Strategy-Aware Contextual Bandits

Anonymous Author(s)
Affiliation
Address
email

Abstract

Algorithmic tools are often used to make decisions about people in high-stakes domains. In the presence of such automated decision making, there is incentive for strategic agents to modify their input to the algorithm in order to receive a more desirable outcome. While previous work on *strategic classification* attempts to capture this phenomenon, these models fail to take into account the *multiple actions* a decision maker usually has at their disposal, and the fact that they often have access only to *bandit feedback*. In contrast, we capture this setting as a *contextual bandit* problem, in which a decision maker must take actions based on a sequence of strategically modified contexts. We provide a low-strategic-regret algorithm for the two action setting, and prove that sublinear strategic regret is generally not possible for settings in which the number of actions is greater than two. Along the way, we obtain impossibility results for multi-class strategic classification which may be of independent interest.

1 Introduction

When subjugated to algorithmic decision making, decision subjects may strategically modify their input to the algorithm in order to receive a better outcome. This may be viewed as a form of *distribution shift* in which the data distribution the machine learning (ML) algorithm is operating on has been strategically altered. It is therefore desirable to design ML algorithms capable of operating in the presence of such *strategic interactions*. While this area of study has received much recent interest in the literature under the name of *strategic classification* (see Appendix A for a list of related work), current work in this area fails to capture several salient features of reality, as most models only consider *binary classification* settings under *full feedback*. In contrast, we study a model in which a decision maker must assign a strategic agent one of *several actions* under *bandit feedback*. Our model captures several settings of practical interest including (1) personalized shopping, in which an online platform (decision maker) must decide which level of discount (action) to give to give to customers (strategic agents) given their previous buying patterns and (2) lending, in which a bank (decision maker) needs to decide which type of loan (e.g., high-interest, low-interest, no loan) to offer to applicants (strategic agents) given their credit history. Our main contributions are as follows: (1) We formalize the *strategy-aware contextual bandit setting*, in which a decision maker must assign a sequence of strategic agents one of n actions, and observes their outcomes under bandit feedback. (2) We provide a low-strategic-regret algorithm for the setting in which $n = 2$, using standard tools from strategic classification and online learning. (3) We show that in sharp contrast to the two-action setting, when $n \geq 3$ situations exist in which achieving sublinear strategic regret is not possible. Along the way, we obtain impossibility results for multi-class strategic classification which may be of independent interest.

Submitted to Workshop on Robustness in Sequence Modeling at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Interaction protocol: Contextual bandits under strategic manipulations

For each round $t \in [T]$:

1. decision maker publicly commits to a mapping $\pi_t : \mathcal{X} \to \mathcal{A}$. Agent Γ_t arrives with private type $\mathbf{x}_t$.
2. Agent t strategically selects context $\mathbf{x}'_t$ according to Assumption 2.3.
3. decision maker observes context $\mathbf{x}'_t$ and plays action $a_t = \pi_t(\mathbf{x}'_t)$.
4. decision maker receives reward $r_t^{(l)}(a_t)$ and agent receives reward $r^{(ag)}(a_t)$.

Figure 1: Description of the strategy-aware contextual bandit setting we consider, which may be viewed as a generalization of the contextual bandit problem with linear rewards.

2 Model and preliminaries

We consider a setting in which a *decision maker* interacts with a sequence of T strategic *agents*. At each time $t \in [T]$, a new agent Γ_t arrives with a *private type* $\mathbf{x}_t \in \mathcal{X} \subset [-L, L]^d$ for $L \in \mathbb{R}_+$, and presents a *context* $\mathbf{x}'_t \in \mathbb{R}^d$ to the decision maker. Note that it is possible for $\mathbf{x}'_t = \mathbf{x}_t$, but the two do not necessarily coincide. Given a context $\mathbf{x}'_t$, the decision maker takes an *action* a_t from some set of actions $\mathcal{A}$, where $\mathcal{A} = \{1, \dots, n\}$. For each action $a \in \mathcal{A}$, we assume that the decision maker's *expected reward* $\mathbb{E}[r_t^{(l)}(a)] \in [-1, 1]$ is a linear function of the agent's private type. In particular, we make the following assumption on the decision maker's reward.

Assumption 2.1 (Linearity of Rewards)**.** *The decision maker reward of each arm $a \in \mathcal{A}$ at time t is given by*

$$r_t^{(l)}(a) = \langle \boldsymbol{\theta}_a, \mathbf{x}_t \rangle + \varepsilon_t$$

for some $\boldsymbol{\theta}_a \in \mathbb{R}^d$ which is unknown to the decision maker, where ε_t is zero-mean sub-Gaussian noise.

Note that if $\mathbf{x}'_t = \mathbf{x}_t \ \forall t \in [T]$, our setting reduces to that of linear contextual bandits (e.g, [20, 2]). In contrast to the decision maker's reward, we assume that each agent's reward $r_t^{(ag)}(a)$ is a function of the action alone, i.e., $r_t^{(ag)}(a) = r^{(ag)}(a), \forall t \in [T]$, and is known to the decision maker. In the personalized shopping example, a consumer's private type can be thought of as the type of customer they are (e.g., loyal customer, new customer, thrifty customer, etc.). If customers knew the online platform was going to offer them a discount based on their buying habits, they may strategically alter their short-term behavior on the platform (i.e., their context) in order to obtain the highest discount possible.

