Causal Feature Discovery through Strategic Modification

Yahav Bechavod* Katrina Ligett† Zhiwei Steven Wu‡ Juba Ziani§

February 18, 2020

Abstract

We consider an online regression setting in which individuals adapt to the regression model: arriving individuals may access the model throughout the process, and invest strategically in modifying their own features so as to improve their assigned score. We find that this strategic manipulation may help a learner recover the causal variables, in settings where an agent can invest in improving impactful features that also improve his true label. We show that even simple behavior on the learner's part (i.e., periodically updating her model based on the observed data so far, via least-square regression) allows her to simultaneously i) accurately recover which features have an impact on an agent's true label, provided they have been invested in significantly, and ii) incentivize agents to invest in these impactful features, rather than in features that have no effect on their true label.

*School of Computer Science and Engineering, The Hebrew University. Email: yahav.bechavod@cs.huji.ac.il.
†School of Computer Science and Engineering, The Hebrew University. Email: katrina@cs.huji.ac.il.
‡Department of Computer Science and Engineering, University of Minnesota. Email: zsw@umn.edu.
§Departmenet of Computer and Information Science, University of Pennsylvania. Email: jziani@seas.upenn.edu.

1 Introduction

As algorithmic decision-making takes a more and more important role in myriad application domains, incentives emerge to change the inputs presented to these algorithms—people may either invest in truly relevant attributes or strategically lie about their data. Recently, a collection of very interesting papers has explored various models of strategic behavior on the part of the classified individuals in learning settings, and ways to mitigate the harms to accuracy that can arise from falsified features [Dalvi et al., 2004, Brückner et al., 2012, Hardt et al., 2016, Dong et al., 2018]. Additionally, some recent work has focused on the design of learning algorithms that incentivize the classified individuals to make "good" investments in true changes to their variables [Kleinberg and Raghavan, 2019].

The present paper takes a different tack, and explores another potential effect of strategic investment in true changes to variables, in an online learning setting: we claim that interaction between the online learning and the strategic individuals may actually aid the learning algorithm in identifying causal variables. By *causal*, we mean, informally, variables such that changes in their true value cause changes in the true label and lead agents to improve. In contrast, *non-causal* variables do not affect the true label; such features are susceptible to gaming, as they can be used to obtain better outcomes with respect to the posted model without improving true labels.

The idea is quite simple. First, if a learning algorithm's hypothesis at a particular round depends heavily on a certain variable, this incentivizes the arriving individual to invest in improving that variable. If that variable were causally related to the true label, then the learner would observe the impact of these changes in the form of improved true labels. If that variable were non-causal, the changes would not have an effect on true labels. Second, if a learning algorithm improves its hypotheses over time, this changing sequence of incentives should encourage investment in a variety of promising variables, exposing those that are causal. This process should naturally induce the learner to shift its dependence towards causal variables, thereby incentivizing individuals to invest in meaningful changes, and resulting in an overall higher-quality population.

The goal of this paper is to highlight this potential beneficial effect of the interaction between online learning and strategic modification. To do so, we focus our study on a simple linear regression setting. In our model, there is a true underlying latent regression parameter vector β^*, and there is an underlying distribution over unmodified feature vectors. On every round t, the learner must announce a regression vector $\hat{\beta}_t$.[1] An individual then appears, with an unmodified feature vector x_t chosen i.i.d. from the distribution. Before presenting himself to the learner, the individual observes $\hat{\beta}_t$ and has the opportunity to invest in changing his true features to some $\bar{x}_t$; we focus on a simple model wherein the individual's investment results in a targeted change to a single variable. The individual then receives utility $\langle \hat{\beta}^t, \bar{x}_t \rangle$, and the learner gets feedback $\bar{y}_t = \beta^{*\intercal} \bar{x}_t + \varepsilon_t$, where ε_t is some noise.

Within this simple model, we consider simple behaviors for both the learner and the individuals: At each time t, the individual modifies his features so as to maximize his utility given the posted $\hat{\beta}_t$; periodically, the learner updates $\hat{\beta}_t$ with her best estimate of β^* given the (modified) features and labels she has observed, via least-square regression. Our main result is that under this simple behavior, the learner recovers β^* accurately, after observing sufficiently many individuals. Our result is divided in two parts: first, we show that least-square regression accurately recovers β^*

[1] Eventually, the learner we will consider does not update its regression vector at every round, but rather periodically, so that individuals can be treated in batches.

with respect to features that many individuals have invested in. Second, we show that these dynamics incentivize investments in every feature, leading to accurate recovery of β^* in its entirety, under an assumption on how the learner breaks ties between multiple least-square solutions. Our accuracy guarantees for a feature improve with the number of times that feature is invested in.

It is important to emphasize that we are studying a setting in which individuals' modifications of their variables can be meaningful investments (e.g., studying to achieve better mastery of material by an exam) rather than deceitful manipulations (e.g., cheating on the exam to achieve a higher assessment of that mastery). Strategic lying about variables would not help to expose causal variables, because such changes would not affect the outcome, regardless of whether the changes were in causal or non-causal variables.

Notice that any discovery of causal variables that occurs in our model is a result of the *interaction* between the online learner and the strategic individuals. On the one hand, online learning with no strategic response has no ability to distinguish non-causal variables from causal ones when the two are correlated. On the other hand, if strategic individuals faced with a static scoring algorithm tried to maximize their scores by investing in a non-causal feature, the resulting information would be insufficient for an observer to draw conclusions about the causality of other features.

For example, historical data might show that both a student's grades in high school and the make of car his parents drive to the university visit day are predictive of success in university. Suppose, for simplicity, that success in high school is causally related to success in university, but that make of parents' car is not. If the university admissions process put large weight on high school grades, that would incentivize students to invest effort in performing well in high school, which would also observably pay off in university, which would reinforce the emphasis on high school grades. If the admissions process put large weight on the make of car in which students arrive to the visit day, that would incentivize renting fancy cars for visits. However, this would result in a different distribution over the observed student variables, and on this modified distribution the correlation between cars and university success would be *weakened*, and therefore the admissions formula would not perform well. In future years, the university would naturally correct the formula to de-emphasize cars.

One reason that we find this natural process of causal variable discovery to be interesting is that discovery of causal variables is notoriously difficult and problematic. Separating correlation from causation in passive-observational data is essentially impossible without very strong assumptions [Eberhardt, 2007]. The gold standard traditional method for detecting causal variables is therefore to perform an intervention, and the protypical intervention is the randomized, controlled trial, a concept that grew out of the foundational work of R. A. Fisher in the 1930's [Fisher, 1935]. Randomly assigning experimental subjects to "treatments" of different variables, however, is often expensive, difficult, impossible, unethical, or not meaningful. If the variable is the neighborhood where the subject lives, what does it mean to assign this at random? Even if it were feasible or ethical to consider reassigning subjects' neighborhood for the purposes of an experiment, perhaps what is relevant is not the value of the variable at the time of the experiment, but the lived experience of having been identified with and experienced that variable and its correlates for a long period of time.

We do not suggest that the interaction between online learning and strategic classification solves all problems relating to causality; far from it. The goal of this paper is simply to bring attention to a natural mechanism for exposing causal variables, that we believe is worthy of further attention.

2 Related Work

Much of the work on decision-making on individuals assumes that an individual's data is a fixed input that is independent of the algorithm used by the decision-maker. In practice, however, individuals may try to adapt to the model in place in order to improve their outcomes. A recent line of work studies such strategic behavior in classification settings.

Part of this line of work concerns itself with the negative consequences of strategic behavior, when individuals aim to game the model in place; for example, individuals may misrepresent their data or features (often at a cost) in an effort to obtain positive qualification outcomes or otherwise manipulate an algorithm's output [Dalvi et al., 2004, Perote and Perote-Pena, 2004, Dekel et al., 2010, Brückner et al., 2012, Ioannidis and Loiseau, 2013, Horel et al., 2014, Cai et al., 2015, Hardt et al., 2016, Dong et al., 2018] or even to protect their privacy [Ghosh et al., 2014, Cummings et al., 2015]. The goal in these works is to provide algorithms whose outputs are robust to such gaming. Milli et al. [2019] and Hu et al. [2019] focus on the social impact of robust classification, and show that i) robust classifiers come at a social cost (by forcing even qualified individuals to invest in costly feature manipulations to be classified positively) and ii) disparate abilities to game the model inevitably lead to unfair outcomes.

Another part of this line of work instead sees strategic manipulation as possibly positive, when the classifier incentivizes individuals to invest in meaningfully improving their features. Instead of cheating on a test to obtain a better score, a student may decide to study and actually improve his actual competence level in a given subject. Kleinberg and Raghavan [2019] study how to induce agents to invest effort into improving meaningful features rather than trying to game the classifier. Ustun et al. [2019] provide optimization tools to compute which action an agent should take to improve his label at minimal cost. Most of this line of work assumes the decision-maker already understands which features are impactful and control an agent's true label or qualification level, and which do not.

In contrast, we consider a setting where the decision-maker does not initially know which features affect an agent's label, and we aim to leverage the agents' strategic behavior to learn the causal relationship between features and labels; in that sense, our work is related to a line of research on causality [Pearl, 2009, Halpern, 2016, Peters et al., 2017]. Most closely related to this paper is the work of Miller et al. [2019]. They formalize the distinction between gaming and actual improvements through the structural causality framework of Pearl [2009], by introducing causal graphs that model the effect of their features and target variables on each other. They show that in such settings, it is in the decision-maker's best interest to incentivize actual improvements rather than gaming. Further, they show that designing good incentives that push agents to improve is at least as hard as causal inference, but leave open the question of how to leverage strategic behavior to learn causality, and hence set good incentives. Our paper provides a first step towards addressing this question, albeit in a simpler model.

