Strategic Instrumental Variable Regression: Recovering Causal Relationships From Strategic Responses

Keegan Harris [1] **Daniel Ngo** [*2] **Logan Stapleton** [*2] **Hoda Heidari** [1] **Zhiwei Steven Wu** [1]

Abstract

In settings where Machine Learning (ML) algorithms automate or inform consequential decisions about people, individual decision subjects are often incentivized to strategically modify their observable attributes to receive more favorable predictions. As a result, the distribution the assessment rule is trained on may differ from the one it operates on in deployment. While such distribution shifts, in general, can 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, even under the presence of unobserved confounding variables. Specifically, our work establishes a novel connection between strategic responses to ML 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. We show that our causal recovery method can be utilized to improve decision-making across several important criteria: individual fairness, agent outcomes, and predictive risk. In particular, we show that if decision subjects differ in their ability to modify non-causal attributes, any decision rule deviating from the causal coefficients can lead to (potentially unbounded) individual-level unfairness.

[*]Equal contribution [1]School of Computer Science, Carnegie Mellon University, Pittsburgh, USA [2]Computer Science Department, University of Minnesota, Minneapolis, USA. Correspondence to: Keegan Harris <keeganh@cs.cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Machine learning (ML) predictions increasingly inform high-stakes decisions for people in areas such as college admissions (Pangburn, 2019; Somvichian-Clausen, 2021), credit scoring (Petrasic et al., 2017; Rice & Swesnik, 2013), employment (Sánchez-Monedero et al., 2020), 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 (Kusner et al., 2017). 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, harmful stereotypes—to achieve seemingly accurate predictions on held-out test data (Sweeney, 2013; Kusner & Loftus, 2020). The resulting models frequently underperform in deployment, and their predictions can negatively impact decision subjects. As an example of the long-term negative consequences of ML-based decision-making systems, they often prompt individuals to modify their observable attributes *strategically* to receive more favorable predictions—and subsequently, decisions (Hardt et al., 2016). These strategic responses are among the primary causes of distribution shifts (and subsequently, the unsatisfactory performance) of ML in high-stakes decision-making 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 (Liu et al., 2020; Heidari et al., 2019; Mouzannar et al., 2019).

The above challenges have led to renewed calls on the ML community to strengthen their understanding of the connections between ML and causality (Pearl, 2019; Schölkopf, 2019). 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 policy goals and societal values (e.g., they can audit the system for unfairness against particular populations (Loftus et al., 2018)). Finally, predictions rooted in

causal associations block undesirable 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 in the presence of 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 high-stakes decision-making settings such as the ones mentioned earlier, 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

Next, we describe our theoretical setup in further detail, then proceed to an overview of our findings. For concreteness, we utilize a stylized university admissions scenario as our running example for the remainder of this section. However, the reader should note that our model is applicable to other real-world applications in which confounders taint the causal interpretation of predictive models. For example, in credit lending, lack of access to affordable credit affects not only the applicant's debt, but also their likelihood of default (Collard & Kempson, 2005). In university admissions (which will be our running example), research has shown that the socioeconomic background of a student can impact both their SAT scores and success in college (Sackett et al., 2009).

With the running example in mind, consider a stylized setting in which a university decides whether to admit or reject applicants on a rolling basis[1] 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 the 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, $\hat{y}_t = \mathbf{x}_t^\top \boldsymbol{\theta}_t + \hat{o}_t$ for all t. (Where $\hat{o}_t$ is the current estimate of the expected offset term $\mathbb{E}[o_t]$.) This linear setup corresponds to an instance of the *partially linear regression model* (originally due to Robinson (1988)), a commonly studied setting in both the causal inference and strategic learning literature (e.g., Shavit et al. (2020); Kleinberg & Raghavan (2020); Bechavod et al. (2021)).

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 more 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 variables: effort investments. We formalize the above hidden investments with a vector $\mathbf{a}_t \in \mathbb{R}^d$, capturing the unobservable efforts agent t invests in d *activities* in response to the assessment rule $\boldsymbol{\theta}_t$. We assume there exists a linear mapping $\mathcal{E}_t$ which translates efforts to changes in observable attributes for agent t. The (k, j)-th entry of this effort conversion matrix defines the change in the k-th observable attribute of agent t, $\mathbf{x}_t$, for one unit increase in the jth coordinate of their effort vector $\mathbf{a}_t$.

Latent variables: agent types. Each agent t has an unobserved private type $\mathbf{u}_t$ 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 type can specify their socioeconomic background factors (including the level of educational support they receive within their immediate family), as well as their interest and skills in specific subjects such as English or Mathematics.[2] Formally, we assume the type $\mathbf{u}_t$ characterizes several relevant latent attributes of the agent, which we refer to using the tuple $\mathbf{u}_t := (\mathbf{b}_t, \mathcal{E}_t, o_t)$:

- $\mathbf{b}_t \in \mathbb{R}^m$ specifies agent t's baseline observable attribute 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

[1] See Safier (2022) for a list of such universities in the United States.

[2] Note that later in Section 5, we use the terminology of agent *subpopulations*. Subpopulations are distinct from types in that subpopulations determine the *distribution* of types, but individual agents belonging to the same subpopulation may have different types. We will elaborate on this in Section 5.

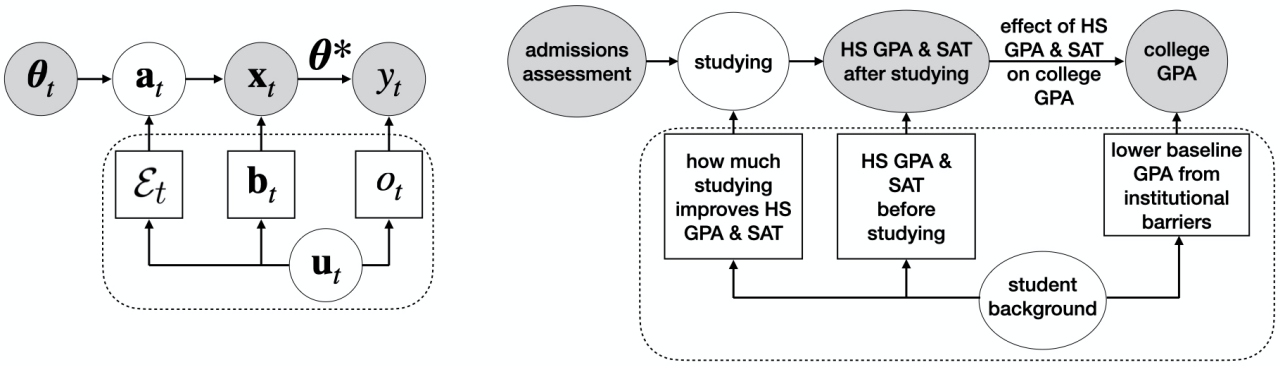

Figure 1. Graphical model for our setting (left) along with the way it corresponds to the admissions running example (right). Grey nodes are observed, white unobserved. Observable features $\mathbf{x}_t$ (e.g. high school GPA, SAT scores, etc.) depend on both the agent's private type $\mathbf{u}_t$ (e.g. a student's background) via initial features $\mathbf{b}_t$ (e.g. the SAT score or HS GPA student t would get without studying) and effort conversion matrix $\mathcal{E}_t$ (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 $\mathbf{u}_t$ (via baseline outcome error term o_t, which could be lower for students from underserved groups due to institutional barriers, discrimination, etc).

or preparing for standardized tests.

- $\mathcal{E}_t$ specifies agent t's *effort conversion matrix*—that is, how various effort investments in unobservable activities translate to changes in observable features.
- o_t summarizes 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 are affected by their type and effort investments. In particular, we assume they take the form $\mathbf{x}_t = \mathbf{b}_t + \mathcal{E}_t \mathbf{a}_t$.