The decision maker's *policy* $\pi_t : \mathcal{X} \to \mathcal{A}$ at time t is a mapping from agent contexts to actions. In particular, the goal of the decision maker is to deploy a sequence of policies $\pi_1, \dots \pi_T$ in order to minimize *strategic regret*, which is defined as follows:

Definition 2.2 (Strategic Regret)**.** *The expected strategic regret of a sequence of policies $\{\pi_t\}_{t=1}^T$ is defined as the cumulative expected reward of $\pi^*(\mathbf{x}) := \arg\max_{a \in \mathcal{A}} \mathbb{E}[r(a)|\mathbf{x}]$, the optimum-in-hindsight policy given the agents' private types, minus the cumulative expected reward of the deployed sequence of policies. Formally,*

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} \underbrace{\mathbb{E}[r_t^{(l)}(\pi^*(\mathbf{x}_t))|\Gamma_t]}_{\textit{reward of optimal policy given private type}} - \underbrace{\mathbb{E}[r_t^{(l)}(\pi_t(\mathbf{x}'_t))|\Gamma_t]}_{\textit{realized reward given context}}$$

A decision maker with sublinear strategic regret will approach the performance of the best-in-hindsight policy if agents are *not* strategic as $T \to \infty$, thereby learning to "account" for the strategic behavior of agents in the long run. Given the decision maker policy π_t, it is natural for an agent to choose their context in a way which maximizes their reward. Specifically, we assume that agent Γ_t determines their context by strategically modifying their private type based on the decision maker's policy, subject to a constraint on the amount of modification which is possible. In addition to being a common assumption in the literature (e.g., [18, 14]), this *budget* constraint on the amount an agent can modify their private type reflects the fact that agents have inherent constraints on the amount of time and resources they can spend on modification.

Assumption 2.3 (Agent Best Response)**.** *We assume that agent Γ_t best-responds to the decision maker's policy π_t in order to maximize their expected reward, subject to the constraint that their*

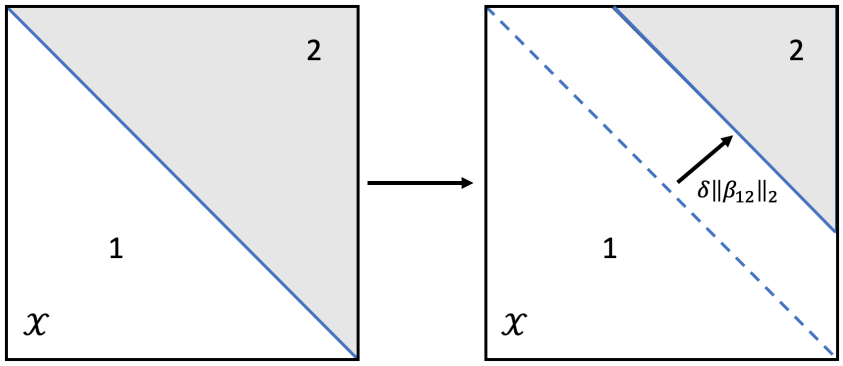

Figure 2: (Left: Proposition 3.1) The optimal policy in the non-strategic setting assigns action 1 to the left of the line $\langle \boldsymbol{\beta}_{12}, \mathbf{x} \rangle = 0$ (blue) and action 2 to the right. (Right: Proposition 3.2) In the strategic setting, the optimal decision boundary is shifted towards the $a = 2$ region by $\delta \|\boldsymbol{\beta}_{12}\|_2$.

context $\mathbf{x}'_t$ is within a δ-radius of their private type $\mathbf{x}_t$. Formally, we assume that the agent solves the following optimization to determine their context:

$$\mathbf{x}'_t \in \arg\max_{\mathbf{x}' \in \mathcal{X}} \; r^{(ag)}(\pi_t(\mathbf{x}'))$$
$$\textit{s.t.} \; \|\mathbf{x}' - \mathbf{x}_t\|_2 \leq \delta$$

Furthermore, we assume that if an agent is indifferent between modifying their context and not modifying, they choose not to modify. See Figure 1 for a summary of the setting we consider.

3 Optimal policies in the presence of strategic interactions

3.1 Sublinear strategic regret for two actions

We now outline an algorithm which achieves sublinear strategic regret in the two-action setting. W.l.o.g., we assume that $r^{(ag)}(2) > r^{(ag)}(1)$ for the remainder of the subsection. While our algorithm is similar to previous algorithms from the "strategic classification" literature (e.g., [12, 3]), we view it as an important structural result which (1) illustrates that sublinear strategic regret *is* possible in the two action setting and (2) provides intuition as to why this is generally not possible in the case of 3+ actions. We begin with a result for the offline, non-strategic setting.

Proposition 3.1. *In the two-action setting in which agents are not strategic (i.e., $\mathbf{x}'_t = \mathbf{x}_t, \forall t \in [T]$) and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are known to the decision maker, the optimal policy assigns action $a_t = 1$ if $\langle \boldsymbol{\beta}_{12}, \mathbf{x}'_t \rangle \geq 0$ and action $a_t = 2$ otherwise, where $\boldsymbol{\beta}_{12} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$.*

Proposition 3.1 follows straightforwardly from Assumption 2.1. While the decision boundary of Proposition 3.1 is optimal in the non-strategic setting, its performance may suffer in the strategic setting (i.e., if $\mathbf{x}'_t \neq \mathbf{x}_t$). In order to account for potential strategic manipulations, the optimal decision boundary needs to be *shifted* by an amount proportional to the agents' ability to manipulate their private type. Intuitively, this shift will cause agents who would've manipulated in order to receive action a_1 not to do so.