3 Model

We consider a linear regression setting where the learner learns the regression parameters based on strategically manipulated data from a sequence of agents over rounds. There is a true latent regression parameter $\beta^* \in [-1, 1]^d$ such that for any agent with feature vector $x \in [-1, 1]^d$, the real-valued label y is given by

$$y = \beta^{*\top} x + \varepsilon,$$

where ε is a noise random variable with $|\varepsilon| \leq \sigma$, and $\mathbb{E}[\varepsilon \mid x] = 0$. We also refer to an individual's features as variables. There is a distribution over the *unmodified* features x in $[-1,1]^d$; we let μ be the mean and Σ be the covariance matrix of this distribution; we note that the distribution of unmodified features may be degenerate, i.e. Σ may not be full-rank. Throughout the paper, we set $\mu = 0$.[2]

The agents and the learner interact in an online fashion. At time t, the learner first posts a regression estimate $\hat{\beta}^t \in \mathbb{R}^d$, then an agent (indexed by t) arrives with their unmodified feature vector x_t. Agent t modifies the feature x_t into $\bar{x}_t$ in response to $\hat{\beta}^t$, in order to improve their assigned score $\langle \hat{\beta}^t, \bar{x}_t \rangle$. Finally, the learner observes the agent's realized label after feature modification, given by $\bar{y}_t = \beta^{*\intercal} \bar{x}_t + \varepsilon_t$.

Causal and non-causal features. When an agent modifies a feature k, this may also affect the agent's true label. We divide the coordinates of any given feature vector x into *causal* and *non-causal*; causal features are features that inform and control an agent's label, while non-causal features are those that do not affect an agent's label. Formally, for any $k \in [d]$, feature k is causal if and only if $\beta^*(k) \neq 0$, and non-causal if and only if $\beta^*(k) = 0$. An agent t can modify his true label by modifying causal features.

Agents' responses. Agents modify their features so as to maximize their regression outcome; modifications are costly and agents are budgeted. We assume agent t incurs a linear cost[3]

$$c_t(\Delta_t) = \sum_{k=1}^{d} c_t(k) \, |\Delta_t(k)|$$

to change his features by Δ_t, and has a total budget of B_t to modify his features. $\left(\{c_t(k)\}_{k \in [d]}, B_t\right)$'s are drawn i.i.d. from a distribution $\mathcal{C}$ that is unknown to the analyst. We assume $\mathcal{C}$ has discrete support $\{\left(c^1, B^1\right), \ldots, \left(c^l, B^l\right)\}$, and we denote by π^i the probability that $(c_t, B_t) = \left(c^i, B^i\right)$. We assume $c^i(k) > 0,\ B^i > 0$ for all $i \in [l],\ k \in [d]$; that is, every agent can modify his features, but no feature can be modified for free.[4]

When facing regression parameters $\hat{\beta}$, agent t solves

$$\begin{aligned} M(\hat{\beta}, c_t, B_t) = \underset{\Delta_t}{\operatorname{argmax}} \quad & \hat{\beta}^\top (x_t + \Delta_t) \\ \text{s.t.} \quad & \sum_{k=1}^{d} c_t(k) \Delta_t(k) \leq B_t. \end{aligned}$$

[2]This can be done whenever the learner can estimate the mean feature vector, since the learner can then center the features. The learner could estimate the mean by using unlabeled historical data; for example, she could collect data during a period when the algorithm does not make any decision on the agents, thus they would have no incentive to modify their features.

[3]We make this assumption for simplicity. Our results extend to more general assumptions on the cost function. It suffices that our cost function does not induces modifications such that several features are modified in a perfectly correlated fashion. When several features are perfectly correlated, said features may become indistinguishable by the learner.

[4]Note that in our model, modifying a feature affects only that feature and the label, but does not affect the values of any other features. We leave exploration of more complex models of feature intervention to future work.

The solution of the above program does not depend on x_t, only on $\hat{\beta}$ and (c_t, B_t), and is given by

$$\Delta_t = \sum_{k=1}^{d} \frac{B_t}{c_t(k)} \cdot \text{sgn}(\hat{\beta}(k)) \mathbb{1} \left\{ k = \underset{j}{\text{argmax}} \frac{\left|\hat{\beta}(j)\right|}{c_t(j)} \right\},$$

up to tie-breaking; when several features maximize $\frac{|\hat{\beta}(j)|}{c_t(j)}$, the agent modifies a single one of these features. We denote by D_τ the set of features that have been modified by at least one agent $t \in [\tau]$.

Natural learner dynamics: least-squares regression. Our goal here is to identify simple, natural learning dynamics that expose causal variables. The dynamics we consider are formally given in Algorithm 1; it is possible that more sophisticated learning algorithms could yield better guarantees with respect to regret and recovery.

When the learner updates his regression parameters, say at time τ, she does so based on the agent data observed up until time τ. We model the learner as picking $\hat{\beta}$ from the set $LSE(\tau)$ of solutions to the least-square regression problem run on the agents' data up until time τ, formally defined as

$$LSE(\tau) = \underset{\beta}{\text{argmin}} \sum_{t=1}^{\tau} \left(\bar{x}_t^\top \beta - \bar{y}_t \right)^2.$$

We introduce notation that will be useful for regression analysis. We let $\bar{X}_\tau \in \mathbb{R}_{\tau \times d}$ be the matrix of (modified) observations up until time τ. Each row corresponds to an agent $t \in [\tau]$, and agent t's row is given by $\bar{x}_t^\top$. Similarly, let $\bar{Y}_\tau = (\bar{y}_t)_{t \in [\tau]}^\top \in \mathbb{R}^{\tau \times 1}$. We can rewrite, for any τ,

$$LSE(\tau) = \underset{\beta}{\text{argmin}} \left(\bar{X}_\tau \beta - \bar{Y}_\tau \right)^\top \left(\bar{X}_\tau \beta - \bar{Y}_\tau \right).$$

Agents are grouped in epochs. The time horizon T is divided into epochs of size n, where n is chosen by the learner. At the start of every epoch E, the learner updates the posted regression parameter vector as a function of the history of $\bar{x}_t, \bar{y}_t$ up until epoch E. We let $\tau(E) = En$ denote the last time step of epoch E. $D_{\tau(E)}$ denotes the set of features that have been modified by at least one agent by the end of epoch E.

Algorithm 1: Online Regression with Epoch-Based Strategic modification (Epoch size n)

Learner picks (any) initial $\hat{\beta}_0$.
for *every epoch* $E \in \mathbb{N}$ **do**
  for $t \in \{(E-1)n + 1, \ldots, En\}$ **do**
    Agent t reports $\bar{x}_t \in M(\hat{\beta}_{E-1}, c_t, B_t)$.
    Learner observes $\bar{y}_t = \beta^{*\top} \bar{x}_t + \varepsilon_t$.
  end
  Learner picks $\hat{\beta}_E \in LSE\left(\tau(E)\right)$.
end

Examples We first illustrate why unmodified observations are insufficient for *any* algorithm to distinguish causal from non-causal features. Consider a setting where non-causal features are convex combinations of the causal features in the underlying (unmodified) distribution. Absent additional information, a learner would be faced with degenerate sets of observations that have rank strictly less than d, which can make accurate recovery of causality impossible:

Example 3.1. *Suppose $d = 2$, $\beta^* = (1, 0)$. Suppose feature 1 is causal and feature 2 is non-causal and correlated with 1: the distribution of unmodified features is such that for any feature vector x, feature 2 is identical to feature 1 as $x(2) = x(1)$. Then, any regression parameter of the form $\beta(\alpha) = (\alpha, 1 - \alpha)$ for $\alpha \in \mathbb{R}$ assigns agents the same score as β^*. Indeed,*

$$\beta^{*\top} x = x(1) = \alpha x(1) + (1 - \alpha)x(2) = \beta(\alpha)^\top x.$$

In turn, in the absence of additional information other than the observed features and labels, β^* is indistinguishable from any $\beta(\alpha)$, many of which recover the causality structure poorly (e.g., consider any α bounded away from 1).

We next illustrate that strategic agent modifications may aid in recovery of causal features, but only for those features that individuals actually invest in changing:

Example 3.2. *Consider a setting where $d = 3$, feature 1 is causal, and features 2 and 3 are non-causal and correlated with feature 1 as follows: for any feature vector x, $x(2), x(3) = x(1)$. Let $\beta^* = (1, 0, 0)$. Consider a situation in which the labels are noiseless (i.e., $\varepsilon = 0$ almost surely). Suppose that agents only modify their causal feature by a (possibly random) amount $\Delta(1)$.*

Note that the difference (in absolute value) between the score obtained by applying a given regression parameter $\hat{\beta}$ and the score obtained by applying β^* to feature vector x is given by

$$\begin{aligned}\left|\hat{\beta}^\top x - \beta^{*\top} x\right| &= \left|\hat{\beta}(1)\left(x(1) + \Delta(1)\right) + \hat{\beta}(2)x(2) + \hat{\beta}(3)x(3) - x(1) - \Delta(1)\right| \\ &= \left|\left(\hat{\beta}(1) + \hat{\beta}(2) + \hat{\beta}(3) - 1\right)x(1) + \left(\hat{\beta}(1) - 1\right)\Delta(1)\right|.\end{aligned}$$

In particular, for appropriate distributions of x and $\Delta(1)$, the predictions of $\hat{\beta}$ and β^* coincide if only if $\hat{\beta}(1) = 1$ and $\hat{\beta}(2) = -\hat{\beta}(3)$. As such, the learner learns after enough observations that necessarily, $\beta^*(1) = 1$. However, any regression parameter vector with $\hat{\beta}(1) = 1$, $\hat{\beta}(2) + \hat{\beta}(3) = 0$ is indistinguishable from β^*, and accurate recovery of $\beta^*(2)$ and $\beta^*(3)$ is impossible.

