follows:*

$$P = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{bmatrix},$$

where A and D are square matrices of arbitrary size, and B and C are conformable for partitioning. Furthermore, A and the Schur complement of A in P ($E = D - CA^{-1}B$) must be invertible.

Let $A = \sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top$, $B = \sum_{t=1}^{T} \boldsymbol{\theta}_t$, $C = \sum_{t=1}^{T} \boldsymbol{\theta}_t^\top$, and $D = \sum_{t=1}^{T} 1 = T$. Note that A is invertible by assumption and E is a scalar, so is trivially invertible unless $CA^{-1}B = T$.

Using this formulation, observe that

$$\bar{\mathbf{z}}^\top = -E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top + E^{-1}\sum_{t=1}^{T} \mathbf{x}_t^\top$$

and

$$\widehat{\Omega} = A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top + A^{-1}BE^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top - A^{-1}BE^{-1}\sum_{t=1}^{T} \mathbf{x}_t^\top$$

Rearranging terms, we see that $\widehat{\boldsymbol{\lambda}}$ can be written as

$$\widehat{\Omega} = A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top - A^{-1}B\bar{\mathbf{z}}^\top$$

Finally, plugging in for A and B, we see that

$$\begin{aligned} \widehat{\Omega} &= \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top - \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t \bar{\mathbf{z}}^\top \\ &= \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{z})^\top \end{aligned}$$

□

Similarly, we can write y_t as $y_t = \begin{bmatrix} \boldsymbol{\theta}_t^\top & 1 \end{bmatrix} \begin{bmatrix} W_t W_t^\top \boldsymbol{\theta}^* \\ g_t + \mathbf{z}_t^\top \boldsymbol{\theta}^* \end{bmatrix}$.

Lemma A.3. *Using OLS, we can estimate* $\begin{bmatrix} \mathbb{E}[W_t W_t^\top]\boldsymbol{\theta}^* \\ \mathbb{E}[g_t] + \mathbb{E}[\mathbf{z}_t^\top]\boldsymbol{\theta}^* \end{bmatrix}$ *as*

$$\begin{aligned} \begin{bmatrix} \widehat{\boldsymbol{\lambda}} \\ \bar{g} + \bar{\mathbf{z}}^\top \boldsymbol{\theta}^* \end{bmatrix} &= \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t y_t \\ &= \left(\sum_{t=1}^{T} \widetilde{\boldsymbol{\theta}}_t \widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1} \begin{bmatrix} \sum_{t=1}^{T} \boldsymbol{\theta}_t y_t^\top \\ \sum_{t=1}^{T} y_t^\top \end{bmatrix}, \end{aligned}$$

where $\widehat{\boldsymbol{\lambda}} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (y_t - \bar{g} - \bar{\mathbf{z}}^\top \boldsymbol{\theta}^*)$.

Proof. The proof follows similarly to the proof of the previous lemma. Let $A = \sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top$, $B = \sum_{t=1}^{T} \boldsymbol{\theta}_t$, $C = \sum_{t=1}^{T} \boldsymbol{\theta}_t^\top$, and $D = \sum_{t=1}^{T} 1 = T$. Note that A is invertible by assumption and E is a scalar, so is trivially invertible unless $CA^{-1}B = T$.

Using this formulation, observe that

$$\bar{g}^\top + \bar{z}^\top \boldsymbol{\theta}^* = -E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t + E^{-1}\sum_{t=1}^{T} y_t$$

and

$$\begin{aligned} \widehat{\boldsymbol{\lambda}} &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t + A^{-1}B\left(E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t - E^{-1}\sum_{t=1}^{T} y_t\right) \\ &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t y_t - A^{-1}B\left(\bar{g}^\top + \bar{z}^\top \boldsymbol{\theta}^*\right) \\ &= \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t \left(y_t - \bar{g}^\top - \bar{z}^\top \boldsymbol{\theta}^*\right) \end{aligned}$$