Proposition 3.2. *In the two-action setting in which agents strategically reveal their contexts according to Assumption 2.3, the optimal policy assigns action $a_t = 1$ if $\langle \boldsymbol{\beta}_{12}, \mathbf{x}'_t \rangle \geq -\delta \|\boldsymbol{\beta}_{12}\|_2$ and action $a_t = 2$ otherwise, where $\boldsymbol{\beta}_{12} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$. Moreover, the performance of such a policy when agents strategically reveal their contexts according to Assumption 2.3 matches that of Proposition 3.1 in the non-strategic setting (i.e., when $\mathbf{x}'_t = \mathbf{x}_t, \forall t \in [T]$).*

See Figure 2 for a visual depiction of the policies of Propositions 3.1 and 3.2. Given the result of Proposition 3.2, it is rather straightforward to achieve sublinear strategic regret in the online setting by using the standard "explore-exploit" technique. In particular, under minor distributional assumptions on the population of agents, the decision maker can apply the following procedure to achieve $\mathcal{O}(T^{2/3})$ strategic regret. See the Appendix for a formal statement.

1. For the first $T^{1/3}$ rounds, assign contexts without using agent contexts. Since agents have no incentive to modify, $\mathbf{x}'_t = \mathbf{x}_t, \forall t \in [T^{1/3}]$.
2. Use observed contexts and rewards to estimate $\boldsymbol{\beta}_{12}$ via ordinary least squares as $\hat{\boldsymbol{\beta}}_{12}$.
3. For $t > T^{1/3}$, assign action $a_t = 1$ if $\langle \hat{\boldsymbol{\beta}}_{12}, \mathbf{x}'_t \rangle \geq \delta \|\hat{\boldsymbol{\beta}}_{12}\|_2$ and action $a_t = 2$ otherwise.

3.2 $\Omega(T)$ strategic regret for more than two actions

In the previous subsection, we showed that it is generally possible to achieve sublinear strategic regret in the two-action strategy-aware contextual bandit setting by applying standard ideas from the strategic classification and online learning literature. However, we will now show that this intuition does not carry over to the 3+ action setting. Furthermore, we show that there exist situations for which *no policy* can achieve sublinear strategic regret if the decision maker has three or more actions. As was the case in Section 3.1, we begin with a result for the offline setting. Note that this result immediately implies that similar "bad examples" exist for settings in which $n > 3$ and $d > 2$.

Theorem 3.3. *Suppose $\mathcal{X} = \mathbb{R}^2$ $n = 3$, and $r^{(ag)}(2) > r^{(ag)}(1) = r^{(ag)}(3)$. For $\boldsymbol{\theta}_1 = [1 \;\; 0.5]^\top$, $\boldsymbol{\theta}_2 = [0 \;\; 1]^\top$, $\boldsymbol{\theta}_3 = [-1 \;\; 0.5]^\top$, there exists a distribution over agent private types such that no policy can achieve perfect classification if agents choose their contexts according to Assumption 2.3.*

Proof Sketch. The proof of Theorem 3.3 proceeds by constructing a "bad" distribution over agents (shown in orange in Figure 3) in which it is *necessary* to shift the separating hyperplanes $\boldsymbol{\beta}_{12}$ and $\boldsymbol{\beta}_{23}$ (defined analogously to $\boldsymbol{\beta}_{12}$ in Proposition 3.2) in order to achieve perfect classification on a certain subpopulation of agents (far-left and far-right orange lines in Figure 3). Next, we show that shifting $\boldsymbol{\beta}_{12}$ and $\boldsymbol{\beta}_{23}$ by such an amount will make it *impossible* to classify another agent subpopulation (orange dot in Figure 3). Finally, we conclude that perfect classification of this distribution of strategic agents is not possible. Note that this is in contrast to the non-strategic setting, in which this agent population is trivially linearly separable. The following corollary follows immediately.

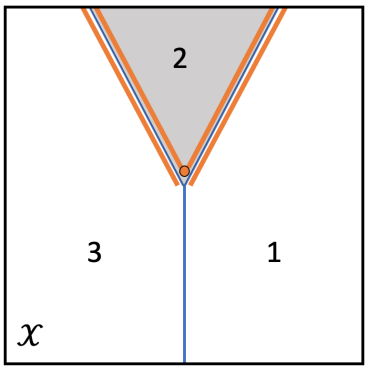

Figure 3: Visualization of the "bad example" from Theorem 3.3. Blue lines represent optimal decision boundaries in the non-strategic setting. The distribution over agent private types for which perfect classification is not possible puts probability mass $1/5$ on the orange dot and each of the orange lines.