Note that even in the noiseless setting of Example 3.2, only the feature that has been modified can be recovered accurately. In more complex settings where the true labels are noisy, one should not hope to recover every feature well, but rather only those that have been modified sufficiently many times.

4 Recovery Guarantees for Modified Features

In this section, we focus on characterizing the recovery guarantees (with respect to the ℓ_2-norm) of Algorithm 1 at time $\tau(E) = En$ for *any* epoch E, with respect to the features that have been modified up until $\tau(E)$ (that is, in epochs 1 to E). We leave discussion of how the dynamics shape the set $D_{\tau(E)}$ of modified features to Section 5.

The main result of this section guarantees the accuracy of the $\hat{\beta}_E$ that the learning process converges to in its interaction with a sequence of strategic agents. The accuracy of the $\hat{\beta}_E$ that is recovered for a particular feature naturally depends on the number of epochs in which that feature is modified by the agents. For a feature that is never modified, we have no ability to distinguish correlation from causation. Recovery improves as the number of observations of the modified variable increases.

Formally, our recovery guarantee is given by the following theorem:

Theorem 4.1 (ℓ_2 Recovery Guarantee for Modified Features)**.** *Pick any epoch E. With probability at least $1-\delta$, for $n \geq \frac{\kappa d^2}{\lambda}\sqrt{\tau(E)\log(12d/\delta)}$,*

$$\sqrt{\sum_{k \in D_{\tau(E)}} \left(\hat{\beta}_E(k) - \beta^*(k)\right)^2} \leq \frac{K\sqrt{d\tau(E)\log(4d/\delta)}}{\lambda n},$$

where K, κ, λ are instance-specific constants that only depend on σ, $\mathcal{C}$, Σ, such that $\lambda > 0$.

When the epoch size is chosen so that $n = \Omega\left(\tau(E)^\alpha\right)$ for $\alpha > 1/2$, our recovery guarantee improves as $\tau(E)$ becomes larger. When $n = \Theta(\tau(E))$, our accuracy bound becomes $O\left(\frac{1}{\sqrt{\tau(E)}}\right)$; this matches the well-known recovery guarantees of least square regression for a single batch of $\tau(E)$ i.i.d observations drawn from a non-degenerate distribution of features. When the epoch size n is sub-linear in $\tau(E)$, the accuracy guarantee degrades to $O\left(\frac{\sqrt{\tau(E)}}{n}\right)$. This is because some features are modified in few epochs,[5] that is, $\Theta(n)$ times, and the number of times such features are modified drives how accurately they can be recovered.

We provide a proof sketch below, and defer the full proof of Theorem 4.1 to Appendix A.

Proof sketch for Theorem 4.1. We focus on the subspace $\mathcal{V}_{\tau(E)}$ of $\mathbb{R}^d$ spanned by the observed features $\bar{x}_1, \ldots, \bar{x}_{\tau(E)}$, and for any $z \in \mathbb{R}^d$, we denote by $z(\mathcal{V}_{\tau(E)})$ the projection of of z onto $\mathcal{V}_{\tau(E)}$.

First, we show via concentration that in this subspace, the mean-square error is strongly convex, with parameter $\Theta(n)$ (see Claim A.6). This strong convexity parameter is controlled by the smallest eigenvalue of $\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)}$ over subspace $\mathcal{V}_{\tau(E)}$. Formally, we lower bound this eigenvalue and show that with probability at least $1-\delta/2$, for n large enough,

$$\left(\hat{\beta}_E(\mathcal{V}_{\tau(E)}) - \beta^*(\mathcal{V}_{\tau(E)})\right)^\top \bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)} \left(\hat{\beta}_E(\mathcal{V}_{\tau(E)}) - \beta^*(\mathcal{V}_{\tau(E)})\right) \geq \frac{\lambda n}{4}. \tag{1}$$

Second, we bound the effect of the noise ε on the mean-squared error by $O(\sqrt{\tau(E)})$ in Lemma A.3, once again via concentration. Formally, we abuse notation and let $\varepsilon_{\tau(E)} \triangleq (\varepsilon_t)_{t \in [\tau(E)]}^\top$, and show that with probability at least $1-\delta/2$,

$$\left(\hat{\beta}_E(\mathcal{V}_{\tau(E)}) - \beta^*(\mathcal{V}_{\tau(E)})\right)^\top \bar{X}_{\tau(E)}^\top \varepsilon_{\tau(E)} \leq \left\|\hat{\beta}_E(\mathcal{V}_{\tau(E)}) - \beta^*(\mathcal{V}_{\tau(E)})\right\|_2 \cdot K\sqrt{d\tau(E)\log(4d/\delta)}. \tag{2}$$

[5] In particular, as we will see, we expect correlated, non-causal features to only be modified in a small number of epochs: once a non-causal feature k has been modified in a few epochs, it is accurately recovered. In further periods E, the learner sets $\hat{\beta}_E(k)$ close to 0. This disincentivizes further modifications of feature k.

Finally, we obtain the result via Lemma A.2, that shows the distance between $\hat{\beta}_E$ and β^* (restricted to $\mathcal{V}_{\tau(E)}$) decreases inversely proportionally to the magnitude of the strong convexity parameter, and increases proportionally to the noise in the mean-squared error. Formally, Lemma A.2 states that taking the first-order conditions on the mean-squared error yields

$$\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)} \left(\hat{\beta}_E(\mathcal{V}_{\tau(E)}) - \beta^*(\mathcal{V}_{\tau(E)}) \right) = \bar{X}_{\tau(E)}^\top \varepsilon_{\tau(E)},$$

which can be combined with Equations (1) and (2) to show our bound with respect to sub-space $\mathcal{V}_{\tau(E)}$. In turn, as the set of features $D_{\tau(E)}$ modified up until time $\tau(E)$ defines a sub-space of $\mathcal{V}_{\tau(E)}$, our accuracy bound applies to $D_{\tau(E)}$. $\square$

Remark 4.2. *We remark that Theorem 4.1 is not a direct consequence of the classical recovery guarantees of least-square regression. Such recovery guarantees leverage strong convexity of the mean-squared error in β; this error is strongly convex if and only if $\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)}$ has rank d, or equivalently the observations $\bar{x}_1, \ldots, \bar{x}_{\tau(E)}$ span $\mathbb{R}^d$. In contrast, our statement can deal with degenerate distributions over modified features, inducing observations that only span a strict sub-space of $\mathbb{R}^d$. Such distributions can arise in our setting, as evidenced by Examples 3.1 and 3.2.*

5 Ensuring Exploration via Least Squares Tie-Breaking

In this section, we focus on ensuring that the interaction between the online learning process and the strategic modification results in modification of a diverse set of variables over time.

Recall we are solving the following least-square problem at time $\tau(E)$, for all epochs E:

$$LSE(\tau(E)) = \operatorname*{argmin}_{\beta} \sum_{t=1}^{\tau(E)} \left(\bar{x}_t^\top \beta - \bar{y}_t \right)^2.$$

An equivalent characterization of $LSE(\tau(E))$ is the set of solutions to the following linear system of equations:

$$\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)} \beta = \bar{X}_{\tau(E)}^\top \bar{Y}_{\tau(E)}. \tag{3}$$

When $\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)}$ is invertible, this has a single solution, given by

$$\hat{\beta}_E = \left(\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)} \right)^{-1} \bar{X}_{\tau(E)}^\top \bar{Y}_{\tau(E)}.$$

However, in our setting, it may be the case that $\bar{X}_{\tau(E)}^\top \bar{X}_{\tau(E)}$ is rank-deficient (see Examples 3.1, 3.2). In this case, the system of (linear) equations (3) is under-determined and admits a continuum of solutions. This gives rise to the question of which least-square solutions are preferable in our setting, and how to break ties between several solutions.

The learner's choice of regression parameters in each epoch affects the distribution of feature modifications in subsequent epochs. As the recovery guarantee of Theorem 4.1 only applies to features that have been modified, we would like our tie-breaking rule to regularly incentivize agents to modify new features. We first show that a natural, commonly used tie-breaking rule—picking the minimum norm solution to the least-square problem—may fail to do so:

Example 5.1. *Consider a setting with $d = 2$, $\beta^* = (1, 2)$ and noiseless labels, i.e., $\varepsilon_t = 0$ always. Suppose that with probability 1, every agent t has features $x_t = (0, 0)$, budget $B_t = 1$, and costs $c_t(1) = c_t(2) = 1$ to modify each feature. We let the tie-breaking pick the solution with the least $\ell 2$ norm among all solutions to the least-square problem.*

Pick any initial regression parameter $\hat{\beta}_0$ with $\hat{\beta}_0(1) > \hat{\beta}_0(2)$. For every agent t in epoch 1, t picks modification vector $\Delta_t = (1, 0)$. This induces observations $\bar{x}_t = (1, 0)$, $\bar{y}_t = 1$. The set of least-square solutions (with error exactly 0) in epoch 1 is then given by $\{(1, \beta_2) : \forall \beta_2 \in \mathbb{R}\}$, and the minimum-norm solution chosen at the end of epoch 1 is $\hat{\beta}_1 = (1, 0)$. This solution incentivizes agents to set $\Delta_t = (1, 0)$, and Algorithm 1 gets stuck in a loop where every agent t reports $\bar{x}_t = (1, 0)$, and the algorithm posts regression parameter vector $\hat{\beta}_E = (1, 0)$ in response, in every epoch E. The second feature is never modified by any agent, and is not recovered accurately.