□

Theorem A.4. *We can estimate $\boldsymbol{\theta}^*$ as*

$$\widehat{\boldsymbol{\theta}} = \widehat{\Omega}^{-1}\widehat{\boldsymbol{\lambda}} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (y_t - \bar{g} - \bar{z}^\top \boldsymbol{\theta}^*)$$

Proof. This follows immediately from the previous two lemmas. □

A.3. 2SLS is consistent

Consider the two-stage least squares (2SLS) estimate of $\boldsymbol{\theta}^*$,

$$\widehat{\boldsymbol{\theta}}_{IV} = \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (y_t - \bar{g} - \bar{z}^\top \boldsymbol{\theta}^*)$$

Plugging in for y_t and simplifying, we get

$$\widehat{\boldsymbol{\theta}}_{IV} = \boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (g_t - \bar{g})$$

To see that $\widehat{\boldsymbol{\theta}}_{IV}$ *is* a consistent estimator of $\boldsymbol{\theta}^*$, we show that $\lim_{T\to\infty} \mathbb{E}\|\widehat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}^*\|_2^2 = 0$.

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}}_{IV} - \boldsymbol{\theta}^*\|_2^2 = \mathbb{E}\left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\theta}_t (g_t - \bar{g})\right\|_2^2$$

$g_t - \bar{g}$ and $\boldsymbol{\theta}_t$ are uncorrelated, so $\sum_{t=1}^{T} \boldsymbol{\theta}_t (g_t - \bar{g})$ will go to zero as $T \to \infty$. On the other hand, $\sum_{t=1}^{T} \boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top$ will approach $T\mathbb{E}[\boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top]$. $\boldsymbol{\theta}_t$ and $\mathbf{x}_t - \bar{\mathbf{z}}$ *are* correlated, so $\mathbb{E}[\boldsymbol{\theta}_t (\mathbf{x}_t - \bar{\mathbf{z}})^\top] \neq \mathbf{0}$ in general.

B. Causal parameter recovery derivations

B.1. Proof of Theorem 2.1

Recall that $\widehat{\boldsymbol{\theta}} = \left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \sum_{t=1}^T \boldsymbol{\theta}_t(y_t - \bar{g} - \bar{\mathbf{z}}^\top \boldsymbol{\theta}^*)$ from Appendix A.2. Plugging this into $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, we get

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 = \left\|\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \left(\sum_{t=1}^T \boldsymbol{\theta}_t(y_t - \bar{g} - \bar{\mathbf{z}}^\top\boldsymbol{\theta}^*)\right) - \boldsymbol{\theta}^*\right\|_2$$

Next, we substitute in our expression for y_t and simplify, obtaining

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 &= \left\|\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1} \left(\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \bar{g})\right)\right\|_2 \\
&\leq \left\|\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)^{-1}\right\|_2 \left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \bar{g})\right\|_2 \\
&\leq \frac{\left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \bar{g})\right\|_2}{\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{z}})^\top\right)}
\end{aligned}$$

We now bound the numerator and denominator separately with high probability.

B.2. Bound on numerator

$$\begin{aligned}
\left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \bar{g})\right\|_2 &= \left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \mathbb{E}[g_t] + \mathbb{E}[g_t] - \bar{g})\right\|_2 \\
&\leq \left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \mathbb{E}[g_t])\right\|_2 + \left\|\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbb{E}[g_t] - \bar{g})\right\|_2
\end{aligned}$$

B.2.1. Bound on first term

$$\left\|\sum_{t=1}^T \boldsymbol{\theta}_t(g_t - \mathbb{E}[g_t])\right\|_2 = \left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(g_t - \mathbb{E}[g_t])\right)^2\right)^{1/2}$$

Since $(g_t - \mathbb{E}[g_t])$ is a zero-mean bounded random variable with variance parameter σ_g^2, the product $\theta_{t,j}(g_t - \mathbb{E}[g_t])$ will also be a zero-mean bounded random variable with variance at most $\beta^2\sigma_g^2$. In order to bound $\left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(g_t - \mathbb{E}[g_t])\right)^2\right)^{1/2}$ with high probability, we make use of the following lemma. Note that bounded random variables are sub-Gaussian random variables.