Corollary 3.4. *In the setting of Theorem 3.3, no online learning algorithm can achieve sublinear strategic regret.*

4 Conclusion and future research

In this extended abstract, we showed a separation between strategic classification and multi-class strategic classification, which immediately implies an impossibility result for the strategy-aware contextual bandit setting with three or more actions. The most important direction for future work is to identify a set of *sufficient* or *necessary* conditions for which the impossibility result of Theorem 3.3 holds. In settings for which it does not hold, it would be interesting to see if ideas similar to the explore-exploit algorithm of Section 3.1 achieve sublinear strategic regret for 3+ actions. Another exciting direction would be to give a lower bound on strategic regret in the two-action setting. In particular, it is unclear if the $\Omega(T^{1/2})$ lower bound from the non-strategic setting applies, or if $\Omega(T^{2/3})$ strategic regret is the best one can do. If the $\Omega(T^{1/2})$ lower bound does indeed hold, deriving an algorithm which obtains this rate would be interesting. Finally, one may argue that strategic regret is too strong of a benchmark given the negative result of Theorem 3.3. Therefore, it may be desirable to derive algorithms with low *Stackelberg regret* [4] in the strategy-aware contextual bandit setting.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[3] Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.

[4] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pages 61–78, 2015.

[5] Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pages 1691–1715. PMLR, 2022.

[6] Yatong Chen, Jialu Wang, and Yang Liu. Linear classifiers that encourage constructive adaptation. *arXiv preprint arXiv:2011.00355*, 2020.

[7] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

[8] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[9] 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.

[10] Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[11] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

[12] 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.

[13] Keegan Harris, Valerie Chen, Joon Sik Kim, Ameet Talwalkar, Hoda Heidari, and Zhiwei Steven Wu. Bayesian persuasion for algorithmic recourse. *arXiv preprint arXiv:2112.06283*, 2021.

[14] Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

[15] Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, and Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR, 2022.

[16] Elad Hazan and Nimrod Megiddo. Online learning with prior knowledge. In *International Conference on Computational Learning Theory*, pages 499–513. Springer, 2007.

[17] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pages 4687–4697. PMLR, 2021.

[18] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

[19] Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pages 6243–6253. PMLR, 2021.

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[21] Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via context multi-armed bandits. In *Proceedings of AISTATS*, 2009.

[22] Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

A Related Work

Our work is closely related to two lines of work: *learning in the presence of strategic behavior* [3, 5, 9, 12, 14, 15, 18, 13, 6, 11, 17, 19, 7] and *contextual bandits* (see [16, 21, 22, 20, 8, 1, 10, 2] for a highly incomplete list).

Strategic learning. The problem of *strategic classification* was first introduced in Hardt et al. [12]. The original formulation considers a sequential game between a "jury" (decision maker), who publishes a classifier, and a "contestant" (strategic agent), who best responds by strategically modifying their observable features. In the years since, there have been extensions to online learning (Chen et al. [7], Dong et al. [9]), repeated interactions (Harris et al. [14]), and social learning settings (Bechavod et al. [5]), although to the best of our knowledge, we are the first to consider multi-class strategic classification under bandit feedback.

Contextual bandits. The contextual bandit setting can be thought of a multi-armed bandit problem, in which the bandit algorithm has access to extra information about the reward of each action (i.e., the *context*) at every round. Various relationships between contexts and rewards have been studied; several of the most common are *Lipschitz contextual bandits* (e.g., Hazan and Megiddo [16], Lu et al. [21], Slivkins [22]), in which the expected reward of each arm is Lipschitz in the context, *linear contextual bandits* (e.g., Li et al. [20], Chu et al. [8]), where the expected reward is linear in the context, and *agnostic algorithms* (e.g., Agarwal et al. [2], Foster and Rakhlin [10]), which make no assumptions on the underlying reward distribution, but assume access to various classification/regression oracles. In particular, our setting is most similar to that of linear contextual bandits, although with the notable exception that in our setting, the expected reward of each action is linear in the agent's *private type*. Since each agent's private type is not observed by the decision maker, this introduces significant additional difficulties in the decision maker's learning problem and leads to our impossibility results in Section 3.2.

B Proofs from Section 3

Proposition B.1. *In the two-action setting in which agents strategically reveal their contexts according to Assumption 2.3, the optimal policy assigns action $a_t = 1$ if $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle \geq -\delta \|\boldsymbol{\beta}_{12}\|_2$ and action $a_t = 2$ otherwise, where $\boldsymbol{\beta}_{12} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$.*

Proof. Denote the unit vector along the direction of $\boldsymbol{\beta}_{12}$ as $\mathbf{n} = \frac{\boldsymbol{\beta}_{12}}{\|\boldsymbol{\beta}_{12}\|_2}$. We proceed on a case-by-case basis.

Case 1: The policy of Proposition 3.1 assigns action $a_t = 1$ to agent Γ_t:

Since $a_t = 1$, $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle \geq 0$. In order to move across the boundary under as little modification as possible, the agent should modify their context in the direction of $-\mathbf{n}$. However, the most they can manipulate their context by is δ, and $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t - \delta \mathbf{n} \rangle \geq -\delta \|\boldsymbol{\beta}_{12}\|_2$ since $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle \geq 0$. Therefore no valid manipulations exist for which the policy of Proposition 3.2 assigns action $a = 2$ to agent Γ_t.