Example 5.1 highlights that a wrong choice of tie-breaking rule can lead Algorithm 1 to explore the same features over and over again. In response, we propose the following tie-breaking rule, described in Algorithm 2:

Algorithm 2: Tie-Breaking Scheme at Time $\tau(E)$.

Input: Epoch E, observations $(\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_{\tau(E)}, \bar{y}_{\tau(E)})$, parameter α
Let $\mathcal{U}_{\tau(E)} = \text{span}\left(\bar{x}_1, \ldots, \bar{x}_{\tau(E)}\right)$.
if $rank\left(\mathcal{U}_{\tau(E)}\right) < d$ **then**
    Find an orthonormal basis $B^{\perp}_{\tau(E)}$ for $\mathcal{U}^{\perp}_{\tau(E)}$.
    Set $v = \sum_{b \in B^{\perp}_{\tau(E)}} b \neq 0$, renormalize $v := \frac{v}{\|v\|_2}$.
    Pick β_E a vector in $LSE(\tau(E))$ with minimal norm.
    Set $\hat{\beta}_E = \beta_E + \alpha v$.
else
    Set $\hat{\beta}_E$ be the unique element in $LSE(\tau(E))$.
end
Output: $\hat{\beta}_E$.

Intuitively, at the end of epoch E, our tie-breaking rule picks a solution in $LSE(\tau(E))$ with large norm. This ensures the existence of a feature $k \notin D_{\tau(E)}$ that has not yet been modified up until time $\tau(E)$, and that is assigned a large weight by our least-square solution. In turn, this feature is more likely to be modified in future epochs.

Our main result in this section shows that the tie-breaking rule of Algorithm 2 eventually incentivizes the agents to modify all d features, allowing for accurate recovery of β^* in its entirety.

Theorem 5.2 (Recovery Guarantee with Tie-Breaking Scheme (Algorithm 2))**.** *Suppose the epoch size satisfies $n \geq \frac{\kappa d^2}{\lambda} \sqrt{2T \log(24d/\delta)}$, and take α to be*

$$\alpha \geq \gamma \left(\sqrt{d} + \frac{Kd\sqrt{2T \log(8d/\delta)}}{\lambda n} \right),$$

where γ, K, κ, λ are instance-specific constants that only depend on σ, $\mathcal{C}$, Σ, and $\lambda > 0$. If $T \geq dn$, we have with probability at least $1 - \delta$ that at the end of the last epoch T/n,

$$\left\| \hat{\beta}_{T/n} - \beta^* \right\|_2 \leq \frac{K\sqrt{2dT \log(8d/\delta)}}{\lambda n},$$

under the tie-breaking rule of Algorithm 2 .

Remark 5.3. *The bound in Theorem 5.2 provides guidance for selecting the epoch length, so as to ensure optimal recovery guarantees. Under the natural assumption that $T >> d$, the optimal recovery rate is achieved when roughly $n = \Theta(\frac{T}{d})$. This results in an $O\left(\frac{d\sqrt{d \log d}}{\sqrt{T}}\right)$ upper bound on the ℓ_2 distance between the recovered regression parameters and β^*.*

We provide a proof sketch below. The full proof is given in Appendix B.

Proof sketch of Theorem 5.2. For α arbitrarily large, the norm of $\hat{\beta}$ becomes arbitrarily large. Because at the end of epoch E, $\hat{\beta}_E$ guarantees accurate recovery of all features modified up until time En, it must be that $\hat{\beta}_E(k)$ is arbitrarily large for some feature k that has not yet been modified. In turn, this feature is modified in epoch $E+1$. After d epochs, and in particular for $T \geq dn$, this leads to $D_T = [d]$. The recovery guarantee of Theorem 4.1 then applies to all features. □

6 Conclusions and Future Directions

This paper provides evidence that interaction between an online learner and individuals who strategically modify their features can result in discovery of causal features, also incentivizing individuals to invest in these features, rather than gaming. In future work, it would be natural to explore this interaction in richer and more complex settings.

7 Acknowledgments

The work of Yahav Bechavod and Katrina Ligett was supported in part by Israel Science Foundation (ISF) grant #1044/16, the United States Air Force and DARPA under contracts FA8750-16-C-0022 and FA8750-19-2-0222, and the Federmann Cyber Security Center in conjunction with the Israel national cyber directorate. Zhiwei Steven Wu was supported in part by the NSF FAI Award #1939606, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant. Juba Ziani was supported in part by the Inaugural PIMCO Graduate Fellowship at Caltech and the National Science Foundation through grant CNS-1518941. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA. We thank Mohammad Fereydounian and Aaron Roth for useful discussions.

References

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. In *COLT*, 2015.

Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In *COLT*, 2015.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759 – 777, 2010.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Frederick Eberhardt. Causation and intervention. *Doctoral dissertation, Carnegie Mellon University*, page 93, 2007.

R.A. Fisher. *The design of experiments. 1935.* Oliver and Boyd, 1935.

Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. Buying private data without verification. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 931–948, 2014.

Joseph Y Halpern. *Actual causality.* MIT Press, 2016.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Thibaut Horel, Stratis Ioannidis, and S. Muthukrishnan. Budget feasible mechanisms for experimental design. In *LATIN 2014: Theoretical Informatics*, Lecture Notes in Computer Science, pages 719–730, 2014.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Stratis Ioannidis and Patrick Loiseau. Linear regression as a non-cooperative game. In *Web and Internet Economics*, pages 277–290, 2013.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the FAT**, pages 825–844, 2019.

John Miller, Smitha Milli, and Moritz Hardt. Strategic adaptation to classifiers: A causal perspective. *arXiv preprint arXiv:1910.10362*, 2019.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Judea Pearl. *Causality.* Cambridge University Press, 2009.

Javier Perote and Juan Perote-Pena. Strategy-proof estimators for simple regression. In *Mathematical Social Sciences 47*, pages 153–176, 2004.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT press, 2017.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

A Proof of Theorem 4.1

A.1 Preliminaries

A.1.1 Useful concentration

Our proof will require applying the following concentration inequality, derived from Azuma's inequality:

Lemma A.1. *Let $W_1, \ldots, W_\tau$ be random variables in $\mathbb{R}$ such that $|W_t| \leq W_{max}$. Suppose for all $t \in [\tau]$, for all $w_1, \ldots, w_{t-1}$,*

$$\mathbb{E}\left[W_t | W_{t-1} = w_{t-1}, \ldots, W_1 = w_1\right] = 0.$$

Then, with at least $1 - \delta$,

$$\left|\sum_{t=1}^{\tau} W_t\right| \leq W_{max}\sqrt{2\tau \log(2/\delta)}.$$

Proof. This is a reformulated version of Azuma's inequality. To see this, define

$$Z_t = \sum_{i=1}^{t} W_i \quad \forall t,$$

and initialize $Z_0 = 0$. We start by noting that for all $t \in [\tau]$, since

$$Z_t = \sum_{i=1}^{t} W_i = W_t + \sum_{i=1}^{t-1} W_i = W_t + Z_{t-1},$$

we have

$$\begin{aligned}\mathbb{E}\left[Z_t | Z_{t-1}, \ldots, Z_1\right] &= \mathbb{E}\left[W_t | Z_{t-1}, \ldots, Z_1\right] + \mathbb{E}\left[Z_{t-1} | Z_{t-1}, \ldots, Z_1\right] \\ &= \mathbb{E}\left[W_t | Z_{t-1}, \ldots, Z_1\right] + Z_{t-1}.\end{aligned}$$

Further, it is easy to see that $Z_i = z_i \; \forall i \in [t-1]$ if and only if $W_i = z_i - z_{i-1} \; \forall i \in [t-1]$, hence

$$\mathbb{E}\left[W_t | Z_{t-1} = z_{t-1}, \ldots, Z_1 = z_1\right] = \mathbb{E}\left[W_t | W_i = z_i - z_{i-1} \; \forall i \in [t-1]\right] = 0.$$

Combining the last two equations implies that

$$\mathbb{E}\left[Z_t | Z_{t-1}, \ldots, Z_1\right] = Z_{t-1},$$

and the Z_t's define a martingale. Since for all t,

$$|Z_t - Z_{t-1}| = |W_t| \leq W_{max},$$

we can apply Azuma's inequality to show that with probability at least $1 - \delta$,

$$|Z_\tau - Z_0| \geq W_{max}\sqrt{2\tau \log(2/\delta)},$$

which immediately gives the result. ∎

A.1.2 Sub-space decomposition and projection

We will also need to divide $\mathbb{R}^d$ in several sub-spaces, and project our observations to said subspaces.

Sub-space decomposition We focus on the sub-space generated by the non-modified features x_t's and the sub-space generated by the feature modifications Δ_t's. We let r be the rank of Σ, and let $\lambda_r \geq \ldots \geq \lambda_1 > 0$ be the non-zero eigenvalues of Σ. Further, we let $f_1, \ldots, f_r$ be the unit eigenvectors (i.e., such that $\|f_1\|_1 = \ldots = \|f_r\|_1 = 1$) corresponding to eigenvalues $\lambda_1, \ldots, \lambda_r$ of Σ. As Σ is a symmetric matrix, $f_1, \ldots, f_r$ are orthonormal. We abuse notations in the proof of Theorem 4.1 and denote $\Sigma = \text{span}(f_1, \ldots, f_r)$ when clear from context.