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 $c(\mathbf{a}_t) = \frac{1}{2}\|\mathbf{a}_t\|_2^2$. (Note that this assumption is common in the strategic learning literature; see, e.g., (Shavit et al., 2020; Mendler-Dünner et al., 2020; Dong et al., 2018)). Formally, we assume agent t selects their effort $\mathbf{a}_t$ by solving the following optimization problem: $\max_{\mathbf{a}} \left\{\hat{y}_t - \frac{1}{2}\|\mathbf{a}\|_2^2\right\}$. It is easy to see that for a given deployed assessment rule $\boldsymbol{\theta}_t$, the agent's best-response effort investment is $\mathbf{a}_t = \mathcal{E}_t^\top \boldsymbol{\theta}_t$.

True causal outcome model. *After* each round, the principal gets to observe the agent's true outcome $y_t \in \mathbb{R}$, which takes the form $y_t = \mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t$. Here $\boldsymbol{\theta}^*$ is the *true causal* relationship between an agent's observable features and outcome. (Recall that $o_t \in \mathbb{R}$ captures the dependence of agent t's outcome y_t on unobservable or unmeasured factors.) We are interested in learning $\boldsymbol{\theta}^* \in \mathbb{R}^m$, which can be interpreted as specifying how interventions impacting the value of $\mathbf{x}$ lead to changes in y. Therefore, we say that an observable feature x_i is *causally relevant* if $\theta_i^* \neq 0$. For convenience, throughout we denote the subset of causally-relevant features by $\mathbf{x}_\mathcal{C}$, where $\mathcal{C} \subseteq [n], \forall i \in \mathcal{C}$ if $\theta_i^* \neq 0$.

1.2. Overview of Results

Strategic regression as instruments. Since $\mathbf{b}_t$, $\mathcal{E}_t$, and o_t may be correlated with one another, ordinary least squares generally will *not* produce a consistent estimator for $\boldsymbol{\theta}^*$ (see Appendix A.1 for details). 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.

Additionally, we show that our recovery of $\boldsymbol{\theta}^*$ can be utilized to improve decision-making across several desired criteria, namely, individual fairness, agent outcomes, and predictive risk.

(Non-)causal assessment rules and fairness. In Section 3, we analyze the individual-level disparities that may result if the assessment rule deviates from $\boldsymbol{\theta}^*$. Unlike most existing definitions of individual fairness, which rely on the observed characteristics of individuals, our definition measures the similarity between two individuals solely by comparing their $\mathbf{b}$'s and $\mathcal{E}$'s—that is, we consider two individuals to be similar if they have the same baseline values for causally relevant observable features and similar potentials for improv-

ing these observable attributes through effort investments. Individual fairness then requires similar individuals to receive similar decisions. (We note that while our notion of individual fairness may not be easy to estimate using observational data, it is a more fine-grained—and arguably better justified—notion of individual fairness, as it distinguishes between the causally relevant and causally irrelevant facets of observable features.) We show that when making predictions using $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, our notion of individual fairness is satisfied, but when the assessment rule deviates from $\boldsymbol{\theta}^*$, i.e., $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, individual fairness may be violated by an arbitrarily large amount.

Agent outcome maximization. Note that a decision-maker can use the assessment rule $\boldsymbol{\theta}$ as a form of intervention to incentivize agents to invest their efforts optimally toward maximizing their outcomes (y). In Section 4, we show that utilizing the causal parameters recovered during our 2SLS procedure, one can find the assessment rule maximizing expected agent outcomes.

Predictive risk minimization. Another commonly-studied goal for decision-makers is *predictive risk minimization*, which aims to minimize $\mathbb{E}[(\hat{y}_t - y_t)^2]$, the expected squared difference between an agent's *true* outcome and the outcome predicted by the assessment. Compared to standard regression, this is a more challenging objective since both the prediction $\hat{y}_t$ and outcome y_t depend on the deployed rule $\boldsymbol{\theta}_t$. This leads to a non-convex risk function. In Section 5.1, we show that the knowledge of $\boldsymbol{\theta}^*$ enables us to compute an unbiased estimate of the gradient of the predictive risk. As a result, we can apply stochastic gradient descent to find a local minimum of predictive risk function.

Empirical observations. In Section 5, we empirically confirm and illustrate the performance of our algorithm. In particular, for a semi-synthetic dataset inspired by our university admissions example, we observe that our methods consistently estimate the true causal relationship between observable features and outcomes (at a rate of $\mathcal{O}(1/\sqrt{T})$), whereas OLS does not. Notably, OLS mistakenly estimates that SAT is causally related to college GPA, even though our experimental setup assumes it is not. On the other hand, our 2SLS-based method avoids this erroneous estimation. We also show that our methods outperform standard SGD methods in the predictive risk minimization setting.

1.3. Related Work

An active area of research on *strategic learning* aims to develop machine learning algorithms that are capable of making accurate predictions about decision subjects even if they respond *strategically* and potentially *untruthfully* to the choice of the predictive model (Dong et al., 2018; Hardt et al., 2016; Mendler-Dünner et al., 2020; Shavit et al., 2020; Chen et al., 2020; Cummings et al., 2015; Cai et al., 2015; Meir et al., 2012; Hu et al., 2019). Generalizing strategic learning, Perdomo et al. (2020) propose a framework called *performative predictions*, which broadly studies settings in which the act of predicting influences the prediction target. Several recent papers have investigated the relationship between strategic learning and causality (Shavit et al., 2020; Bechavod et al., 2021; Miller et al., 2020).

The setting most similar to ours is that of Shavit et al. (2020). 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 8 in the appendix for a graphical representation of their model). While they assume that an agent's latent 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 related to safe driving are modifiable. On the other hand, our model captures settings like university admissions, where confounding factors (e.g., socioeconomic background) are not easily modifiable. Both models are special cases of a broader causal graph (described in Appendix G). Note that in the model of Shavit et al., $\boldsymbol{\theta}_t$ violates the backdoor criterion and therefore cannot serve as a valid instrument. (Bechavod et al., 2021) consider a setting simpler than ours in which there are no confounding effects from agents' unobserved types on their observable features and outcomes. As a result, the authors can apply standard least squares regression techniques to recover causal parameters.

Our work is also related to Miller et al. (2020), which shows that designing good incentives for agent improvement in strategic classification is at least as hard as orienting edges in the corresponding causal graph. In contrast to their work, we make the observation that the assessment rule deployed by the principal can be actively used as a valid *instrument*, which allows us to circumvent this hardness result by performing an *intervention* on the causal graph of the model.

Instrumental variable (IV) regression (Angrist & Krueger, 2001; Angrist & Imbens, 1995; Angrist et al., 1996) has mostly been used for observational studies (see e.g., (Angrist, 2005; Bloom et al., 1997)). Similar to ours, there is recent work on constructing instruments through dynamic action recommendations in multi-armed bandits settings (Ngo et al., 2021; Kallus, 2018). We consider an orthogonal direction: constructing instruments through assessment rules in the strategic learning setting.

2. IV Regression through Strategic Learning

Instrumental variable (IV) regression allows for consistent estimation of the relationship between an outcome and observable features in the presence of confounding terms. In this setting, we view the assessment rules $\{\boldsymbol{\theta}_t\}_{t=1}^T$ as algorithmic instruments and perform IV regression to estimate the true causal relationship $\boldsymbol{\theta}^*$. There are three criteria for $\boldsymbol{\theta}_t$ to be a valid instrument: (1) $\boldsymbol{\theta}_t$ influences the observable features $\mathbf{x}_t$, (2) $\boldsymbol{\theta}_t$ only influences the outcome y_t through $\mathbf{x}_t$, and (3) $\boldsymbol{\theta}_t$ is independent from the private type $\mathbf{u}_t$. By design, criterion (1) and (2) are satisfied. We aim to design a mechanism that satisfies criterion (3) by choosing the assessment rule $\boldsymbol{\theta}_t$ independently of the private type $\mathbf{u}_t$. As can be seen by Figure 1, the principal's assessment rule $\boldsymbol{\theta}_t$ satisfies these criteria.