Case 2: The policy of Proposition 3.1 assigns action $a_t = 2$ to agent Γ_t:

Since $a_t = 2$, $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle < 0$. If agent Γ_t modifies their private from $\mathbf{x}_t$ to context $\mathbf{x}_t - \delta \mathbf{n}$, $\langle \boldsymbol{\beta}_{12}, \mathbf{x}'_t \rangle = \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \mathbf{x}_t \rangle - \delta \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \mathbf{n} \rangle < -\delta \|\boldsymbol{\beta}_{12}\|_2$. Therefore, they can receive action $a = 2$ under the policy of Proposition 3.2 by strategically modifying their private type from $\mathbf{x}_t$ to context $\mathbf{x}_t - \delta \mathbf{n}$.

□

Theorem B.2. *Suppose $\mathcal{X} = \mathbb{R}^2$ and $n = 3$. For $\boldsymbol{\theta}_1 = [1 \;\; 0.5]^\top$, $\boldsymbol{\theta}_2 = [0 \;\; 1]^\top$, $\boldsymbol{\theta}_3 = [-1 \;\; 0.5]^\top$, there exists a distribution over agent private types such that no policy can achieve perfect classification if agents choose their contexts according to Assumption 2.3.*

Proof. Recall that assigning $a_t = \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\theta}_a, \mathbf{x}_t \rangle$ will achieve perfect classification in the non-strategic setting. Define $\boldsymbol{\beta}_{12} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = [1 \quad -0.5]^\top$, $\boldsymbol{\beta}_{13} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_3 = [2 \quad 0]^\top$, and

$\boldsymbol{\beta}_{23} := \boldsymbol{\theta}_2 - \boldsymbol{\theta}_3 = [1 \;\; 0.5]^\top$ Assigning actions according to the following set of decision boundaries is equivalent to assigning $a_t = \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\theta}_a, \mathbf{x}_t \rangle$:

$$a_t = \begin{cases} a_1 & \text{if } \langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle \geq 0 \text{ and } \langle \boldsymbol{\beta}_{13}, \mathbf{x}_t \rangle \geq 0 \\ a_2 & \text{if } \langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle < 0 \text{ and } \langle \boldsymbol{\beta}_{23}, \mathbf{x}_t \rangle \geq 0 \\ a_3 & \text{if } \langle \boldsymbol{\beta}_{13}, \mathbf{x}_t \rangle < 0 \text{ and } \langle \boldsymbol{\beta}_{23}, \mathbf{x}_t \rangle < 0 \end{cases}$$

See Figure 3 for a visual depiction of the decision space (blue lines).

Consider the following set over agent private types: $\mathbf{x}_1 = [0 \;\; \alpha_1]^\top$, $\mathcal{X}_2 = \{\mathbf{x} : \langle \boldsymbol{\beta}_{12}, \mathbf{x} \rangle = \alpha_2, \; \mathbf{x}[1] > 0\}$, $\mathcal{X}_2' = \{\mathbf{x} : \langle \boldsymbol{\beta}_{12}, \mathbf{x} \rangle = \boldsymbol{\beta}_{12}[2] \cdot \alpha_1, \; \mathbf{x}[1] > 0\}$, $\mathcal{X}_3 = \{\mathbf{x} : \langle \boldsymbol{\beta}_{23}, \mathbf{x} \rangle = -\alpha_2, \; \mathbf{x}[1] < 0\}$, $\mathcal{X}_3' = \{\mathbf{x} : \langle \boldsymbol{\beta}_{23}, \mathbf{x} \rangle = \boldsymbol{\beta}_{23}[2] \cdot \alpha_1, \; \mathbf{x}[1] < 0\}$, where $\alpha_1, \alpha_2 > 0$ and the corresponding distribution over $\mathcal{X}$: $\mathbb{P}(\mathbf{x}_t = \mathbf{x}_1) = \frac{1}{5}$, $\mathbb{P}(\mathbf{x}_t \in \mathcal{X}_2) = \frac{1}{5}$,$\mathbb{P}(\mathbf{x}_t \in \mathcal{X}_2') = \frac{1}{5}$, $\mathbb{P}(\mathbf{x}_t \in \mathcal{X}_3) = \frac{1}{5}$, $\mathbb{P}(\mathbf{x}_t \in \mathcal{X}_3') = \frac{1}{5}$. Suppose that $r(a_2) > r(a_1) = r(a_3)$. Under this setting, we make use of the following two lemmas to identify necessary conditions to achieve perfect classification in the strategic setting:

Lemma B.3. *To prevent gaming, the decision maker must assign action a_2 only if $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle + \varepsilon_{12} < 0$ and $\langle \boldsymbol{\beta}_{23}, \mathbf{x}_t \rangle - \varepsilon_{23} \geq 0$, where $\varepsilon_{12} \geq \delta \|\boldsymbol{\beta}_{12}\|_2 - \alpha_2$ and $\varepsilon_{23} > \delta \|\boldsymbol{\beta}_{23}\|_2 - \alpha_2$.*

Lemma B.4. *To correctly classify agents with $\mathbf{x}_t = \mathbf{x}_1$, the decision maker must assign action a_2 only if $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle + \varepsilon_{12} < 0$ and $\langle \boldsymbol{\beta}_{23}, \mathbf{x}_t \rangle - \varepsilon_{23} \geq 0$, where $\varepsilon_{12} \leq -\boldsymbol{\beta}_{12}[2](\alpha_1 + \delta)$ and $\varepsilon_{23} \leq \boldsymbol{\beta}_{23}[2](\alpha_1 + \delta)$.*