Lemma B.1 (High probability bound on the sum of unbounded sub-Gaussian random variables)**.** *Let $x_t \sim subG(0, \sigma^2)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\sum_{t=1}^T x_t\right| \leq \sigma\sqrt{2T\log(1/\delta)}$$

Applying Lemma B.1 to $\left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(g_t - \mathbb{E}[g_t])\right)^2\right)^{1/2}$, we get

$$\begin{aligned}
\sqrt{\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(g_t - \mathbb{E}[g_t])\right)^2} &\leq \sqrt{\sum_{j=1}^m \left(\beta\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2} \\
&\leq \sqrt{\sum_{j=1}^m \beta^2\sigma_g^2 2T\log(m/\delta)} \\
&\qquad \text{(by a union bound, where } \delta_j = \delta/m \text{ for all } j) \\
&\leq \beta\sigma_g\sqrt{2Tm\log(m/\delta)}
\end{aligned}$$

with probability at least $1-\delta$.

B.2.2. Bound on second term

$$\begin{aligned}
\left\|\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbb{E}[g_t] - \bar{g})\right\|_2 &= \left\|\sum_{t=1}^T \boldsymbol{\theta}_t\left(\mathbb{E}[g_t] - \frac{1}{T}\sum_{s=1}^T g_s\right)\right\|_2 \\
&= \left\|\sum_{t=1}^T \boldsymbol{\theta}_t \frac{1}{T}\sum_{s=1}^T (\mathbb{E}[g_t] - g_s)\right\|_2 \\
&= \left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}\frac{1}{T}\sum_{s=1}^T (\mathbb{E}[g_t] - g_s)\right)^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^m \left(\sum_{t=1}^T |\theta_{t,j}|\frac{1}{T}\left|\sum_{s=1}^T \mathbb{E}[g_t] - g_s\right|\right)^2\right)^{1/2}
\end{aligned}$$

After applying Lemma B.1, we get

$$\begin{aligned}
\left\|\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbb{E}[g_t] - \bar{g})\right\|_2 &\leq \left(\sum_{j=1}^m \left(\sum_{t=1}^T |\theta_{t,j}|\frac{1}{T}\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^m \left(\beta\sigma_g\sqrt{2T\log(1/\delta_j)}\right)^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^m \beta^2\sigma_g^2 2T\log(m/\delta)\right)^{1/2} \\
&\leq \beta\sigma_g\sqrt{2Tm\log(m/\delta)}
\end{aligned}$$

with probability at least $1-\delta$

B.3. Proof of Corollary 2.2

Next let's bound the denominator. By plugging in the expression for $\mathbf{x}_t$, we see that

$$\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{x}_t-\bar{\mathbf{z}})^\top\right)=\sigma_{min}\left(A+B\right),$$

where $A = \sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t - \bar{\mathbf{z}})^\top$ and $B = \sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top W_tW_t^\top$. By definition,

$$\sigma_{min}(A+B)=\min_{\mathbf{a},\|\mathbf{a}\|_2=1}\|(A+B)\mathbf{a}\|_2.$$

Via the triangle inequality,

$$\begin{aligned}\sigma_{min}(A+B)&\geq\min_{\mathbf{a},\|\mathbf{a}\|_2=1}(\|B\mathbf{a}\|_2-\|A\mathbf{a}\|_2)\\&\geq\min_{\mathbf{a},\|\mathbf{a}\|_2=1}\|B\mathbf{a}\|_2-\|A\|_2\\&\geq\sigma_{min}(B)-\|A\|_2\\&.\end{aligned}$$

B.3.1. Bounding $\|A\|_2$

$$\begin{aligned}\|A\|_2&=\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t]+\mathbb{E}[\mathbf{z}_t]-\bar{\mathbf{z}})^\top\right\|_2\\&\leq\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t])^\top\right\|_2+\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[\mathbf{z}_t]-\bar{\mathbf{z}})^\top\right\|_2\end{aligned}$$