We focus on two-stage least-squares regression (2SLS), a family of techniques for IV estimation. Intuitively, 2SLS can be thought of as estimating the causal relationship $\boldsymbol{\theta}^*$ between $\mathbf{x}_t$ and y_t by perturbing the instrument $\boldsymbol{\theta}$ and measuring the change in $\mathbf{x}_t$ and y_t. This enables us to account for the change in y_t *as a result of* the change in $\mathbf{x}_t$. 2SLS does this by independently estimating 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. For more background on the specific version of 2SLS we use, see Section 4.8 of (Cameron & Trivedi, 2005).

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 $\Omega = \mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top], \mathbb{E}[\mathbf{b}_t^\top]$ using
$$\begin{bmatrix}\widehat{\Omega} \\ \bar{\mathbf{b}}^\top\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\mathbf{x}_t^\top$$
2. Estimate $\boldsymbol{\lambda} = \Omega\boldsymbol{\theta}^*$, $(\mathbb{E}[o_t] + \mathbb{E}[\mathbf{b}_t^\top]\boldsymbol{\theta}^*)$ using
$$\begin{bmatrix}\widehat{\boldsymbol{\lambda}} \\ \bar{o} + \bar{\mathbf{b}}^\top\boldsymbol{\theta}^*\end{bmatrix} = \left(\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t\widetilde{\boldsymbol{\theta}}_t^\top\right)^{-1}\sum_{t=1}^T \widetilde{\boldsymbol{\theta}}_t y_t$$
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 $\boldsymbol{\theta}^*$ under our setting, 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 relationship $\boldsymbol{\theta}^*$ and the estimate $\widehat{\boldsymbol{\theta}}$ obtained via IV regression is bounded as*

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

with probability $1-\delta$, if o_t is a bounded random variable.

Proof Sketch. While similar bounds exist for traditional IV regression problems, they do not apply to the strategic learning setting we consider. 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 $\{(\boldsymbol{\theta}_t, o_t)\}_{t=1}^T$), and using a Chernoff bound to bound the term dependent on $\{(\boldsymbol{\theta}_t, o_t)\}_{t=1}^T$ with high probability.

While in some settings, the principal may only have access to *observational* (e.g., batch) 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, $\mathbf{b}_t$ and $\mathcal{E}_t$ are bounded random variables, $\mathcal{E}_t\mathcal{E}_t^\top$ is full-rank, and $\sigma_{min}(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]) > 0$, then*

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

with probability $1-\delta$.

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

3. (Un)fairness of (Non-)causal Assessments

While making predictions based on causal relationships is important from an ML perspective for reasons of generalization and robustness, the societal implications of using

non-causal relationships to make decisions are perhaps an even more persuasive reason to use causally-relevant assessments. In particular, it could be the case that a certain individual is worse at strategically manipulating features which are not causally relevant when compared to their peers. If these attributes are used in the decision-making process, this agent may be unfairly seen by the decision-maker as less qualified than their peers, even if their initial features and ability for improvement is similar to others.

One important criterion for assessing the fairness of a machine learning model at the individual level is that two individuals who have similar merit should receive similar predictions. Dwork et al. (2012) formalize this intuition through the notion of *individual fairness*, which is formally defined as follows.

Definition 3.1 (Individual Fairness (Dwork et al., 2012))**.** A mapping $M : \mathcal{U} \to \Delta(Y)$ is individually fair if for every $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have

$$D(M(\mathbf{u}), M(\mathbf{u}')) \leq d(\mathbf{u}, \mathbf{u}'),$$

where $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$ are individuals in population $\mathcal{U}$, $\Delta(Y)$ is the probability distribution over predictions Y, $D(M(\mathbf{u}), M(\mathbf{u}'))$ is a distance function which measures the similarity of the predictions received by $\mathbf{u}$ and $\mathbf{u}'$, and $d(\mathbf{u}, \mathbf{u}')$ is a distance function which measures the similarity of the two individuals.

Recall that in the setting we consider, the mapping between individuals and predictions is defined to be $M(\mathbf{u}) := \mathbf{x}^\top \boldsymbol{\theta} + \hat{o} = (\mathbf{b} + \mathcal{E}\mathcal{E}^\top)^\top \boldsymbol{\theta} + \hat{o}$. The prediction an individual receives is deterministic, so a natural choice for $D(M(\mathbf{u}), M(\mathbf{u}'))$ is $|\hat{y} - \hat{y}'|$. We take a causal perspective when defining a metric $d(\mathbf{u}, \mathbf{u}')$ to measure the similarity of two individuals $\mathbf{u}$ and $\mathbf{u}'$. Intuitively, individuals that have similar initial causally-relevant features and ability to modify causally-relevant features should be treated similarly. Therefore, we define $d(\mathbf{u}, \mathbf{u}')$ to reflect the difference in causally-relevant components of $\mathbf{b}$ & $\mathbf{b}'$ (initial feature values) and $\mathcal{E}\mathcal{E}^\top$ & $\mathcal{E}'\mathcal{E}'^\top$ (ability to manipulate features). With this in mind, we are now ready to define the criterion for individual fairness to be satisfied in the strategic learning setting.

Definition 3.2. In the strategic learning setting, individual fairness is satisfied if

$$\begin{aligned} |\hat{y} - \hat{y}'| &\leq d(\mathbf{u}, \mathbf{u}') \\ &= \|\mathbf{b}_{\mathcal{C}} - \mathbf{b}'_{\mathcal{C}}\|_2 + \|(\mathcal{E}\mathcal{E}^\top)_{\mathcal{C}} - (\mathcal{E}'\mathcal{E}'^\top)_{\mathcal{C}}\|_2, \end{aligned}$$

where

$$b_{\mathcal{C},i} = \begin{cases} b_i & \text{if } i \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases},$$

$$(\mathcal{E}\mathcal{E}^\top)_{\mathcal{C},ij} = \begin{cases} (\mathcal{E}\mathcal{E}^\top)_{ij} & \text{if } i \in \mathcal{C} \text{ or } j \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $\mathcal{C} \subseteq \{1, \ldots, n\}$ denotes the set of indices of observable features $\mathbf{x}$ which are causally relevant to y (i.e., $\theta^*_i \neq 0$ for $i \in \mathcal{C}$).

Theorem 3.3. *Assessment $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ satisfies individual fairness for any two agents $\mathbf{u}$ and $\mathbf{u}'$.*

Proof Sketch. See Appendix C for the full proof, which follows straightforwardly from the Cauchy-Schwarz inequality and the definition of the matrix operator norm. (Our results are not dependent upon the specific matrix or vector norms used, analogous results will hold for other popular choices of norm.) Throughout the proof we assume that $\|\boldsymbol{\theta}^*\|_2 = 1$ by definition, although our results hold up to constant multiplicative factors if this is not the case.

While $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ satisfies the criterion for individual fairness, this will generally not be the case for an arbitrary assessment $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. For instance, consider the case where $d(\mathbf{u}, \mathbf{u}') = 0$ for two agents $\mathbf{u}$ and $\mathbf{u}'$. Under this setting, it is possible to express $|\hat{y} - \hat{y}'|$ using quantities which do not depend on $d(\mathbf{u}, \mathbf{u}')$. As these quantities increase, $|\hat{y} - \hat{y}'|$ increases as well, despite the fact that $d(\mathbf{u}, \mathbf{u}')$ remains constant.

Theorem 3.4. *For any deployed assessment rule $\boldsymbol{\theta}$, the gap in predictions between two agents $\mathbf{u}$ and $\mathbf{u}'$ such that $d(\mathbf{u}, \mathbf{u}') = 0$ is*

$$\begin{aligned} |\hat{y} - \hat{y}'| = \Bigg| &\sum_{i \notin \mathcal{C}} (b_i - b'_i)\theta_i \\ &+ \sum_{i \notin \mathcal{C}} \sum_{j \notin \mathcal{C}} ((\mathcal{E}\mathcal{E}^\top)_{ij} - (\mathcal{E}'\mathcal{E}'^\top)_{ij})\theta_i\theta_j \Bigg| \end{aligned}$$

See Appendix C for the full derivation. Note that all components of $\boldsymbol{\theta}$ which appear in Theorem 3.4 are outside of the support of $\boldsymbol{\theta}^*$.