For all k, let e_k be the unit vector such that $e_k(k) = 1$ and $e_k(j) = 0 \ \forall j \neq k$. At time τ, we denote $\mathcal{D}_\tau = \text{span}\left(e_k\right)_{k \in D_\tau}$ the sub-space of $\mathbb{R}^d$ spanned by the features in D_τ.

Finally, we let

$$\mathcal{V}_\tau = \Sigma + \mathcal{D}_\tau = \text{span}\left(f_1, \ldots, f_r\right) + \text{span}\left(e_k\right)_{k \in D_\tau}$$

be the Minkowski sum of sub-spaces Σ and $\mathcal{D}_\tau$.

Projection onto sub-spaces For any vector z, sub-space $\mathcal{H}$ of $\mathbb{R}^d$, we write $z = z(\mathcal{H}) + z(\mathcal{H}^\perp)$ where $z(\mathcal{H})$ is the projection of z onto sub-space $\mathcal{H}$, i.e. is uniquely defined as

$$z(\mathcal{H}) = \sum_{q \in B} (z^\top q) q$$

for any orthonormal basis B of $\mathcal{H}$. We also let $z(\mathcal{H}^\perp)$ be the projection on the orthogonal complement $\mathcal{H}^\perp$. In particular, $z(\mathcal{H})$ is orthogonal to $z(\mathcal{H}^\perp)$. Further, we write $\bar{X}_\tau(\mathcal{H})$ the matrix whose rows are given by $\bar{x}_t(\mathcal{H})^\top$ for all $t \in [\tau]$.

A.2 Main Proof

Characterization of the least-square estimate via first-order conditions First, for any least square solution $\hat{\beta}_E$ at time $\tau(E)$, we write the first order conditions solved by $\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)$, the projection of $\hat{\beta}_E$ on sub-space $\mathcal{V}_{\tau(E)}$. We abuse notations to let $\varepsilon_{\tau(E)} \triangleq \left(\varepsilon_t\right)_{t \in [\tau(E)]}$ the vector of all ε_t's up until time $\tau(E)$, and state the result as follows:

Lemma A.2 (First-order conditions projected onto $\mathcal{V}_{\tau(E)}$)**.** *Suppose $\hat{\beta}_E \in LSE(\tau(E))$. Then,*

$$\left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) = \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}.$$

Proof. For simplicity of notations, we drop all $\tau(E)$ indices and subscripts in this proof. Remember that

$$LSE = \operatorname*{argmin}_\beta \left(\bar{X}\beta - \bar{Y}\right)^\top \left(\bar{X}\beta - \bar{Y}\right).$$

Since $\hat{\beta}_E \in LSE$, it must satisfy the first order conditions given by

$$2\bar{X}^\top \left(\bar{X}\hat{\beta}_E - \bar{Y}\right) = 0,$$

which can be rewritten as

$$\bar{X}^\top \bar{X}\hat{\beta}_E = \bar{X}^\top \bar{Y}.$$

Second, we note that for all t, $x_t \in \mathrm{span}(f_1, \ldots, f_r)$ and $\Delta_t \in \mathrm{span}\left((e_k)_{k \in D}\right)$ (by definition of D). This immediately implies, in particular, that $\bar{x}_t = x_t + \Delta_t \in \mathcal{V}$. In turn, $\bar{x}_t\left(\mathcal{V}\right) = \bar{x}_t$ for all t, and

$$\bar{X} = \bar{X}\left(\mathcal{V}\right).$$

As such, the first order condition can be written

$$\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right) \hat{\beta}_E = \bar{X}\left(\mathcal{V}\right)^\top \bar{Y}.$$

Now, we remark that

$$\begin{aligned}
\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right) \hat{\beta}_E &= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right) \bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E \\
&= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right) \bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E\left(\mathcal{V}\right) + \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right) \bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E(\mathcal{V}^\perp) \\
&= \sum_{t \in S} \bar{x}_t\left(\mathcal{V}\right) \bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E\left(\mathcal{V}\right) \\
&= \bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right) \hat{\beta}_E\left(\mathcal{V}\right),
\end{aligned}$$

where the second-to-last equality follows from the fact that $\mathcal{V}$ and $\mathcal{V}^\perp$ are orthogonal, which immediately implies $\bar{x}_t\left(\mathcal{V}\right)^\top \hat{\beta}_E(\mathcal{V}^\perp) = 0$ for all t. To conclude the proof, we note that $\bar{Y} = \bar{X}^\top \beta^* + \varepsilon = \bar{X}\left(\mathcal{V}\right)^\top \beta^*\left(\mathcal{V}\right) + \varepsilon$. Plugging this in the above equation, we obtain that

$$\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right) \hat{\beta}_E\left(\mathcal{V}\right) = \bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)^\top \beta^*\left(\mathcal{V}\right) + \bar{X}\left(\mathcal{V}\right)^\top \varepsilon.$$

This can be rewritten

$$\left(\bar{X}\left(\mathcal{V}\right)^\top \bar{X}\left(\mathcal{V}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}\right) - \beta^*\left(\mathcal{V}\right)\right) = \bar{X}\left(\mathcal{V}\right)^\top \varepsilon,$$

which completes the proof. ∎

Upper-bounding the right-hand side of the first order conditions We now use concentration to give an upper bound on a function of the right-hand side of the first order conditions,

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}.$$

Lemma A.3. *With probability at least $1 - \delta$,*

$$\begin{aligned}
&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon \\
&\le \left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \cdot K' \sqrt{d\tau(E)\log(2d/\delta)}.
\end{aligned}$$

where K' is a constant that only depends on the distribution of costs and the bound σ on the noise.

Proof. Pick any $k \in [d]$, and define $W_t = \bar{x}_t(k)\varepsilon_t$. First, we remark that

$$|\bar{x}_t(k)| \le |x_t(k)| + |\Delta_t(k)| \le 1 + \max_{k \in [d],\ i \in [l]} \frac{B^i}{c^i(k)}.$$

In turn, $|W_t| \leq K'$ where

$$K' \triangleq \left(1 + \max_{k\in[d],\ i\in[l]} \frac{B^i}{c^i(k)}\right)\sigma.$$

Further, note that both $x_t(k)$ and ε_t are independent of the history of play up through time $t-1$, hence of $W_1, \ldots, W_{t-1}$, and that ε_t is further independent of Δ_t (the distribution of Δ_t is a function of the currently posted $\hat{\beta}_{E-1}$ only, which only depends on the previous time steps). Noting that if A, B, C are random variables, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{A,B}[AB|C=c] &= \sum_a \sum_b ab \Pr[A=a, B=b|C=c] \\
&= \sum_a \sum_b ab \Pr[A=a|B=b, C=c] \Pr[B=b|C=c] \\
&= \sum_b b \left(\sum_a a \Pr[A=a|B=b, C=c]\right) \Pr[B=b|C=c] \\
&= \sum_b b \mathop{\mathbb{E}}_A[A|B=b, C=c] \Pr[B=b|C=c] \\
&= \mathop{\mathbb{E}}_B\left[\mathop{\mathbb{E}}_A[A|B, C=c]\, B|C=c\right],
\end{aligned}$$

and applying this with $A = \varepsilon_t$, $B = \Delta_t(k)$, $C = W_1 \cap \ldots \cap W_{t-1}$, we obtain

$$\begin{aligned}
\mathbb{E}[W_t|W_{t-1}, \ldots, W_1] &= \mathbb{E}[\bar{x}_t(k)\varepsilon_t|W_{t-1}, \ldots, W_1] \\
&= \mathbb{E}[x_t(k)\varepsilon_t|W_{t-1}, \ldots, W_1] + \mathbb{E}[\Delta_t(k)\varepsilon_t|W_{t-1}, \ldots, W_1] \\
&= \mathbb{E}[x_t(k)\varepsilon_t] + \mathop{\mathbb{E}}_{\Delta_t}\left[\mathop{\mathbb{E}}_{\varepsilon_t}[\varepsilon_t|\Delta_t(k), W_{t-1}, \ldots, W_1] \cdot \Delta_t(k) \middle| W_{t-1}, \ldots, W_1\right] \\
&= \mathop{\mathbb{E}}_{x_t}\left[x_t(k) \cdot \mathop{\mathbb{E}}_{\varepsilon}[\varepsilon_t|x_t(k)]\right] + \mathop{\mathbb{E}}_{\Delta_t}\left[\Delta_t(k) \cdot \mathop{\mathbb{E}}_{\varepsilon_t}[\varepsilon_t] \middle| W_{t-1}, \ldots, W_1\right] \\
&= 0,
\end{aligned}$$

since $\mathbb{E}_{\varepsilon_t}[\varepsilon_t] = 0$ and $\mathbb{E}_{\varepsilon}[\varepsilon_t|x_t(k)] = 0$. Hence, we can apply Lemma A.1 and a union bound over all d features to show that with probability at least $1-\delta$,

$$\sum_{t=1}^{\tau(E)} \bar{x}_t(k)\varepsilon_t \geq -K'\sqrt{2\tau(E)\log(2d/\delta)} \quad \forall k \in [d].$$