Bound on first term

$$\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t])^\top\right\|_2\leq\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t])^\top\right\|_F$$

$$\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t])^\top\right\|_2\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,i}(z_{t,j}-\mathbb{E}[z_{t,j}])\right)^2\right)^{1/2}$$

Notice that $\theta_{t,i}(z_{t,j}-\mathbb{E}[z_{t,j}])$ is a zero-mean bounded random variable with variance at most $\beta^2\sigma_z^2$. Applying Lemma B.1, we can see that

$$\begin{aligned}\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbf{z}_t-\mathbb{E}[\mathbf{z}_t])^\top\right\|_2&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2}\\&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2}\\&\leq\left(m^2\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2}\\&\leq m\beta\sigma_z\sqrt{2T\log(m^2/\delta)}\end{aligned}$$

with probability at least $1-\delta$.

Bound on second term

$$\begin{aligned}\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[\mathbf{z}_t]-\bar{\mathbf{z}})^\top\right\|_2&=\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[\mathbf{z}_t]-\mathbf{z}_j)^\top\right\|_2\\&\leq\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[\mathbf{z}_t]-\mathbf{z}_j)^\top\right\|_F\\&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}\theta_{t,i}\frac{1}{T}\sum_{s=1}^{T}(\mathbb{E}[z_{t,j}]-z_j)\right)^2\right)^{1/2}\\&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,i}|\frac{1}{T}\left|\sum_{s=1}^{T}(\mathbb{E}[z_{t,j}]-z_j)\right|\right)^2\right)^{1/2}\end{aligned}$$

By applying Lemma B.1, we obtain

$$\begin{aligned}\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t(\mathbb{E}[\mathbf{z}_t]-\bar{\mathbf{z}})^\top\right\|_2&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T}|\theta_{t,i}|\frac{1}{T}\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2}\\&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta\sigma_z\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2}\\&\leq\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\beta^2\sigma_z^2 2T\log(m^2/\delta)\right)^{1/2}\\&\leq m\beta\sigma_z\sqrt{2T\log(m^2/\delta)}\end{aligned}$$

B.3.2. Bounding $\sigma_{min}(B)$

Next we bound $\sigma_{min}(B) = \sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top W_tW_t^\top)$. We can write $W_tW_t^\top$ as $\mathbb{E}[W_tW_t^\top]+\epsilon_t$. Note that since each element of W_t is bounded, each element of $\epsilon_t\in\mathbb{R}^{m\times m}$ will be bounded as well. Using this formulation,

$$\begin{aligned}\sigma_{min}(B)&=\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top(\mathbb{E}[W_tW_t^\top]+\epsilon_t)\right)\\&=\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[W_tW_t^\top]+\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\epsilon_t\right)\\&\geq\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[W_tW_t^\top]\right)-\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\epsilon_t\right\|_2\\&\geq\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[W_tW_t^\top]\right)-\left\|\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\epsilon_t\right\|_F\end{aligned}$$

We proceed by bounding each term separately.

Bound on first term

$$\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[W_tW_t^\top]\right)\geq\sigma_{min}(\mathbb{E}[W_tW_t^\top])\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top)$$

Let $c = \sigma_{min}(\mathbb{E}[W_tW_t^\top])$. We assume that W_t is distributed such that $c>0$. Therefore,

$$\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[W_tW_t^\top]\right)\geq c\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top).$$

Next, we use the matrix Chernoff bound to bound $c\lambda_{min}(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top)$ with high probability.