An immediate consequence of the above two lemmas is that unless $\delta \|\boldsymbol{\beta}_{12}\|_2 - \alpha_2 \leq -\boldsymbol{\beta}_{12}[2](\alpha_1 + \delta)$ and $\delta \|\boldsymbol{\beta}_{23}\|_2 - \alpha_2 \leq \boldsymbol{\beta}_{23}[2](\alpha_1 + \delta)$, perfect classification is not possible. In the above setting, this implies that perfect classification is not possible unless $\frac{1}{2}\alpha_1 + \alpha_2 \geq \delta(\sqrt{1.25} - 0.5)$, which does not hold for sufficiently small α_1, α_2. □

B.1 Proof of Lemma B.3

Proof. Since $r^{(ag)}(1) = r^{(ag)}(3)$, agents are indifferent between receiving actions 1 and 3, and so $\boldsymbol{\beta}_{13}$ is the optimal decision boundary between the $a = 1$ and $a = 3$ best-response regions. Due to the structure of the distribution over agents, it is w.l.o.g. to consider linear decision boundaries which are parallel to $\boldsymbol{\beta}_{12}$ and $\boldsymbol{\beta}_{23}$.

Since $r(a_2) > r(a_1)$, an agent with private type $\mathbf{x}_t \in \mathcal{X}_2$ has incentive to choose their context such that $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t' \rangle + \varepsilon_{12} < 0$, where $\mathbf{x}_t'$ is agent t's context. It is w.l.o.g. to assume the agent modifies their private type in the direction perpendicular to $\boldsymbol{\beta}_{12}$. The most the agent can modify in this direction is $\frac{-\delta \boldsymbol{\beta}_{12}}{\|\boldsymbol{\beta}_{12}\|_2}$. Therefore if we want $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t' \rangle + \varepsilon_{12} \geq 0$,

$$\begin{aligned} \langle \boldsymbol{\beta}_{12}, \mathbf{x}_t - \frac{\delta \boldsymbol{\beta}_{12}}{\|\boldsymbol{\beta}_{12}\|_2} \rangle + \varepsilon_{12} &\geq 0 \\ \langle \boldsymbol{\beta}_{12}, \mathbf{x}_t \rangle - \delta \|\boldsymbol{\beta}_{12}\|_2 + \varepsilon_{12} &\geq 0 \\ \alpha_2 - \delta \|\boldsymbol{\beta}_{12}\|_2 + \varepsilon_{12} &\geq 0 \end{aligned}$$

Similarly for an agent $\mathbf{x}_t \in \mathcal{X}_3$,

$$\begin{aligned} \langle \boldsymbol{\beta}_{23}, \mathbf{x}_t + \frac{\delta \boldsymbol{\beta}_{23}}{\|\boldsymbol{\beta}_{23}\|_2} \rangle - \varepsilon_{23} &< 0 \\ \langle \boldsymbol{\beta}_{23}, \mathbf{x}_t \rangle + \delta \|\boldsymbol{\beta}_{23}\|_2 - \varepsilon_{23}' &< 0 \\ -\alpha_2 + \delta \|\boldsymbol{\beta}_{23}\|_2 - \varepsilon_{23} &< 0 \end{aligned}$$

□

B.2 Proof of Lemma B.4

Proof. It is w.l.o.g. to assume that $\mathbf{x}_t' = [0 \;\; \alpha_1 + \delta]^\top$. If we want $\langle \boldsymbol{\beta}_{12}, \mathbf{x}_t' \rangle + \varepsilon_{12} < 0$, then

$$\boldsymbol{\beta}_{12}[2](\alpha_1 + \delta) + \varepsilon_{12} < 0$$

Similarly, if we want $\langle \boldsymbol{\beta}_{23}, \mathbf{x}_t' \rangle - \varepsilon_{23} \geq 0$, then

$$\boldsymbol{\beta}_{23}[2](\alpha_1 + \delta) - \varepsilon_{23} \geq 0$$

□

B.3 Proof of sublinear strategic regret for two actions

ALGORITHM 1: Explore-first for online strategic classification with bandit feedback (`Explore-First`)

for $t = 1, \ldots, T_0$ **do**
  if $t \leq T_0/2$ **then**
    Assign action a_1
  else
    Assign action a_2
  end
end
Estimate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ via OLS as $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$
for $t = T_0 + 1, \ldots, T$ **do**
  Assign action

$$a_t = \begin{cases} a_1 & \text{if } \langle \hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2, \mathbf{x}_t \rangle - \delta \|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2\|_2 \geq 0 \\ a_2 & \text{o.w.} \end{cases}$$

end

Lemma B.5. *If* $T_0 \geq \frac{4dL}{\sigma_x^2 \log \frac{1}{2}e} \log \frac{2d}{\delta}$,

$$|\langle \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1, \mathbf{x} \rangle| \leq \frac{4L\sigma_\varepsilon \sqrt{d \log(2d/\delta)}}{\sqrt{T_0}\sigma_x^2}$$

with probability at least $1 - \delta$*. An identical bound holds for* $|\langle \hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2, \mathbf{x} \rangle|$.