In order to illustrate how $|\hat{y} - \hat{y}'|$ can grow while $d(\mathbf{u}, \mathbf{u}')$ remains constant, consider the following example.

Example 3.5. Consider a setting in which the distance $d(\mathbf{u}, \mathbf{u}') = 0$ between agents $\mathbf{u}$ and $\mathbf{u}'$, and there is a one-to-one mapping between actions and observable features for each agent, with one agent having an advantage when it comes to manipulating features which are not causally relevant. Formally, let $\boldsymbol{\theta}^* = [\mathbf{0}^\top_{n/2} \;\; \sqrt{\frac{2}{n}}\mathbf{1}^\top_{n/2}]^\top$, $\mathbf{b} = \mathbf{b}'$, $\mathcal{E} = \boldsymbol{\delta} I_{n \times n}$, and $\mathcal{E}' = \boldsymbol{\delta}' I_{n \times n}$, where $\boldsymbol{\delta} = [\sqrt{n}\mathbf{1}^\top_{n/2} \;\; \mathbf{1}^\top_{n/2}]^\top$ and $\boldsymbol{\delta}' = [\mathbf{0}^\top_{n/2} \;\; \mathbf{1}^\top_{n/2}]^\top$.

Under such a setting, the equation in Theorem 3.4 simplifies to

$$|\hat{y} - \hat{y}'| = n \sum_{i \notin \mathcal{C}} \theta_i^2 = n \sum_{i=1}^{n/2} \theta_i^2.$$

For the full derivation, see Appendix C. Suppose now that the assessment $\boldsymbol{\theta}$ puts weight at least $1/\sqrt{n}$ on each observable feature which is not causally relevant. Under such a setting, $|\hat{y} - \hat{y}'| \geq n/2$, meaning that the difference in predictions tends towards infinity as n grows large, despite the fact that $d(\mathbf{u}, \mathbf{u}') = 0$ and $y = y'$!

4. Agent Outcome Improvements

In the strategic learning setting, the goal of each agent is clear: they aim to achieve the highest prediction $\hat{y}$ possible, regardless of their true label y. On the other hand, what the goal should be for the principal is less clear, and depends on the specific setting being considered. For example, in some settings it may be enough to discover the causal relationships between observable features and outcomes. However in other settings, the principal may wish to take a more active role. In particular, when making decisions which have real-world consequences, it may be in the principal's best interest to use a decision rule which promotes desirable behavior (Kleinberg & Raghavan, 2020; Shavit et al., 2020; Harris et al., 2021), i.e., behavior which has the potential to improve the *actual outcome* of an agent.

In the agent outcome improvement setting, the goal of the principal is to maximize the expected outcome $\mathbb{E}[y]$ of an agent drawn from the agent population. In our college admissions example, this would correspond to deploying an assessment rule with the goal of *maximizing* expected student college GPA. Formally, we aim to find $\boldsymbol{\theta}^{AO}$ in a convex set $\mathcal{S}$ of feasible assessment rules such that the induced expected agent outcome $\mathbb{E}[y]$ is maximized.

After some algebraic manipulation, the optimization becomes $\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \boldsymbol{\theta}^\top \boldsymbol{\lambda}$, where $\boldsymbol{\lambda} = \mathbb{E}[\mathcal{E}_t \mathcal{E}_t^\top] \boldsymbol{\theta}^*$.

For the full derivation, see Appendix D.1. Note that while the principal never directly observes $\mathbb{E}[\mathcal{E}_t \mathcal{E}_t^\top]$ nor $\boldsymbol{\theta}^*$, they estimate $\boldsymbol{\lambda} = \mathbb{E}[\mathcal{E}_t \mathcal{E}_t^\top] \boldsymbol{\theta}^*$ during the second stage of the 2SLS procedure. Therefore, if the principal has already run 2SLS to recover a sufficiently accurate estimate of the causal parameters $\boldsymbol{\theta}^*$, they can estimate the agent outcome-maximizing decision rule by solving the above optimization.

5. Experiments

We empirically evaluate our model on a semi-synthetic dataset inspired by our running university admissions example. We compare our 2SLS-based method against ordinary least squares (OLS), which directly regresses observed outcomes y on observable features $\mathbf{x}$. We show that even in our stylized setting with just two observable features, OLS does not recover $\boldsymbol{\theta}^*$, whereas our method does.

University admissions experimental description We constructed a semi-synthetic dataset based on the SATGPA dataset, a collection of real university admissions data.[3] The SATGPA dataset contains 6 variables on 1000 students. We use the following: two features (high school (HS) GPA and SAT score) and an outcome (college GPA). Using OLS (which is assumed to be consistent since we have yet to modify the data to include confounding), we find that the effect of $[\text{SAT}, \text{HS GPA}]$ on college GPA in this dataset is $\boldsymbol{\theta}^* = [0.0015, 0.5895]^\top$. We then construct synthetic data that is based on this original data, yet incorporates confounding factors. For simplicity, we let the true effect $\boldsymbol{\theta}^* = [0, 0.5]^\top$. That is, we assume HS GPA is causally related to college GPA, but SAT score is not.[4] We consider two private types of applicant backgrounds: *disadvantaged* and *advantaged*. Disadvantaged applicants have lower initial HS GPA and SAT ($\mathbf{b}$), lower baseline college GPA (o), and need more effort to improve observable features ($\mathcal{E}$).[5] Each applicants' initial features are randomly drawn from one of two Gaussian distributions, depending on back-

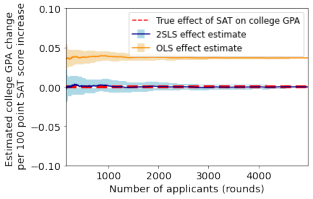

Figure 2. OLS versus 2SLS estimates for SAT effect on college GPA over 5000 rounds. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation. The red dashed line is the true causal relationship between SAT score and college GPA.

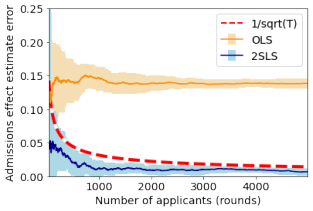

Figure 3. OLS effect estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{OLS}} - \boldsymbol{\theta}^*\|_2$ (in orange) and 2SLS estimate error $\|\widehat{\boldsymbol{\theta}}_{\text{2SLS}} - \boldsymbol{\theta}^*\|_2$ (in blue) over 5000 rounds. Results are averaged over 10 runs. Error bars (in lighter colors) represent one standard deviation. 2SLS estimate error decreases at a rate of about $\frac{1}{\sqrt{T}}$ (red dashed line).

[3] Originally collected by the Educational Testing Service, the SATGPA dataset is publicly available and can be found here: https://www.openintro.org/data/index.php?data=satgpa.

[4] Though this assumption may be contentious, it is based on existing research (e.g., Allensworth & Clark (2020)).

[5] For example, this could be due to the disadvantaged group being systemically underserved or marginalized (and the converse for advantaged group).

ground. Applicants may manipulate both of their features. See Appendix F for a full experimental description.

Results. In Figure 2, we compare the true effect of SAT score on college GPA ($\boldsymbol{\theta}^*$) with the estimates of these quantities given by our method of 2SLS from Section 2 ($\hat{\boldsymbol{\theta}}_{\text{2SLS}}$) and with the estimates given by OLS ($\hat{\boldsymbol{\theta}}_{\text{OLS}}$). (An analogous figure for the effects of HS GPA is included in the appendix.) In Figure 3, 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 causal relationship (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 relationship with college GPA, OLS mistakenly predicts that, on average, a 100 point increase in SAT score leads to about a 0.05 point increase in college GPA. If SAT were not causally related to collegiate 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 using a naive parameter estimation method like OLS in the presence of confounding could cause decision-making institutions to deploy assessments which don't accurately reflect the characteristics they are trying to measure.