Theorem B.2 (Matrix Chernoff)**.** *Consider a finite sequence $\{X_t\}_{t=1}^{T}$ of independent, random, Hermitian matrices with common dimension d. Assume that*

$$0 \le \lambda_{min}(X_t) \text{ and } \lambda_{max}(X_t) \le L \text{ for each index } t$$

Introduce the random matrix

$$Y = \sum_{t=1}^{T} X_t.$$

Define the minimum eigenvalue μ_{min} of the expectation $\mathbb{E}[Y]$:

$$\mu_{min} = \lambda_{min}(\mathbb{E}[Y]) = \lambda_{min}\left(\sum_{t=1}^{T} \mathbb{E}[X_t]\right)$$

Then,

$$P(\lambda_{min}(Y) \le (1-\epsilon)\mu_{min}) \le d\left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{\mu_{min}/L}$$

for $\epsilon \in [0, 1)$.

Let $Y = \sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top$. In our setting,

$$\begin{aligned}
\mu_{min} &= \lambda_{min}\left(\sum_{t=1}^{T} \mathbb{E}[\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top]\right) \\
&= T\lambda_{min}\left(\mathbb{E}[\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top]\right) \\
&= T\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m} + \mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]\right)
\end{aligned}$$

$\sigma_\theta^2 \mathbb{I}_{m\times m}$ and $\mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]$ commute, so

$$\begin{aligned}
\mu_{min} &= T\left(\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m}\right) + \lambda_{min}\left(\mathbb{E}[\boldsymbol{\theta}_t]\mathbb{E}[\boldsymbol{\theta}_t^\top]\right)\right) \\
&= T\lambda_{min}\left(\sigma_\theta^2 \mathbb{I}_{m\times m}\right) \\
&= T\sigma_\theta^2 \lambda_{min}\left(\mathbb{I}_{m\times m}\right) \\
&= T\sigma_\theta^2
\end{aligned}$$

$$\lambda_{max}(\boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top) = \beta m,$$

so let $L = \beta m$.

Picking $\epsilon = 1/2$ and applying the matrix Chernoff bound to $\lambda_{min}(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top)$, we obtain

$$P\left(\lambda_{min}\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right) \le \frac{1}{2}T\sigma_\theta^2\right) \le d\left(\frac{1}{2}e\right)^{-\frac{T\sigma_\theta^2}{2\beta m}}$$

By rearranging terms, we see that if $T \ge \frac{2\beta m}{\sigma_\theta^2 \log \frac{1}{2}e} \log \frac{d}{\delta}$, then

$$\lambda_{min}\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top\right) \ge \frac{1}{2}T\sigma_\theta^2$$

with probability at least $1 - \delta$.

Bound on second term

$$\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top \epsilon_t\right\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=1}^{T} \theta_{t,i}\theta_{t,j}\epsilon_{t,i,j}\right)^2\right)^{1/2}$$

Since each $\epsilon_{t,i,j}$ is a bounded zero-mean random variable, $\theta_{t,i}\theta_{t,j}\epsilon_{t,i,j}$ is also a bounded zero-mean random variable, with variance at most $\beta^4\sigma_W^2$ We can now apply Lemma B.1:

$$\begin{aligned}
\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t \boldsymbol{\theta}_t^\top \epsilon_t\right\|_F &\le \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\beta^2\sigma_W\sqrt{2T\log(1/\delta_{i,j})}\right)^2\right)^{1/2} \\
&\le \left(\sum_{i=1}^{m}\sum_{j=1}^{m} \beta^4\sigma_W^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\le \left(m^2\beta^4\sigma_W^2 2T\log(m^2/\delta)\right)^{1/2} \\
&\le m\beta^2\sigma_W\sqrt{2T\log(m^2/\delta)}
\end{aligned}$$

with probability at least $1 - \delta$.

Putting everything together

Putting everything together, we have that

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le \frac{2\beta\sigma_g\sqrt{2m\log(m/\delta)}}{\frac{1}{2}c\sqrt{T}\sigma_\theta^2 - m\beta^2\sigma_W\sqrt{2\log(m^2/\delta)} - 2m\beta\sigma_z\sqrt{2\log(m^2/\delta)}}$$

with probability at least $1 - 6\delta$.

C. Omitted experiments

In this section, we present additional details for our experiments in Section 3. At the end, we provide more information regarding the dataset and computation resources used.