By Cauchy-Schwarz, we have

$$\begin{aligned}
\left(\hat{\beta}_E(\mathcal{V}) - \beta^*(\mathcal{V})\right)^\top \sum_{t=1}^{\tau(E)} \bar{x}_t\varepsilon_t &\leq \left\|\hat{\beta}_E(\mathcal{V}) - \beta^*(\mathcal{V})\right\|_2 \cdot \left\|\sum_{t=1}^{\tau(E)} \bar{x}_t\varepsilon_t\right\|_2 \\
&\leq \left\|\hat{\beta}_E(\mathcal{V}) - \beta^*(\mathcal{V})\right\|_2 \sqrt{\sum_{k=1}^{d}\left(\sum_t \bar{x}_t(k)\varepsilon_t\right)^2} \\
&\leq \left\|\hat{\beta}_E(\mathcal{V}) - \beta^*(\mathcal{V})\right\|_2 \cdot K'\sqrt{2d\tau(E)\log(2d/\delta)}.
\end{aligned}$$

□

Strong convexity of the mean-squared error in sub-space $\mathcal{V}(\tau(E))$ We give a lower bound on the eigenvalues of $\bar{X}^\top \bar{X}$ on sub-space $\mathcal{V}(\tau(E))$, so as to show that at time $\tau(E)$, any least square solution $\hat{\beta}_E$ satisfies

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) \geq \Omega(n)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2.$$

To do so, we will need the following concentration inequalities:

Lemma A.4. *Suppose $\mathbb{E}[x_t] = 0$. Fix $\tau(E) = En$ for some $E \in \mathbb{N}$. With probability at least $1-\delta$, we have that*

$$\sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z \geq \left(\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(6r/\delta)}\right)\|z\|_2^2 \quad \forall z \in \Sigma,$$

and

$$\sum_{t=1}^{\tau(E)} z^\top \Delta_t \Delta_t^\top z \geq \left(\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} n - \left(\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\right)^2 \sqrt{2n\log(6d/\delta)}\right)\|z\|_2^2 \quad \forall z \in \mathcal{D}_{\tau(E)}$$

and

$$\sum_{t=1}^{\tau(E)} z^\top x_t \Delta_t^\top z \geq -2\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\} d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2 \quad \forall z \in \mathbb{R}^d.$$

Proof. Deferred to Appendix A.2.1. □

We will also need the following statement on the norm of the projections of any $z \in \mathcal{V}$ to $\mathcal{D}$ and Σ:

Lemma A.5. *Let*

$$\lambda(\mathcal{D}, \Sigma) = \inf_{z \in \mathcal{D}+\Sigma} \|z(\mathcal{D})\|_2 + \|z(\Sigma)\|_2 \quad s.t. \quad \|z\|_2 = 1.$$

Then, $\lambda(\mathcal{D}, \Sigma) > 0$.

Proof. With respect to the Euclidean metric, the objective function is continuous in z (the orthogonal projection operators are linear hence continuous functions of z and $z \to \|z\|_2$ also is a continuous function), and its feasible set is compact (as it is a sphere in a bounded-dimensional space over real values). By the extreme value theorem, the optimization problem admits an optimal solution, i.e., there exists z^* with $\|z^*\|_2 = 1$ such that $\lambda(\mathcal{D}, \Sigma) = \|z^*(\mathcal{D})\|_2 + \|z^*(\Sigma)\|_2$. Now, supposing $\lambda(\mathcal{D}, \Sigma) \leq 0$, it must necessarily be the case that $z(\mathcal{D}) = 0$, $z(\Sigma) = 0$. In particular, this means z is orthogonal to both $\mathcal{D}$ and Σ. In turn, z must be orthogonal to every vector in $\mathcal{D} + \Sigma$; since $z \in \mathcal{D} + \Sigma$, this is only possible when $z = 0$, contradicting $\|z\|_2 = 1$. □

We can now move onto the proof of our lower bound for

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right).$$

Corollary A.6. *Fix $\tau(E)=En$ for some $E\in\mathbb{N}$. With probability at least $1-\delta$,*

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)\\ &\geq\left(\frac{\lambda n}{2}-\kappa' d^2\sqrt{\tau(E)\log(6d/\delta)}\right)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2,\end{aligned}$$

for some constants κ', λ that only depend on σ, $\mathcal{C}$, and Σ, with $\lambda>0$.

Proof. Since it is clear from context, we drop all $\tau(E)$ subscripts in the notation of this proof. First, we remark that

$$\begin{aligned}z^\top\bar{X}^\top\bar{X}z&=\sum_t z^\top\bar{x}_t\bar{x}_t^\top z\\&=\sum_t z^\top x_t x_t^\top z+\sum_t z^\top\Delta_t\Delta_t^\top z+2\sum_t z^\top\Delta_t z^\top x_t.\end{aligned}$$

We have by Lemma A.5 that for all $z\in\mathcal{V}=\mathcal{D}+\Sigma$,

$$\|z(\mathcal{D})\|_2+\|z(\Sigma)\|_2\geq\lambda(\mathcal{D},\Sigma)\|z\|_2.$$

Let $\lambda(\Sigma)\triangleq\min_{D\subset[d]}\lambda(\mathcal{D},\Sigma)$. Since there are finitely many subsets D of $[d]$ (and corresponding sub-spaces $\mathcal{D}$) and since for all such subsets, $\lambda(\mathcal{D},\Sigma)>0$, we have that $\lambda(\Sigma)>0$. Further,

$$\|z(\mathcal{D})\|_2+\|z(\Sigma)\|_2\geq\lambda(\Sigma)\|z\|_2.$$

Therefore, it must be the case that either $\|z(\mathcal{D})\|_2\geq\frac{\lambda(\Sigma)}{2}\|z\|_2$ or $\|z(\Sigma)\|_2\geq\frac{\lambda(\Sigma)}{2}\|z\|_2$. We divide our proof into the corresponding two cases:

1. The first case is when $\|z(\Sigma)\|_2\geq\frac{\lambda(\Sigma)}{2}\|z\|_2$. Then, note that since $z^\top\Delta_t\Delta_t^\top z\geq 0$ always, we have

$$\begin{aligned}\sum_t z^\top\bar{x}_t\bar{x}_t^\top z&\geq\sum_t z^\top x_t x_t^\top z+2\sum_t z^\top\Delta_t z^\top x_t\\&=\sum_t z(\Sigma)^\top x_t x_t^\top z(\Sigma)+2\sum_t z^\top\Delta_t z^\top x_t,\end{aligned}$$

where the last equality follows from the fact that $x_t\in\Sigma$ and $z=z(\Sigma)+z(\Sigma^\perp)$. By Lemma A.4, we get that for some constant C_1 that depends only on $\mathcal{C}$,

$$\begin{aligned}&\sum_t z^\top\bar{x}_t\bar{x}_t^\top z\\&\geq\left(\lambda_r\tau(E)-2rd\sqrt{\tau(E)\log(6r/\delta)}\right)\|z(\Sigma)\|_2^2-C_1 d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2\\&\geq\left(\frac{\lambda(\Sigma)\lambda_r}{2}\tau(E)-\lambda(\Sigma)rd\sqrt{\tau(E)\log(6r/\delta)}-C_1 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2\\&\geq\left(\frac{\lambda(\Sigma)\lambda_r}{2}\tau(E)-\lambda(\Sigma)d^2\sqrt{\tau(E)\log(6d/\delta)}-C_1 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2.\end{aligned}$$

(The second step assumes $\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(6r/\delta)} \geq 0$. When this is negative, the bound trivially holds as $\sum_t z^\top \bar{x}_t \bar{x}_t^\top z \geq 0$.)

2. The second case arises when $\|z(\mathcal{D})\|_2 \geq \frac{\lambda(\Sigma)}{2}\|z\|_2$. Note that

$$\begin{aligned}\sum_t z^\top \bar{x}_t \bar{x}_t^\top z &\geq \sum_t z^\top \Delta_t \Delta_t^\top z + 2\sum_t z^\top \Delta_t z^\top x_t \\ &= \sum_t z(\mathcal{D})^\top \Delta_t \Delta_t^\top z(\mathcal{D}) + 2\sum_t z^\top \Delta_t z^\top x_t,\end{aligned}$$

as $\Delta_t \in \mathcal{D}$ and $z = z(\mathcal{D}) + z(\mathcal{D}^\perp)$. By Lemma A.4, it follows that for some constants $C_2,\ C_3$ that only depend on $\mathcal{C}$,

$$\begin{aligned}&\sum_t z^\top \bar{x}_t \bar{x}_t^\top z \\ &\geq \left(n \min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} - C_2\sqrt{n\log(6d/\delta)}\right)\|z(\mathcal{D})\|_2^2 - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\|z\|_2^2 \\ &\geq \left(\frac{\lambda(\Sigma)n}{2}\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} - \frac{\lambda(\Sigma)C_2}{2}\sqrt{n\log(6d/\delta)} - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2 \\ &\geq \left(\frac{\lambda(\Sigma)n}{2}\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} - \frac{\lambda(\Sigma)C_2}{2}\sqrt{\tau(E)\log(6d/\delta)} - C_3 d\sqrt{\tau(E)\log(6d/\delta)}\right)\|z\|_2^2.\end{aligned}$$

Noting that by definition $\lambda_r > 0$ and $\min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\} > 0$, and picking the worse of the two above bounds on $\sum_t z^\top \bar{x}_t \bar{x}_t^\top z$ concludes the proof with

$$\lambda = \frac{\lambda(\Sigma)}{2}\min\left(\lambda_r, \min_{i,k}\left\{\pi^i\left(\frac{B^i}{c^i(k)}\right)^2\right\}\right) > 0.$$

□

We can now prove Theorem 4.1. By Lemma A.2, we have that

$$\left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) = \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)},$$

which immediately yields

$$\begin{aligned}&\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \left(\bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right) \\ &= \left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon_{\tau(E)}\end{aligned}$$

by performing matrix multiplication with $\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right) - \beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top$ on both sides on the first-order conditions. Further, by Lemma A.3, Corollary A.6, and a union bound, we get that with probability at least $1-\delta$,