Proof.

$$\begin{aligned}
\left| \left\langle \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1, \mathbf{x} \right\rangle \right| &\leq \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 \|\mathbf{x}\|_2 \\
&\leq L\sqrt{d} \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 \\
&= L\sqrt{d} \left\| \left(\sum_{t=1}^{T_0/2} \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} \sum_{t=1}^{T_0/2} \mathbf{x}_t (\mathbf{x}_t^\top \boldsymbol{\theta}_1 + \varepsilon_t) - \boldsymbol{\theta}_1 \right\|_2 \\
&= L\sqrt{d} \left\| \left(\sum_{t=1}^{T_0/2} \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} \sum_{t=1}^{T_0/2} \mathbf{x}_t \varepsilon_t \right\|_2 \\
&\leq L\sqrt{d} \left\| \left(\sum_{t=1}^{T_0/2} \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} \right\|_2 \left\| \sum_{t=1}^{T_0/2} \mathbf{x}_t \varepsilon_t \right\|_2 \\
&= \frac{L\sqrt{d} \left\| \sum_{t=1}^{T_0/2} \mathbf{x}_t \varepsilon_t \right\|_2}{\sigma_{min}(\sum_{t=1}^{T_0/2} \mathbf{x}_t \mathbf{x}_t^\top)} \\
&= \frac{L\sqrt{d} \left\| \sum_{t=1}^{T_0/2} \mathbf{x}_t \varepsilon_t \right\|_2}{\lambda_{min}(\sum_{t=1}^{T_0/2} \mathbf{x}_t \mathbf{x}_t^\top)}
\end{aligned}$$

Bounding $\left\| \sum_{t=1}^{T_0/2} \mathbf{x}_t \varepsilon_t \right\|_2$

Since $\mathbf{x}_t(j) \in [-L, L]$ and ε_t is a sub-Guassian random variable with variance σ_ε^2, $\mathbf{x}_t(j)\varepsilon_t$ is a zero-mean sub-Gaussian random variable with variance at most $L^2\sigma_\varepsilon^2$. Therefore, we can use the following result to bound $\left\|\sum_{t=1}^{T_0/2} \mathbf{x}_t\varepsilon_t\right\|_2$ with high probability.

Theorem B.6 (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 $1 - \delta$,*

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

Using Theorem B.6,

$$\begin{aligned}
\left\|\sum_{t=1}^{T_0/2} \mathbf{x}_t\varepsilon_t\right\|_2 &= \sqrt{\sum_{j=1}^{d}\left(\sum_{t=1}^{T_0/2} \mathbf{x}_t(j)\varepsilon_t\right)^2} \\
&\leq \sqrt{\sum_{j=1}^{d}\left(L\sigma_\varepsilon\sqrt{T_0\log(1/\delta_j)}\right)^2} \\
&= L\sigma_\varepsilon\sqrt{\sum_{j=1}^{d} T_0\log(1/\delta_j)} \\
&= L\sigma_\varepsilon\sqrt{dT_0\log(d/\delta)},
\end{aligned}$$

with probability at least $1 - \delta$, where the last line follows from a union bound where $\delta_j = \delta/d$ for all $j \in [d]$.

Bounding $\lambda_{min}\left(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top\right)$

Let $Y = \sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top$.

$$\begin{aligned}
\mu_{min} &= \lambda_{min}(\mathbb{E}[Y]) \\
&= \lambda_{min}(\mathbb{E}[\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top]) \\
&= \lambda_{min}(\frac{1}{2}T_0\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^\top]) \\
&= \frac{1}{2}T_0\lambda_{min}(\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^\top]) \\
&= \frac{1}{2}T_0(\lambda_{min}(\sigma_x^2\mathbb{I}_{d\times d} + \mathbb{E}[\mathbf{x}_1]\mathbb{E}[\mathbf{x}_1^\top]))
\end{aligned}$$

$\lambda_{min}(\sigma_x^2\mathbb{I}_{d\times d})$ and $\lambda_{min}(\mathbb{E}[\mathbf{x}_1]\mathbb{E}[\mathbf{x}_1^\top])$ commute, so

$$\begin{aligned}
\mu_{min} &= \frac{1}{2}T_0(\lambda_{min}(\sigma_x^2\mathbb{I}_{d\times d}) + \lambda_{min}(\mathbb{E}[\mathbf{x}_1]\mathbb{E}[\mathbf{x}_1^\top])) \\
&= \frac{1}{2}T_0\lambda_{min}(\sigma_x^2\mathbb{I}_{d\times d}) \\
&= \frac{1}{2}T_0\sigma_x^2\lambda_{min}(\mathbb{I}_{d\times d}) \\
&= \frac{1}{2}T_0\sigma_x^2
\end{aligned}$$

We use the following result to bound $\lambda_{min}\left(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top\right)$ with high probability:

Theorem B.7 (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,

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

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

$\lambda_{max}(\mathbf{x}_t\mathbf{x}_t^\top) \leq dL$, so $L' = dL$. Pick $\varepsilon = \frac{1}{2}$. Applying Theorem B.7 to $\lambda_{min}(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top)$,

$$\mathbb{P}(\lambda_{min}(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top) \leq \frac{1}{4}T_0\sigma_x^2) \leq d\left(\frac{1}{2}e\right)^{-\frac{T_0\sigma_x^2}{4dL}} \tag{1}$$

After inverting the bound, we see that $\lambda_{min}(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top) \geq \frac{1}{4}T_0\sigma_x^2$ with probability at least $1-\delta$ if $T_0 \geq \frac{4dL}{\sigma_x^2 \log \frac{1}{2}e} \log \frac{d}{\delta}$.