5.1. Predictive Risk Minimization

Analogous to recovering causal relationships and improving agent outcomes, another common goal of the principal in the strategic learning setting is to *minimize predictive risk*. Formally, the goal of the principal in the predictive risk minimization setting is to learn the assessment rule that minimizes the expected squared difference between an agent's true outcome and the outcome predicted by the principal, i.e., $f(\boldsymbol{\theta}_t) = \mathbb{E}[(\widehat{y}_t - y_t)^2]$.

Due to the dependence of $\mathbf{x}_t$ and y_t on $\boldsymbol{\theta}_t$, $f(\boldsymbol{\theta}_t)$ will be non-convex in general, and can have several extrema which are not global minima, even in the case of just one observable feature. When faced with such non-convex optimization problems, gradient descent is often a popular approach due to its simplicity and convergence to local minima in practice.

If the effort conversion matrix $\mathcal{E}$ is the same for all agents, the gradient of population risk function can be written as

$$\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t) = 2(\mathbb{E}[(\widehat{y}_t - y_t)\mathbf{x}_t] + \mathbb{E}[\widehat{y}_t - y_t]\mathcal{E}\mathcal{E}^\top(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*).$$

See Appendix E.1 for the derivation. In our college admissions example, this would correspond to the setting in which all students' math GPA, SAT scores, etc. improve the same amount given the same effort: this may be a reasonable assumption if the students being considered have the same ability to learn, despite other differences in background they may have. If $\mathcal{E}\mathcal{E}^\top$ is known to the principal (e.g. through the 2SLS procedure in Section 2), then each $(\boldsymbol{\theta}_t, \mathbf{x}_t, y_t)$ tuple can be used to compute an unbiased estimate of $\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t)$ for use in online gradient descent.

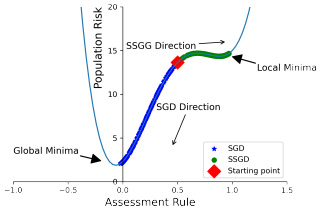

Figure 4. Stochastic Gradient Descent (SGD, takes into account x_t, y_t, $\widehat{y}_t$'s dependence on θ_t) vs Simple Stochastic Gradient Descent (SSGD, does not). In the 1D setting, it is possible for the gradient of SSGD to have the wrong sign. When both are initialized at $\theta_0 = 0.5$, SGD is able to follow the gradient and converge to the global minima, while SSGD is not. We ran each method for 1000 time-steps with a decaying learning rate of $\frac{0.001}{\sqrt{T}}$.

Recent work on *performative prediction* (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Miller et al., 2021) examines the use of repeated gradient descent in the strategic learning setting and finds that repeated gradient descent generally converges to *performatively stable* points. There is no direct comparison between performatively stable points and local minima in our setting. In fact, performatively stable points can actually *maximize* predictive risk under some settings. (See Miller et al. (2021) for such an example.) Our methods differ from this line of work because we take $\mathbf{x}_t$, y_t, and $\widehat{y}_t$'s *direct dependence* on the assessment rule $\boldsymbol{\theta}_t$ into account when calculating the gradient of the risk function, whereas these performative prediction models (henceforth *simple stochastic gradient descent* or *SSGD*) do not. While SSGD may be satisfactory for some settings, it produces a biased estimate of the gradient in general, which can lead to unexpected behavior under our setting; by contrast, our gradient estimate is unbiased (see Figure 4). Even in situations which SSGD *does* get the sign of the gradient correct, it may converge at a much slower rate, due to its incomplete estimate of the gradient (see Figure 12 in Appendix H).

6. Conclusion

In this work, we establish 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 is that in strategic settings, assessment rules serve as valid instruments (because they causally impact observable attributes but do not directly affect the outcome). This observation enables us to present a 2SLS method to correct for confounding bias in causal estimates. We then demonstrate the potential of the recovered

causal coefficients to be utilized for preventing individual-level disparities, improving agent outcomes, and reducing predictive risk minimization.

While our work offers an initial step toward extracting causal knowledge from automated assessment rules, we rely on several simplifying assumptions—all of which mark critical directions for future work. In particular, we assume all assessment rules and the underlying causal model are linear. This assumption allows us to utilize linear IV methods. Extending our work to *non-linear* assessment rules and IV methods is necessary for the applicability of our method to real-world settings. Another critical assumption is the agent's *full knowledge* of the assessment rule and their *rational* response to it, subject to a *quadratic effort cost*. While these are standard assumptions in economic modeling, they need to be empirically verified in the particular decision-making context at hand before our method's outputs can be viewed as reliable estimates of causal relationships.

7. Acknowledgements

ZSW, KH, DN, LS were supported in part by the NSF FAI Award #1939606, NSF SCC Award #1952085, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant. HH acknowledges support from NSF (#IIS2040929), J.P. Morgan, CyLab, Meta, and PwC. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and other funding agencies. Finally, the authors would like to thank Yonadav Shavit and John Miller for insightful conversations about Shavit et al. (2020) and Miller et al. (2021), respectively.

References

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. URL https://doi.org/10.3102/0013189X20902110.

Angrist, J. Instrumental variables methods in experimental criminological research: What, why, and how? Working Paper 314, National Bureau of Economic Research, September 2005. URL http://www.nber.org/papers/t0314.

Angrist, J. D. and Imbens, G. W. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535.

Angrist, J. D. and Krueger, A. B. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, December 2001. doi: 10.1257/jep.15.4.69. URL https://www.aeaweb.org/articles?id=10.1257/jep.15.4.69.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291629.

Bechavod, Y., Ligett, K., Wu, Z. S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1234–1242. PMLR, 2021. URL http://proceedings.mlr.press/v130/bechavod21a.html.

Bernstein, D. S. *Matrix mathematics: theory, facts, and formulas*. Princeton University Press, 2009.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *The Journal of Human Resources*, 32(3):549–576, 1997. ISSN 0022166X. URL http://www.jstor.org/stable/146183.

Cai, Y., Daskalakis, C., and Papadimitriou, C. H. Optimum statistical estimation with strategic data sources. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pp. 280–296. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Cai15.html.

Cameron, A. C. and Trivedi, P. K. *Microeconometrics: methods and applications*. Cambridge University Press, 2005.

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15265–15276. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae87a54e183c075c494c4d397d126a66-Paper.pdf.

Collard, S. and Kempson, E. *Affordable credit: The way forward*. Policy Press, 2005.

Cummings, R., Ioannidis, S., and Ligett, K. Truthful linear regression. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pp. 448–483. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Cummings15.html.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In Tardos, É., Elkind, E., and Vohra, R. (eds.), *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, pp. 55–70. ACM, 2018. doi: 10.1145/3219166.3219193. URL https://doi.org/10.1145/3219166.3219193.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In Goldwasser, S. (ed.), *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pp. 214–226. ACM, 2012. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Goodman, J., Gurantz, O., and Smith, J. Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–58, May 2020. doi: 10.1257/pol.20170503. URL https://www.aeaweb.org/articles?id=10.1257/pol.20170503.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pp. 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Harris, K., Heidari, H., and Wu, S. Z. Stateful strategic regression. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 28728–28741, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f1404c2624fa7f2507ba04fd9dfc5fb1-Abstract.html.

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. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2692–2701. PMLR, 2019. URL http://proceedings.mlr.press/v97/heidari19a.html.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In danah boyd and Morgenstern, J. H. (eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 259–268. ACM, 2019. doi: 10.1145/3287560.3287597. URL https://doi.org/10.1145/3287560.3287597.

Kallus, N. Instrument-armed bandits. In Janoos, F., Mohri, M., and Sridharan, K. (eds.), *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, volume 83 of *Proceedings of Machine Learning Research*, pp. 529–546. PMLR, 2018. URL http://proceedings.mlr.press/v83/kallus18a.html.

Kleinberg, J. M. and Raghavan, M. How do classifiers induce agents to invest effort strategically? *ACM Trans. Economics and Comput.*, 8(4):19:1–19:23, 2020. doi: 10.1145/3417742. URL https://doi.org/10.1145/3417742.