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)$$
$$\geq \left(\frac{\lambda n}{2}-\kappa' d^2\sqrt{\tau(E)\log(12d/\delta)}\right)\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2^2,$$

and

$$\left(\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right)^\top \bar{X}_{\tau(E)}\left(\mathcal{V}_{\tau(E)}\right)^\top \varepsilon$$
$$\leq \left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \cdot K'\sqrt{d\tau(E)\log(4d/\delta)}.$$

Combining the two above inequalities with the first-order conditions yields

$$\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \leq \frac{K'\sqrt{d\tau(E)\log(4d/\delta)}}{\frac{\lambda n}{2}-\kappa' d^2\sqrt{\tau(E)\log(12d/\delta)}}.$$

For

$$n \geq \frac{4\kappa' d^2}{\lambda}\sqrt{\tau(E)\log(12d/\delta)},$$

the bound becomes

$$\left\|\hat{\beta}_E\left(\mathcal{V}_{\tau(E)}\right)-\beta^*\left(\mathcal{V}_{\tau(E)}\right)\right\|_2 \leq \frac{4K'\sqrt{d\tau(E)\log(4d/\delta)}}{\lambda n}.$$

The proof concludes by letting $K \triangleq 4K'$, $\kappa \triangleq 4\kappa'$ and noting that since $\mathcal{D}_{\tau(E)} \subset \mathcal{V}_{\tau(E)}$ by construction, the statement holds true over $\mathcal{D}_{\tau(E)}$ (projecting onto a subspace cannot increase the $\ell 2$-norm).

A.2.1 Proof of Lemma A.4

For the first statement, note that for all $k \neq j \leq r$,

$$\mathbb{E}\left[f_k^\top x_t x_t^\top f_j\right] = f_k^\top \mathbb{E}\left[x_t x_t^\top\right] f_j = \lambda_j f_k^\top f_j,$$

as f_j is (by definition) an eigenvector of $\Sigma = \mathbb{E}\left[x_t x_t^\top\right]$ for eigenvalue λ_j. Note that the $f_j^\top x_t x_t^\top f_k = (f_j^\top x_t)(f_k^\top x_t)$ are random variables that are independent across t. Further, by Cauchy-Schwarz,

$$\left|(f_k^\top x_t)(f_j^\top x_t)\right| \leq \|f_k\|_2\|f_j\|_2\|x_t\|_2^2 = \|x_t\|_2^2 \leq d.$$

Therefore, we can apply Hoeffding with a union bound over the r^2 choices of (f_k, f_j) to show that with probability at least $1-\delta'$,

$$\left|\sum_{t=1}^{\tau(E)} f_k^\top x_t x_t^\top f_j - \lambda_j \tau(E) f_k^\top f_j\right| \leq d\sqrt{2\tau(E)\log(2r^2/\delta')}.$$

Note now that for all $z \in \Sigma$, we can write $z = \sum_{k=1}^{r} \left(z^\top f_k\right) f_k$, and as such

$$\begin{aligned}
&\left| \sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z - \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) \lambda_j \tau(E) f_k^\top f_j \right| \\
&= \left| \sum_{t=1}^{\tau(E)} \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) f_k^\top x_t x_t^\top f_j - \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) \lambda_j \tau(E) f_k^\top f_j \right| \\
&= \left| \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) \left(\sum_t f_k^\top x_t x_t^\top f_j - \lambda_j \tau(E) f_k^\top f_j \right) \right| \\
&\le d\sqrt{2\tau(E)\log(2r^2/\delta')} \sum_{k,j=1}^{r} |z^\top f_k||z^\top f_j| \\
&\le rd\sqrt{2\tau(E)\log(2r^2/\delta')} \|z\|_2^2,
\end{aligned}$$

where the last step follows from the fact that by Cauchy-Schwarz,

$$\sum_{k=1}^{r} |z^\top f_k| \le \sqrt{\sum_{k=1}^{r} 1^2} \sqrt{\sum_{k=1}^{r} (z^\top f_k)^2} = \sqrt{r}\|z\|_2.$$

Hence, for $z \in \Sigma$, remembering $f_k^\top f_j = 0$ when $k \neq j$ and $f_k^\top f_k = 1$, and noting $\|z\|_2^2 = \sum_{k=1}^{r} (z^\top f_k)^2$, we get that

$$\begin{aligned}
\sum_{t=1}^{\tau(E)} z^\top x_t x_t^\top z &\ge \sum_{k,j=1}^{r} (z^\top f_k)(z^\top f_j) \lambda_j \tau(E) f_k^\top f_j - rd\sqrt{2\tau(E)\log(2r^2/\delta')} \|z\|_2^2 \\
&= \sum_{k=1}^{r} \lambda_k \tau(E) (z^\top f_k)^2 - rd\sqrt{2\tau(E)\log(2r^2/\delta')} \|z\|_2^2 \\
&\ge \lambda_r \tau(E) \sum_{k=1}^{r} (z^\top f_k)^2 - rd\sqrt{2\tau(E)\log(2r^2/\delta')} \|z\|_2^2 \\
&= \left(\lambda_r \tau(E) - 2rd\sqrt{\tau(E)\log(2r/\delta')} \right) \|z\|_2^2.
\end{aligned}$$

For the second statement, we remind the reader that the costs of modification are such that $\left|\Delta_t(k)^2\right| \le \left(\max_{i,j} \left\{ \frac{B^i}{c^i(j)} \right\} \right)^2$, and that within any epoch ϕ, the Δ_t's are independent of each other. We can therefore apply Hoeffding's inequality and a union bound (over $k \in D_{\tau(E)} \subset [d]$) to show that with probability at least $1 - \delta'$, for any $k \in D_{\tau(E)}$, there exists an epoch $\phi(k) \le E$ (pick any ϕ in which k is modified) such that

$$\begin{aligned}
\sum_{t \in \phi(k)} e_k^\top \Delta_t \Delta_t^\top e_k &\ge n \, \mathbb{E}\left[\Delta_t(k)^2\right] - \left(\max_{i,j} \left\{ \frac{B^i}{c^i(j)} \right\} \right)^2 \sqrt{2n \log(d/\delta')} \\
&\ge n \min_{i \in [l], j \in [d]} \left\{ \pi^i \left(\frac{B^i}{c^i(j)} \right)^2 \right\} - \left(\max_{i,j} \left\{ \frac{B^i}{c^i(j)} \right\} \right)^2 \sqrt{2n \log(d/\delta')}.
\end{aligned}$$

The last inequality holds noting that k can be modified in period $\phi(k)$ only if there exists a cost type i on the support of $\mathcal{C}$ such that k is a best response to $\hat{\beta}_{\phi(k)-1}$; in turn, k is modified with probability π^i by amount $\Delta(k) = B^i/c^i(k)$, leading to

$$\mathbb{E}\left[\Delta_t(k)^2\right] \geq \pi^i \left(\frac{B^i}{c^i(k)}\right)^2.$$

Since $\Delta_t(k)\Delta_t(j) = 0$ when $k \neq j$ as a single direction is modified at a time, note that for all $z \in \mathcal{D}_{\tau(E)}$, we have

$$\begin{aligned}
&\sum_{t \leq \tau(E)} z^\top \Delta_t \Delta_t^\top z \\
&= \sum_{t \leq \tau(E)} \sum_{k=1}^{d} \Delta_t(k)^2 z^\top e_k e_k^\top z \\
&= \sum_{k=1}^{d} \sum_{t \leq \tau(E)} \Delta_t(k)^2 (z^\top e_k)^2 \\
&\geq \sum_{k \in D_{\tau(E)}} \sum_{t \in \phi(k)} \Delta_t(k)^2 (z^\top e_k)^2 \\
&\geq \sum_{k \in D_{\tau(E)}} \left(n \min_{i \in [l], j \in [d]} \left\{ \pi^i \left(\frac{B^i}{c^i(j)}\right)^2 \right\} - \left(\max_{i,j} \left\{ \frac{B^i}{c^i(j)} \right\} \right)^2 \sqrt{2n \log(d/\delta')} \right) (z^\top e_k)^2 \\
&= \left(n \min_{i \in [l], j \in [d]} \left\{ \pi^i \left(\frac{B^i}{c^i(j)}\right)^2 \right\} - \left(\max_{i,j} \left\{ \frac{B^i}{c^i(j)} \right\} \right)^2 \sqrt{2n \log(d/\delta')} \right) \sum_{k \in D_{\tau(E)}} (z^\top e_k)^2.
\end{aligned}$$

For $z \in \mathcal{D}_{\tau(E)}$, $\sum_{k \in D_{\tau(E)}} (z^\top e_k)^2 = \|z\|_2^2$, and the second inequality immediately holds.