Combining the bounds for $\left\|\sum_{t=1}^{T_0/2} \mathbf{x}_t\varepsilon_t\right\|_2$ and $\lambda_{min}\left(\sum_{t=1}^{T_0/2} \mathbf{x}_t\mathbf{x}_t^\top\right)$ via a union bound, we obtain our stated result.

□

Theorem B.8. *If $T_0 \geq \frac{4dL}{\sigma_x^2 \log \frac{1}{2}e} \log(2dT^4)$, Algorithm `Explore-First` achieves expected strategic regret*

$$\mathbb{E}[R(T)] \leq T_0 + T\frac{16L\sigma_\varepsilon\sqrt{d\log(4dT^4)}}{\sqrt{T_0}\sigma_x^2}$$

Proof.

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{t=1}^{T} \mathbb{E}[r(\pi_S(\mathbf{x}_t))|\mathbf{x}_t] - \sum_{t=1}^{T} \mathbb{E}[r(\pi_t(\mathbf{x}_t'))|\mathbf{x}_t'] \\
&\leq T_0 + \sum_{t=T_0+1}^{T} \mathbb{E}[r(\pi_S(\mathbf{x}_t))|\mathbf{x}_t] - \sum_{t=1}^{T} \mathbb{E}[r(\pi_t(\mathbf{x}_t'))|\mathbf{x}_t'] \\
&= T_0 + \sum_{t=T_0+1}^{T} \langle \boldsymbol{\theta}_{a_t^*} - \boldsymbol{\theta}_{a_t}, \mathbf{x}_t\rangle
\end{aligned}$$

If $T_0 \geq \frac{8dL}{\sigma_x^2 \log \frac{1}{2}e} \log \frac{2d}{\delta}$, the following two events hold simultaneously with probability at least $1-\delta$:

$$\begin{aligned}
|\langle \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1, \mathbf{x}\rangle| &\leq \frac{4L\sigma_\varepsilon\sqrt{d\log(4d/\delta)}}{\sqrt{T_0}\sigma_x^2} \\
|\langle \hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2, \mathbf{x}\rangle| &\leq \frac{4L\sigma_\varepsilon\sqrt{d\log(4d/\delta)}}{\sqrt{T_0}\sigma_x^2}
\end{aligned}$$

If a_t is the suboptimal action to assign to agent t, then with probability at least $1-\delta$,

$$\begin{aligned}
&\langle \boldsymbol{\theta}_{a_t}, \mathbf{x}_t\rangle + \frac{4L\sigma_\varepsilon\sqrt{d\log(4d/\delta)}}{\sqrt{T_0}\sigma_x^2} \geq \left\langle \hat{\boldsymbol{\theta}}_{a_t}, \mathbf{x}_t\right\rangle \geq \left\langle \hat{\boldsymbol{\theta}}_{a_t^*}, \mathbf{x}_t\right\rangle \\
&\geq \left\langle \boldsymbol{\theta}_{a_t^*}, \mathbf{x}_t\right\rangle - \frac{4L\sigma_\varepsilon\sqrt{d\log(4d/\delta)}}{\sqrt{T_0}\sigma_x^2} \geq \left\langle \hat{\boldsymbol{\theta}}_{a_t^*}, \mathbf{x}_t\right\rangle
\end{aligned} \tag{2}$$

Therefore, $\left\langle \boldsymbol{\theta}_{a_t^*} - \boldsymbol{\theta}_{a_t}, \mathbf{x}_t \right\rangle \leq \frac{8L\sigma_\varepsilon\sqrt{d\log(4d/\delta)}}{\sqrt{T_0}\sigma_x^2}$ with probability at least $1-\delta$.

By an argument similar to the proof of Proposition 3.2, the `Explore-First`'s strategic regret is the same as that of making decisions according to $\hat{\boldsymbol{\beta}}_{12}$ on non-strategic agents. Using this observation, we can set $\delta = \frac{1}{T^4}$ and upper-bound expected strategic regret by

$$\mathbb{E}[R(T)] \leq T_0 + T\frac{16L\sigma_\varepsilon\sqrt{d\log(4dT^4)}}{\sqrt{T_0}\sigma_x^2}$$

□

Corollary B.9. *If* $T_0 = \frac{4L^{2/3}\sigma_\varepsilon^{2/3}(d\log(4dT^4))^{1/3}}{\sigma_x^{4/3}}T^{2/3}$ *and* $T \geq \frac{2L^{1/2}d}{\sigma_\varepsilon\sigma_x(\log\frac{1}{2}e)^{3/2}}\log\frac{16L^{1/2}d^2}{\sigma_\varepsilon\sigma_x(\log\frac{1}{2}e)^{3/2}}$, *then*

$$\mathbb{E}[R(T)] \leq \frac{16L^{2/3}\sigma_\varepsilon^{2/3}(d\log(4dT^4))^{1/3}}{\sigma_x^{4/3}}T^{2/3}$$

The proof of Corollary B.9 follows straightforwardly from Theorem B.8 and the inequality $\log x \leq \alpha x - \log\alpha - 1$ for any $\alpha, x > 0$.