Kusner, M. and Loftus, J. The long road to fairer algorithms. *Nature*, 578:34–36, 02 2020. doi: 10.1038/d41586-020-00274-3.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4066–4076, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5\-Abstract.html.

Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. T. The disparate equilibria of algorithmic decision making when individuals invest rationally. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G. (eds.), *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January*

27-30, 2020, pp. 381–391. ACM, 2020. doi: 10.1145/3351095.3372861. URL https://doi.org/10.1145/3351095.3372861.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018. URL http://arxiv.org/abs/1805.05859.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.03.008. URL https://www.sciencedirect.com/science/article/pii/S000437021200029X.

Mendler-Dünner, C., Perdomo, J. C., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6917–6926. PMLR, 2020. URL http://proceedings.mlr.press/v119/miller20b.html.

Miller, J., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7710–7720. PMLR, 2021. URL http://proceedings.mlr.press/v139/miller21a.html.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In danah boyd and Morgenstern, J. H. (eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 359–368. ACM, 2019. doi: 10.1145/3287560.3287599. URL https://doi.org/10.1145/3287560.3287599.

Ngo, D. D. T., Stapleton, L., Syrgkanis, V., and Wu, S. Incentivizing compliance with algorithmic instruments. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8045–8055. PMLR, 2021. URL http://proceedings.mlr.press/v139/ngo21a.html.

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.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL http://proceedings.mlr.press/v119/perdomo20a.html.

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

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

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912705.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., and Waters, S. D. Socioeconomic status and the relationship between the sat® and freshman gpa: An analysis of data from 41 colleges and universities. research report no. 2009-1. *College Board*, 2009.

Safier, R. Complete list: Colleges with rolling admissions, 2022. URL https://blog.prepscholar.com/colleges-with-rolling-admissions.

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 Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G. (eds.), *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pp. 458–468. ACM, 2020.

doi: 10.1145/3351095.3372849. URL https://doi.org/10.1145/3351095.3372849.

Schölkopf, B. Causality for machine learning. *CoRR*, abs/1911.10500, 2019. URL http://arxiv.org/abs/1911.10500.

Shavit, Y., Edelman, B. L., and Axelrod, B. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8676–8686. PMLR, 2020. URL http://proceedings.mlr.press/v119/shavit20a.html.

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

Sweeney, L. Discrimination in online ad delivery. *ACM Queue*, 11(3):10, 2013. doi: 10.1145/2460276.2460278. URL https://doi.org/10.1145/2460276.2460278.

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

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

After distributing terms and simplifying, we get

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

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

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 $\mathbf{x}_t = \begin{bmatrix} \mathcal{E}_t \mathcal{E}_t^\top & \mathbf{b}_t \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_t \\ 1 \end{bmatrix}$.

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

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

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

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

Fact A.2 (Block Matrix Inversion ((Bernstein, 2009))). 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

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

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

$$\begin{aligned}\widehat{\Omega} &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top \\ &\quad + A^{-1}B(E^{-1}CA^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top - E^{-1}\sum_{t=1}^{T} \mathbf{x}_t^\top) \\ &= A^{-1}\sum_{t=1}^{T} \boldsymbol{\theta}_t \mathbf{x}_t^\top - A^{-1}B\bar{\mathbf{b}}^\top\end{aligned}$$

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

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

□

Similarly, we can write y_t as $y_t = \begin{bmatrix} \boldsymbol{\theta}_t^\top & 1 \end{bmatrix} \begin{bmatrix} \mathcal{E}_t \mathcal{E}_t^\top \boldsymbol{\theta}^* \\ o_t + \mathbf{b}_t^\top \boldsymbol{\theta}^* \end{bmatrix}$.

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

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

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

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

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

and

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

□

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

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

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

A.3. 2SLS is consistent

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

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

Plugging in for y_t and simplifying, we get

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

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

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

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

B. Causal parameter recovery derivations

B.1. Proof of Theorem 2.1

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

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

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

$$\begin{aligned}
&\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \\
&= \left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \right. \\
&\quad \left.\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t^\top\boldsymbol{\theta}^* + o_t - \bar{o} - \bar{\mathbf{b}}^\top\boldsymbol{\theta}^*)\right) - \boldsymbol{\theta}^*\right\|_2 \\
&= \left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\boldsymbol{\theta}^*\right)\right. \\
&\quad \left. + \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right) - \boldsymbol{\theta}^*\right\|_2 \\
&= \left\|\boldsymbol{\theta}^* + \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right) - \boldsymbol{\theta}^*\right\|_2 \\
&= \left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right)\right\|_2 \\
&\leq \left\|\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)^{-1}\right\|_2 \left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right\|_2 \\
&\leq \frac{\left\|\sum_{t=1}^{T} \boldsymbol{\theta}_t(o_t - \bar{o})\right\|_2}{\sigma_{min}\left(\sum_{t=1}^{T} \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right)}
\end{aligned}$$

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

B.2. Bound on numerator

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

B.2.1. Bound on first term

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

Since $(o_t - \mathbb{E}[o_t])$ is a zero-mean bounded random variable with variance parameter σ_g^2, the product $\theta_{t,j}(o_t -$

$\mathbb{E}[o_t]$) will also be a zero-mean bounded random variable with variance at most $\beta^2\sigma_g^2$. In order to bound $\left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(o_t - \mathbb{E}[o_t])\right)^2\right)^{1/2}$ 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 $\left(\sum_{j=1}^m \left(\sum_{t=1}^T \theta_{t,j}(o_t - \mathbb{E}[o_t])\right)^2\right)^{1/2}$, we get

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

with probability at least $1 - \delta$.

B.2.2. Bound on second term

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

After applying Lemma B.1, we get

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

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

$$\begin{aligned}
&\sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{x}_t - \bar{\mathbf{b}})^\top\right) \\
&= \sigma_{min}\left(\sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{b}_t - \bar{\mathbf{b}})^\top + \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathcal{E}_t\mathcal{E}_t^\top\right) \\
&= \sigma_{min}\left(A + B\right),
\end{aligned}$$

where $A = \sum_{t=1}^T \boldsymbol{\theta}_t(\mathbf{b}_t - \bar{\mathbf{b}})^\top$ and $B = \sum_{t=1}^T \boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathcal{E}_t\mathcal{E}_t^\top$. 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 $\|A\|_2$

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

Bound on first term

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

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

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

with probability at least $1-\delta$.

Bound on second term

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

By applying Lemma B.1, we obtain

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

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

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

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

We proceed by bounding each term separately.

Bound on first term

$$\begin{aligned}&\sigma_{min}\left(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top]\right)\\ &\ge \sigma_{min}(\mathbb{E}[\mathcal{E}_t\mathcal{E}_t^\top])\sigma_{min}(\sum_{t=1}^{T}\boldsymbol{\theta}_t\boldsymbol{\theta}_t^\top)\end{aligned}$$

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

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

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

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

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,

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

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

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

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

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

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

with probability at least $1-\delta$.

Bound on second term

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

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

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

with probability at least $1-\delta$.