C.1. University admissions full experimental description

We construct a semi-synthetic dataset based on an example of university admissions with disadvantaged and advantaged students from Hu et al. (7). From a real dataset of the high school (HS) GPA, SAT score, and college GPA of 1000 college students, we estimate the causal effect of observed features $[\text{SAT}, \text{HS GPA}]$ on college GPA to be $\boldsymbol{\theta}^* = [0.00085, 0.49262]^\top$ using OLS (which is assumed to be consistent, since we have yet to modify the data to include confounding). We then use this dataset to construct

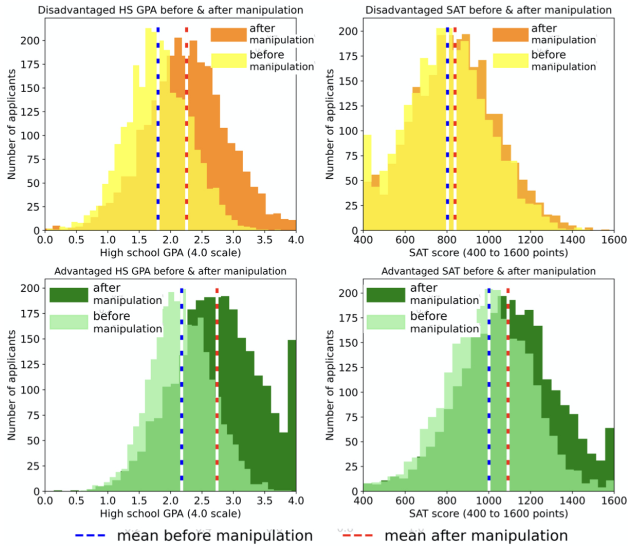


Figure 3: Distributions of unobserved features $\mathbf{z}$ (in lighter colors), i.e. initial HS GPA (two left figures) and SAT (two right figures), and observed features $\mathbf{x}$ (darker colors) for disadvantaged (two top figures in yellow and orange) and advantaged students (two bottom figures in green).

synthetic data which looks similar, yet incorporates confounding factors. For simplicity, we let the true causal effect parameters $\boldsymbol{\theta}^* = [0, 0.5]^\top$. That is, we assume there is a significant causal relationship between college performance and HS GPA, but not SAT score.[1] We consider two types of student backgrounds, those from a *disadvantaged group* and those from an *advantaged group*. We assume disadvantaged applicants have, on average, lower HS GPA and SAT $\mathbf{z}_t$, lower baseline college GPA g_t, and require more effort to improve observable features (reflected in W_t): this could be due to disadvantaged groups being systemically underserved, marginalized, or abjectly discriminated against (and the converse for advantaged groups). Initial features $\mathbf{z}_t$ are constructed as such: For any disadvantaged applicant t, their initial SAT features $z_t^{\text{SAT}} \sim \mathcal{N}(800, 200)$ and initial HS GPA $z_t^{\text{HS GPA}} \sim \mathcal{N}(1.8, 0.5)$. For any advantaged applicant t, $z_t^{\text{SAT}} \sim \mathcal{N}(1000, 200)$ and $z_t^{\text{HS GPA}} \sim \mathcal{N}(2.2, 0.5)$. We truncate SAT scores between 400 to 1600 and HS GPA between 0 to 4. For any applicant t, we randomly deploy assessment rule $\boldsymbol{\theta}_t = [\theta_t^{\text{SAT}}, \theta_t^{\text{HS GPA}}]^\top$ where $\theta_t^{\text{SAT}} \sim \mathcal{N}(1, 1)$ and $\theta_t^{\text{HS GPA}} \sim \mathcal{N}(7.5, 56.25)$. $\boldsymbol{\theta}_t$ need not be zero-mean, so universities can play a reasonable assessment rule with slight perturbations while still being able to perform unbiased causal estimation. Components of the average effort conversion matrix $\mathbb{E}[W_t]$ are smaller for disadvantaged applicants, which makes their mean improvement worse (see Figure 3). We set the expected effort conversion term

[1]Though this assumption may be contentious, it is based on existing research (1).