Finally, let us prove the last inequality. Take $(k, j) \in [d]^2$, and let us write $W_t = e_k^\top x_t \Delta_t^\top e_j$. First, note that x_t and Δ_t are independent: in epoch ϕ, the distribution of Δ_t is a function of $\hat{\beta}_{\phi-1}$ (and $\mathcal{C}$) only, which only depends on the realizations of x, ε, Δ in previous time steps. Further, x_t is independent of the history of features and modifications up until time $t-1$ included. Hence, it must be the case that

$$\begin{aligned}
\mathbb{E}\left[W_t | W_{t-1}, \ldots, W_1\right] &= \mathbb{E}\left[\mathbb{E}\left[e_k^\top x_t \middle| \Delta_t, W_{t-1}, \ldots, W_1\right] \Delta_t^\top e_j \middle| W_{t-1}, \ldots, W_1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e_k^\top x_t\right] \Delta_t^\top e_j \middle| W_{t-1}, \ldots, W_1\right] \\
&= \mathbb{E}\left[e_k^\top x_t\right] \cdot \mathbb{E}\left[\Delta_t^\top e_j \middle| W_{t-1}, \ldots, W_1\right] \\
&= 0,
\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}\left[x_t\right] = 0$. Further,

$$\left| e_k^\top x_t \Delta_t^\top e_j \right| = |x_t(k)| |\Delta_t(j)| \leq \max_{i,k} \left\{ \frac{B^i}{c^i(k)} \right\}.$$

We can therefore apply Lemma A.1 and a union bound over all $(k, j) \in [d]^2$ to show that with probability at least $1 - \delta'$,

$$\left|\sum_{t=1}^{\tau(E)} e_k^\top x_t \Delta_t^\top e_j\right| \leq \max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{2\tau(E)\log(2d^2/\delta')}.$$

In particular, we get that for all $z \in \mathbb{R}^d$,

$$\begin{aligned}
\left|\sum_{t\in E} z^\top x_t \Delta_t^\top z\right| &= \left|\sum_{k,j}\sum_{t\in E}(z^\top e_k)(z^\top e_j)e_k^\top x_t \Delta_t^\top e_j\right| \\
&\leq \sum_{k,j}|z^\top e_k||z^\top e_j|\left|\sum_{t\in E} e_k^\top x_t \Delta_t^\top e_j\right| \\
&\leq \max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{2\tau(E)\log(2d^2/\delta')}\left(\sum_k |z^\top e_k|\right)^2 \\
&\leq 2d\max_{i,k}\left\{\frac{B^i}{c^i(k)}\right\}\sqrt{\tau(E)\log(2d/\delta')}\|z\|_2^2,
\end{aligned}$$

where the last step follows from the fact that by Cauchy-Schwarz,

$$\left(\sum_k |z^\top e_k|\right)^2 = \left(\sum_k |z(k)|\right)^2 \leq \sum_k 1^2 \cdot \sum_k z(k)^2 = d\cdot\|z\|_2^2.$$

We conclude the proof with a union bound over all three inequalities, taking $\delta' = 3\delta$.

B Proof of Theorem 5.2

We drop the $\tau(E)$ subscripts when clear from context. We first note that $\hat{\beta}_E$ is a least-square solution.

Claim B.1.

$$\hat{\beta}_E \in LSE(\tau(E)).$$

Proof. This follows immediately from noting that

$$\left(\bar{X}\hat{\beta}_E - \bar{Y}\right)^\top\left(\bar{X}\hat{\beta}_E - \bar{Y}\right) = \left(\bar{X}\beta_E - \bar{Y}\right)^\top\left(\bar{X}\beta_E - \bar{Y}\right),$$

as $\bar{X}^\top v = \bar{X}(\mathcal{U})^\top v = 0$ by definition of $\mathcal{U}$, and since $v \in \mathcal{U}^\perp$. □

Second, we show that $\hat{\beta}_E$ has large norm:

Claim B.2.

$$\left\|\hat{\beta}_E\right\|_2 \geq \alpha.$$

Proof. First, we note that necessarily, $\beta_E \in \mathcal{U}_{\tau(E)}$. Suppose not, then we can write

$$\beta_E = \beta_E\left(\mathcal{U}_{\tau(E)}\right) + \beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right),$$

with $\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right) \neq 0$. By the same argument as in Claim B.1, $\beta_E\left(\mathcal{U}_{\tau(E)}\right)$ is a least-square solution. Using orthogonality of $\mathcal{U}_{\tau(E)}$ and $\mathcal{U}^{\perp}_{\tau(E)}$ and the fact that $\left\|\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right)\right\|_2 > 0$, we have

$$\|\beta_E\|^2 = \left\|\beta_E\left(\mathcal{U}_{\tau(E)}\right)\right\|_2^2 + \left\|\beta_E\left(\mathcal{U}^{\perp}_{\tau(E)}\right)\right\|_2^2 > \left\|\beta_E\left(\mathcal{U}_{\tau(E)}\right)\right\|_2^2.$$

This contradicts β_E being a minimum norm least-square solution. Hence, it must be the case that $\beta_E \in \mathcal{U}_{\tau(E)}$. Since $v \in \mathcal{U}^{\perp}_{\tau(E)}$, we have that β_E and v are orthogonal with $\|v\|_2 = 1$, implying

$$\left\|\hat{\beta}_E\right\|_2^2 = \|\beta_E\|_2^2 + \alpha^2\|v\|_2^2 \geq \alpha^2.$$

This concludes the proof. ∎

We argue that such a solution places a large amount of weight on currently unexplored features:

Lemma B.3. *At time $\tau(E)$, suppose rank $\left(\mathcal{U}_{\tau(E)}\right) \leq [d]$. Suppose $n \geq \frac{\kappa d^2}{\lambda}\sqrt{\tau(E)\log(12d/\delta')}$. Take any α with*

$$\alpha \geq \gamma\left(\sqrt{d} + \frac{Kd\sqrt{T\log(4d/\delta')}}{\lambda n}\right),$$

where γ is a constant that depends only on $\mathcal{C}$. With probability at least $1-\delta'$, there exists $i \in [l]$ and a feature $k \notin D_{\tau(E)}$ with

$$\frac{\left|\hat{\beta}_E(k)\right|}{c^i(k)} > \frac{\left|\hat{\beta}_E(j)\right|}{c^i(j)},\ \forall j \in D_{\tau(E)}.$$

Proof. Since $\hat{\beta}_E \in LSE(\tau(E))$, it must be by Theorem 4.1 that with probability at least $1-\delta'$,

$$\begin{aligned}\sqrt{\sum_{k\in D}\left(\hat{\beta}_E(k) - \beta^*(k)\right)^2} &\leq \frac{K\sqrt{d\tau(E)\log(4d/\delta')}}{\lambda n} \\ &\leq \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.\end{aligned} \tag{4}$$

First, since $z \to \sqrt{\sum_{k\in D} z(k)^2}$ defines a norm (in fact, the $\ell 2$-norm in $\mathbb{R}^{|D|}$), it must be the case that

$$\sqrt{\sum_{k\in D}(z(k) - z'(k))^2} \geq \sqrt{\sum_{k\in D} z(k)^2} - \sqrt{\sum_{k\in D} z'(k)^2}.$$

In turn, plugging this in Equation (4), we obtain

$$\begin{aligned}\sqrt{\sum_{k\in D}\hat{\beta}_E(k)^2} &\leq \sqrt{\sum_{k\in D}\beta^*(k)^2} + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n} \\ &\leq \|\beta^*\|_2 + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n} \\ &\leq \sqrt{d} + \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.\end{aligned}$$

By the triangle inequality and the lemma's assumption, we also have that

$$\sqrt{\sum_{k\in D}\hat{\beta}_E(k)^2}+\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}\geq\|\hat{\beta}\|_2\geq\alpha.$$

Combining the last two equations, we obtain

$$\sqrt{d}+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}+\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2},\geq\alpha$$

which implies that for $\alpha\geq\gamma\left(\sqrt{d}+\frac{Kd\sqrt{T\log(4d/\delta')}}{\lambda n}\right)$, we have:

$$\begin{aligned}\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}&\geq\alpha-\sqrt{d}-\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\\&\geq\alpha-\sqrt{d}-\frac{Kd\sqrt{T\log(4d/\delta')}}{\lambda n}\\&\geq\sqrt{d}\,(\gamma-1)\left(1+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\right).\end{aligned}$$

Second, note that Equation (4) implies immediately that for any $j\in D_T$,

$$\left|\hat{\beta}_E(j)-\beta^*(j)\right|\leq\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n},$$

and in turn,

$$\left|\hat{\beta}_E(j)\right|\leq|\beta^*(j)|+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}\leq 1+\frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.$$

Therefore,

$$\sqrt{\sum_{k\notin D}\hat{\beta}_E(k)^2}\geq\sqrt{d}\,(\gamma-1)\max_{j\in D}\hat{\beta}_E(j).$$

Hence, there must exist feature $k\notin D$ with

$$\hat{\beta}_E(k)\geq(\gamma-1)\max_{j\in D}\hat{\beta}_E(j).$$

Picking γ such that for some $i\in[l]$,

$$\gamma-1\geq\max_{j\in D}\frac{c^i(k)}{c^i(j)}$$

yields the result immediately. $\square$

The proof of Theorem 5.2 follows directly from Lemma B.3 and a union bound over the first d epochs. With probability at least $1 - d\delta'$, for every epoch $E \in [d]$, there is a feature $k \notin D_{\tau(E)}$ such that for some $i \in [l]$,

$$\frac{\left|\hat{\beta}_E(k)\right|}{c^i(k)} > \frac{\left|\hat{\beta}_E(j)\right|}{c^i(j)} \;\forall j \in D_{\tau(E)}.$$

This implies that there exists $k \in D_{\tau(E+1)}$ but $k \notin D_{\tau(E)}$. Applying this d times, we have that if $T \geq dn$, necessarily $D_T = [d]$. We can then apply Theorem 4.1 to then show that with probability at least $1 - \delta'$

$$\left\|\hat{\beta}_{T/n} - \beta^*\right\|_2 \leq \frac{K\sqrt{dT\log(4d/\delta')}}{\lambda n}.$$

Taking a union bound over the two above events and $\delta = 2d\delta'$, we get the theorem statement with probability at least $1 - \delta'(d+1) \geq 1 - \delta$.