Putting everything together

Putting everything together, we have that

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

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

C. Individual Fairness Derivations

C.1. Proof of Theorem 3.3

Proof.

$$\begin{aligned}&|\hat{y} - \hat{y}'| = \\ &|(\mathbf{b}-\mathbf{b}')^\top\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\top}(\mathcal{E}\mathcal{E}^\top - \mathcal{E}'\mathcal{E}'^\top)\boldsymbol{\theta}^*| \\ &= |(\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C})^\top\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\top}((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*| \\ &\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2\|\boldsymbol{\theta}^*\|_2 + \|\boldsymbol{\theta}^*\|_2\|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*\|_2 \\ &\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 + \max_{\boldsymbol{\theta},\|\boldsymbol{\theta}\|_2=1}\|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\boldsymbol{\theta}^*\|_2 \\ &\leq \|\mathbf{b}_\mathcal{C} - \mathbf{b}'_\mathcal{C}\|_2 + \|((\mathcal{E}\mathcal{E}^\top)_\mathcal{C} - (\mathcal{E}'\mathcal{E}'^\top)_\mathcal{C})\|_2\end{aligned}$$

□

C.2. Proof of Theorem 3.4

Proof. Let

$$b_{\widetilde{\mathcal{C}},i} = \begin{cases} b_i & \text{if } i \notin \mathcal{C} \\ 0 & \text{otherwise} \end{cases},$$

$$(\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}},ij} = \begin{cases} (\mathcal{E}\mathcal{E}^\top)_{ij} & \text{if } i,j \notin \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned}
&|\hat{y} - \hat{y}'| = |(\mathbf{b}_{\mathcal{C}} - \mathbf{b}'_{\mathcal{C}})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_{\mathcal{C}} - (\mathcal{E}'\mathcal{E}'^\top)_{\mathcal{C}})\boldsymbol{\theta} \\
&+ (\mathbf{b}_{\widetilde{\mathcal{C}}} - \mathbf{b}'_{\widetilde{\mathcal{C}}})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}}} - (\mathcal{E}'\mathcal{E}'^\top)_{\widetilde{\mathcal{C}}})\boldsymbol{\theta}| \\
&= |(\mathbf{b}_{\widetilde{\mathcal{C}}} - \mathbf{b}'_{\widetilde{\mathcal{C}}})^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top((\mathcal{E}\mathcal{E}^\top)_{\widetilde{\mathcal{C}}} - (\mathcal{E}'\mathcal{E}'^\top)_{\widetilde{\mathcal{C}}})\boldsymbol{\theta}| \\
&= \left| \sum_{i \notin \mathcal{C}} (b_i - b'_i)\theta_i + \sum_{i \notin \mathcal{C}} \sum_{j \notin \mathcal{C}} ((\mathcal{E}\mathcal{E}^\top)_{ij} - (\mathcal{E}'\mathcal{E}'^\top)_{ij})\theta_i\theta_j \right|
\end{aligned}$$

□

C.3. Example 3.5 Derivations

$$\begin{aligned}
d(\mathbf{u}, \mathbf{u}') &= \|b_{\mathcal{C}} - b'_{\mathcal{C}}\|_2 + \|(\mathcal{E}\mathcal{E}^\top)_{\mathcal{C}} - (\mathcal{E}'\mathcal{E}'^\top)_{\mathcal{C}}\|_2 \\
&= \|(\boldsymbol{\delta} - \boldsymbol{\delta}')_{\mathcal{C}} I_{n\times n}\|_2 = \|\mathbf{0}_{n\times n}\|_2 = 0,
\end{aligned}$$

where

$$\delta_{\mathcal{C},i} = \begin{cases} \delta_i & \text{if } i \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$$

$$|\hat{y} - \hat{y}'| = \left| 0 + \sum_{i \notin \mathcal{C}} (n-0)\theta_i^2 \right| = n \sum_{i=1}^{n/2} \theta_i^2$$

D. Agent outcome maximization derivations

D.1. Derivation of $\boldsymbol{\theta}^{AO}$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[y_t]$$

Substituting in for y_t:

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^* + o_t]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^*] + \mathbb{E}[o_t]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[\mathbf{x}_t^\top \boldsymbol{\theta}^*]$$

Substitute in for $\mathbf{x}_t$:

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[(\mathbf{b}_t^\top + \boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top)\boldsymbol{\theta}^*]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[\mathbf{b}_t^\top \boldsymbol{\theta}^*] + \mathbb{E}[\boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top \boldsymbol{\theta}^*]$$

$$\boldsymbol{\theta}^{AO} = \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \mathbb{E}[\boldsymbol{\theta}^\top \mathcal{E}_t \mathcal{E}_t^\top \boldsymbol{\theta}^*]$$

$$\begin{aligned}
\boldsymbol{\theta}^{AO} &= \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \boldsymbol{\theta}^\top \mathbb{E}[\mathcal{E}_t \mathcal{E}_t^\top]\boldsymbol{\theta}^* \\
&= \arg\max_{\boldsymbol{\theta} \in \mathcal{S}} \quad \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{C}} \sum_{k=1}^{d} \mathbb{E}[w_{ik} w_{jk}]\theta_i \theta_j^* \\
&\quad + \sum_{i \notin \mathcal{C}} \sum_{j \in \mathcal{C}} \sum_{k=1}^{d} \mathbb{E}[w_{ik} w_{jk}]\theta_i \theta_j^*
\end{aligned}$$

E. Predictive risk minimization derivations

E.1. Population gradient derivation

The gradient of the population risk function $f(\boldsymbol{\theta}_t) = \mathbb{E}[(\widehat{y}_t - y_t)^2]$ can be derived as follows

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_t} f(\boldsymbol{\theta}_t) &= \mathbb{E}[\nabla_{\boldsymbol{\theta}_t} (\widehat{y}_t - y_t)^2] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t} (\widehat{y}_t - y_t)] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t} (\mathbf{x}_t^\top \boldsymbol{\theta}_t - \mathbf{x}_t^\top \boldsymbol{\theta}^* - o_t)] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t} (\mathbf{x}_t^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)\nabla_{\boldsymbol{\theta}_t} ((\mathbf{b}_t^\top + \boldsymbol{\theta}_t \mathcal{E}_t \mathcal{E}_t^\top)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)(\mathbf{b}_t + \mathcal{E}_t \mathcal{E}_t^\top (2\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))] \\
&= 2\mathbb{E}[(\widehat{y}_t - y_t)(\mathbf{x}_t + \mathcal{E}_t \mathcal{E}_t^\top (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*))]
\end{aligned}$$

F. Omitted experiments

In this section, we present additional details for our experiments in Section 5. At the end, we provide more information regarding the dataset and computation resources used.

F.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. (2019). 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 [SAT, 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 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

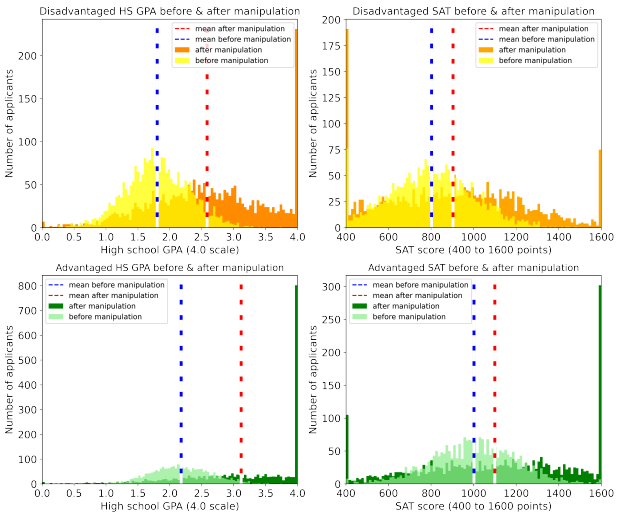

Figure 5. Distributions of unobserved features $\mathbf{b}$ (in lighter colors), i.e. initial HS GPA (two left figures) and SAT (two right figures), and observed features $\mathbf{x}$ (darker colors) for disadvantaged (two top figures in yellow and orange) and advantaged students (two bottom figures in green).