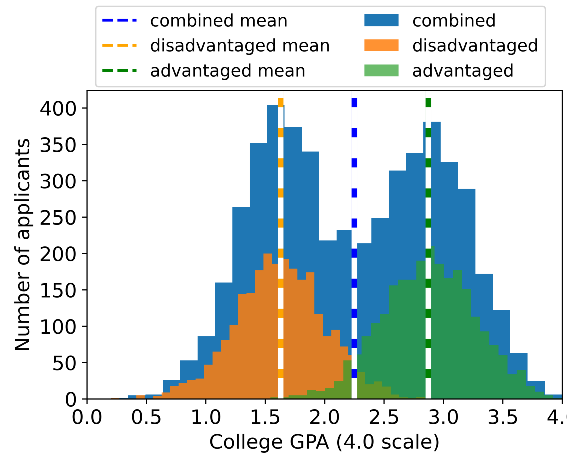


Figure 4: Outcome distributions for semi-synthetic datasets: college GPA for university admissions data. Distribution of college GPAs (outcomes y) for disadvantaged students (orange), advantaged students (green), and both combined (blue).

$\mathbb{E}[W_t W_t^\top] = \begin{pmatrix} 5 & 0.05 \\ 0.1 & 0.4 \end{pmatrix}$. Each row of $\mathbb{E}[W_t]$ corresponds to effort expended to change a specific feature. For example, entries in the first row of $\mathbb{E}[W_t]$ correspond to effort expended to change one's SAT score. For each applicant t, we perturb $\mathbb{E}[W_t W_t^\top]$ with random noise to produce $W_t W_t^\top$. We add noise to $\mathbb{E}[W_t W_t^\top]$ to produce $W_t W_t^\top$ for advantaged applicants and subtract for disadvantaged applicants: thus, it takes more effort, on average, for members of disadvantaged groups to improve their HS GPA and SAT scores than members of advantaged groups. Finally, we construct college GPA (true outcome y_t) by multiplying observed features $\mathbf{x}_t$ by the true effect parameters $\boldsymbol{\theta}^*$. We then add confounding error g_t where $g_t \sim \mathcal{N}(0.5, 0.2)$ for disadvantaged applicants and $g_t \sim \mathcal{N}(1.5, 0.2)$ for advantaged applicants. Disadvantage applicants could have lower baseline outcomes, e.g. due to institutional barriers or discrimination. While the setting we consider is simplistic, Figures 3 and 4 demonstrate that our semi-synthetic admissions data is somewhat realistic.[2]

C.2. Experimental Details

We evaluate our model on two semi-synthetic datasets: one based on our running university admission example (4) and the TAIWAN-CREDIT dataset obtained from the UCI Machine Learning Repository (25). These datasets are publicly available at `www.openintro.org/`

[2]For example, the mean shift in SAT scores from the first to second exam is 46 points (5). In our data, the mean shift for disadvantaged and advantaged applicants is about 36 points and 91 points, respectively.

data/index.php?data=satgpa and https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, respectively. These datasets do not contain personally identifiable information or offensive content. Since this is a publically available dataset, no consent from the people whose data we are using was required. We ran our experiments on a 2020 MacBook Air laptop with 16GB of RAM.

D. Comparison with Shavit et al.

The setting most similar to ours is that of Shavit et al.. They consider a strategic classification setting in which an agent's outcome is a linear function of features –some observable and some not (see Figure 5 for a graphical representation of their model). While they assume that an agent's hidden attributes can be modified strategically, we choose to model the agent as having an unmodifiable private *type*. Both of these assumptions are reasonable, and some domains may be better described by one model than the other. For example, the model of Shavit et al. may be useful in a setting such as car insurance pricing, where some unobservable factors which lead to safe driving are modifiable. On the other hand, settings like our college admissions example in which the unobservable features which contribute to college success (i.e. socioeconomic status, lack of resources, etc., captured in g_t) are not easily modifiable.