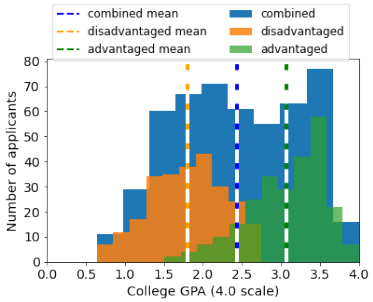

Figure 6. Distribution of college GPAs (outcomes y) for disadvantaged students (orange), advantaged students (green), and both combined (blue).

and HS GPA, but not SAT score.[6] 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 $\mathbf{b}_t$, lower baseline college GPA o_t, and require more effort to improve observable features (reflected in $\mathcal{E}_t$): this could be due to disadvantaged groups being systemically underserved, marginalized, or abjectly discriminated against (and the converse for advantaged groups). Initial features $\mathbf{b}_t$ 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 $\theta_t^{\text{SAT}} \sim \mathcal{N}(1, 10)$ and $\theta_t^{\text{HS GPA}} \sim \mathcal{N}(1, 2)$. $\boldsymbol{\theta}_t$ 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 $\mathbb{E}[\mathcal{E}_t]$ are smaller for disadvantaged applicants, which makes their mean improvement worse (see Figure 5). We set the expected effort conversion term $\mathbb{E}[\mathcal{E}_t] = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$ for simplicity. Each row of $\mathbb{E}[\mathcal{E}_t]$ corresponds to effort expended to change a specific feature. For example, entries in the first row of $\mathbb{E}[\mathcal{E}_t]$ correspond to effort expended to change one's SAT score. For each applicant t, we perturb $\mathbb{E}[\mathcal{E}_t]$ with random noise drawn from $\mathcal{N}(0.5, 0.25)$ to the top left entry and noise drawn from $\mathcal{N}(0.1, 0.01)$ the bottom right entry to produce $\mathcal{E}_t$. We add this noise to $\mathbb{E}[\mathcal{E}_t]$ to produce $\mathcal{E}_t$ 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 o_t where $o_t \sim \mathcal{N}(0.5, 0.2)$ for disadvantaged applicants and $o_t \sim \mathcal{N}(1.5, 0.2)$ 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 5 and 6 demonstrate that our semi-synthetic admissions data behaves reasonably.[7]

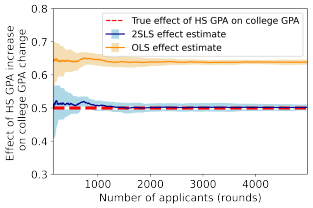

Figure 7. OLS versus 2SLS estimates for high school GPA effect on college GPA over 5000 rounds. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation. The red dashed line is the true causal effect of each high school GPA on college GPA.

F.2. Experimental Details

We evaluate our model on a semi-synthetic dataset based on our running university admission example (Dua &

[6] Though this assumption may be contentious, it is based on existing research (Allensworth & Clark, 2020).

[7] For example, the mean shift in SAT scores from the first to second exam is 46 points (Goodman et al., 2020). In our data, the mean shift for disadvantaged and advantaged applicants is about 36 points and 91 points, respectively.

Graff, 2019). The dataset we base our experiments off of is publicly available at www.openintro.org/data/index.php?data=satgpa. This dataset does not contain personally identifiable information or offensive content. Since this is a publicly 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.

G. 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 8 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 o_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 (Shavit et al., 2020), 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 9. 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.

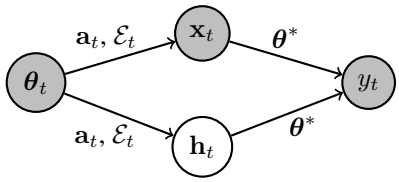

Figure 8. Graphical model of Shavit et al.. Observable features $\mathbf{x}_t$ (e.g. the type of car a person drives) and unobservable features $\mathbf{h}_t$ (e.g. how defensive of a diver someone is) are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$ (e.g. buying a new car) and common action conversion matrix $\mathcal{E}$ (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.

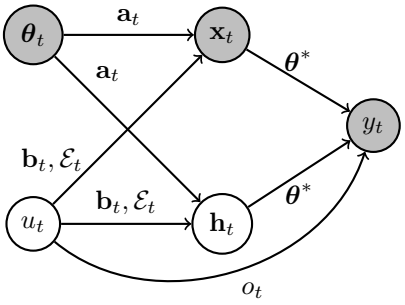

Figure 9. Graphical model which captures both our setting and that of Shavit et al.. In this setting, observable features $\mathbf{x}_t$ *and* unobservable features $\mathbf{h}_t$ are affected by $\boldsymbol{\theta}_t$ through action $\mathbf{a}_t$. The agent's private type u_t affects $\mathbf{x}_t$ and $\mathbf{h}_t$ through initial feature values $\mathbf{b}_t$ and action conversion matrix $\mathcal{E}_t$. 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 o_t. Note that much like the setting of (Shavit et al., 2020), causal parameter recovery is not possible in this setting unless all features are observable.

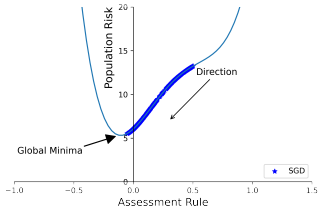

Figure 10. SGD with invex function

Example G.1. Consider the one-dimensional setting

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

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.

H. Setting of SGD comparison

In 1D, the derivative of $\mathbb{E}[(\widehat{y}_t - y_t)]^2$ which accounts for x_t and y_t's dependence on $\boldsymbol{\theta}_t$ takes the form

$$\Delta = 2(\mathbb{E}[(\widehat{y}_t - y_t)x_t] + \mathbb{E}_{x_t}[\widehat{y}_t - y_t]\mathcal{E}^2(\theta_t - \theta^*)).$$

By plugging in for x_t, y_t, $\widehat{y}_t$ and simplifying, we can write the derivative as

$$\Delta = 2\left(\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t] + \mathcal{E}^4\theta_t(\theta_t - \theta^*)^2\right).$$

The derivative of $\mathbb{E}[(\widehat{y}_t - y_t)]^2$ which does *not* account for x_t and y_t's dependence on $\boldsymbol{\theta}_t$ can be written as

$$\begin{aligned}\Delta' &= 2\left(\mathbb{E}[(\widehat{y}_t - y_t)x_t]\right) \\ &= 2\left(\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t]\right).\end{aligned}$$

As can be seen by comparing the two equations, there is an extra $\mathcal{E}^4\theta_t(\theta_t - \theta^*)^2$ term present in Δ that is not in Δ'. This can cause Δ and Δ' to have opposite signs under certain scenarios, e.g. when $\mathbb{E}[b_t^2] + \mathcal{E}^4\theta_t^2(\theta_t - \theta^*) - \mathbb{E}[o_t b_t]$ is negative and $\mathcal{E}^4\theta_t(\theta_t - \theta^*)^2$ is sufficiently large. To generate Figure 4, we set $\mathbb{E}[b_t] = 0$, $\mathbb{E}[b_t^2] = 0.3$, $\mathbb{E}[o_t] = 0$, $\mathbb{E}[g_t^2] = 15$, $\mathbb{E}[o_t b_t] = -6.5$, $\mathcal{E} = 3$, $\theta^* = 1$, and $\theta_0 = 0.5$. To generate Figure 10 and 11, we changed θ^* to be 0.7.

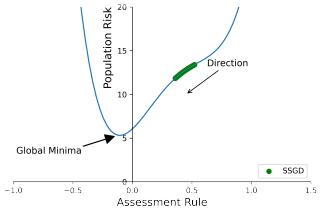

Figure 11. SSGD with invex function

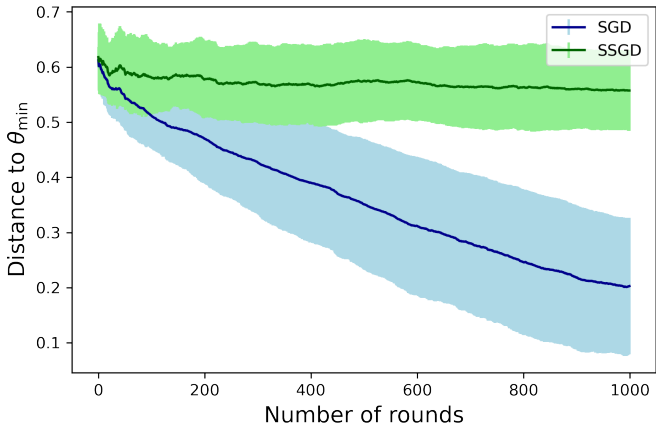

Figure 12. Convergence rate of Stochastic Gradient Descent vs Simple Stochastic Gradient Descent for simple 1D setting. Even when SSGD converges, it may do so at a much slower rate, due to the inexact measure of the gradient. We ran both methods for 1000 time-steps with a decaying learning rate of $\frac{0.001}{\sqrt{T}}$. Results are averaged over 10 runs, with the error bars (in lighter colors) representing one standard deviation.