One benefit of our setting is that we are able to use $\boldsymbol{\theta}_t$ as a valid instrument to recover the true relationship $\boldsymbol{\theta}^*$ between observable features and outcomes. This is generally not possible in the model of (22), since $\boldsymbol{\theta}_t$ violates the backdoor criterion as long as there exists any hidden features $\mathbf{h}_t$ and is therefore not a valid instrument. Another difference between our setting and theirs is that we allow for a heterogeneous population of agents, while they do not. Specifically, they assume that each agent's mapping from actions to features is the same, while our model is capable of handling mappings which vary from agent-to-agent.

A natural question is whether or not there exists a general model which captures the setting of both Shavit et al. and ours. We provide such a model in Figure 6. In this setting, an agent has both observable and unobservable features, both of which are affected by the assessment rule $\boldsymbol{\theta}_t$ deployed and the agent's private type u_t. However, much like the setting of Shavit et al., $\boldsymbol{\theta}_t$ violates the backdoor criterion, so it cannot be used as a valid instrument in order to recover the true relationship between observable features and outcomes. Moreover, the following toy example illustrates that *no* form of true parameter recovery can be performed when an agent's unobservable features are modifiable.

Example D.1. *Consider the one-dimensional setting*

$$y_t = \theta^* x_t + \beta^* h_t,$$

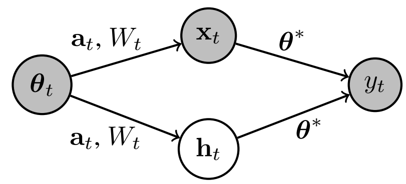


Figure 5: Graphical model of Shavit et al.. Observable features $\mathbf{x}_t$ (e.g. the type of car a person drives) and unobservable features $\mathbf{h}_t$ (e.g. how defensive of a diver someone is) are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$ (e.g. buying a new car) and common action conversion matrix W (representing, in part, the cost to a person of buying a new car). Outcome y_t (in this example, the person's chance of getting in an accident) is affected by $\mathbf{x}_t$ and $\mathbf{h}_t$ through the true causal relationship $\boldsymbol{\theta}_t$. Note that causal parameter recovery is not possible in this setting unless all features are observable.

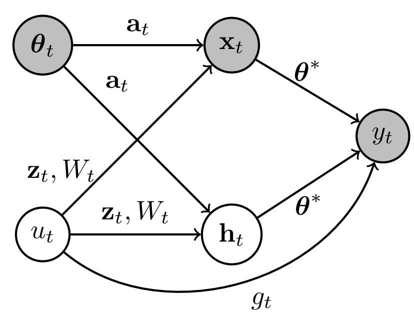


Figure 6: Graphical model which captures both our setting and that of Shavit et al.. In this setting, observable features $\mathbf{x}_t$ *and* unobservable features $\mathbf{h}_t$ are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$. The agent's private type u_t affects $\mathbf{x}_t$ and $\mathbf{h}_t$ through initial feature values $\mathbf{z}_t$ and action conversion matrix W_t. The agent's outcome y_t depends on $\mathbf{x}_t$ and $\mathbf{h}_t$ through the causal relationship $\boldsymbol{\theta}^*$ and u_t through confounding term g_t. Note that much like the setting of (22), causal parameter recovery is not possible in this setting unless all features are observable.

where x_t is an agent's observable, modifiable feature and h_t is an unobservable, modifiable feature. If the relationship between x_t and h_t is unknown, then it is generally impossible to recover the true relationship between x_t, h_t, and outcome y_t. To see this, consider the setting where h_t and x_t are highly correlated. In the extreme case, take $h_t = x_t, \ \forall t$. (Note we use equality to indicate identical feature values, not a causal relationship.) In this setting, the models $\theta^* = 1$, $\beta^* = 1$ and $\theta^* = 2$, $\beta^* = 0$ produce the same outcome y_t for all $x \in \mathbb{R}$, making it impossible to distinguish between the two models, even in the limit of infinite data.