Stackelberg Games with Side Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study an online learning setting in which a *leader* interacts with a sequence of *followers* over the course of T rounds. At each round, the leader commits to a mixed strategy over actions, after which the follower best-responds. Such settings are referred to in the literature as *Stackelberg games*. Stackelberg games have received much interest from the community, in part due to their applicability to real-world security settings such as wildlife preservation and airport security. However despite this recent interest, current models of Stackelberg games fail to take into consideration the fact that the players' optimal strategies often depend on *external factors* such as weather patterns, airport traffic, etc. We address this gap by allowing for player payoffs to depend on an external *context*, in addition to the actions taken by each player. We formalize this setting as a repeated Stackelberg game with *side information* and show that under this setting, it is impossible to achieve sublinear regret if both the sequence of contexts *and* the sequence of followers is chosen adversarially. Motivated by this impossibility result, we consider two natural relaxations: (1) stochastically chosen contexts with adversarially chosen followers and (2) stochastically chosen followers with adversarially chosen contexts. In both of these settings, we provide simple algorithms which obtain no-regret guarantees.

1 Introduction

A *Stackelberg game* [22, 7] is a strategic interaction between two utility-maximizing players in which one player (the *leader*) is able to *commit* to a strategy before the other player (the *follower*) takes an action. While Stackelberg's original formulation was used to model economic competition between firms, Stackelberg games have been used to study a wide range of topics ranging from incentives in algorithmic decision-making [12] to radio spectrum utilization [23]. Perhaps the most successful application of Stackelberg games to solve real-world problems is in the domain of security, where the analysis of *Stackelberg security games* has led to new methods in domains such as passenger screening at airports [6], wildlife protection efforts in conservation areas [8], the deployment of US Federal Air Marshals on board commercial flights [15], and patrol boat schedules for the US Coast Guard [1]. However in many real-world (security) settings, the payoffs of the players often depend on additional *contextual information* which is not captured by the Stackelberg (security) game framework. For example, in airport security the severity of an attack depends on factors such as the arrival and departure city of a flight, the number of passengers on board, and the amount of valuable cargo on the aircraft. Additionally, there may be information in the time leading up to the attack attempt which may help the security service determine the type of attack which is coming [14]. In wildlife protection settings, factors such as the weather or time of year may make certain species of wildlife easier or harder to defend from poaching, and information such as the location of tire tracks may provide context about which animals are being targeted.

In order to capture this additional information that the leader may have at their disposal, we formalize such settings as *Stackelberg games with side information*. Specifically, we consider a setting in which

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

a leader interacts with a sequence of followers in an online setting. At each time-step, the leader gets to see payoff-relevant information about the current round in the form of a *context*. After observing the context, the leader commits to a mixed strategy, and the follower best-responds in order to maximize their utility. While we show that it is impossible for the leader to achieve good performance (measured through *regret*) whenever the sequence of followers and side information are chosen by an *adversary*, we show that effective learning is possible whenever the power of the adversary is restricted.

2 Setting and background

Notation We use $[N] := \{1, \ldots, N\}$ to denote the set of natural numbers up to and including $N \in \mathbb{N}$ and $\text{cl}(\mathcal{P})$ to denote the closure of the set $\mathcal{P}$. $\mathbf{x}[a]$ denotes the a-th component of vector $\mathbf{x}$, and $\Delta(\mathcal{A})$ denotes the probability simplex over the set $\mathcal{A}$. Finally, while we present our results for general Stackelberg games with side information, our results are readily applicable to the special case of Stackelberg *security* games with side information. See Appendix A for a discussion on related work.

We consider a repeated Stackelberg game between a leader and a sequence of followers. At each time-step $t \in [T]$, the leader moves first by playing some mixed strategy $\mathbf{x}_t$ over a set of (finite) leader actions $\mathcal{A}_l$, i.e., $\mathbf{x}_t \in \Delta(\mathcal{A}_l)$. Having observed the leader's mixed strategy, the follower *best-responds* by playing some action $a_f \in \mathcal{A}_f$, where $\mathcal{A}_f$ is the (finite) set of follower actions. We assume that each follower is one of K *follower types* $\{\alpha_1, \ldots, \alpha_K\}$. Each follower type α_i is characterized by a payoff matrix $M_{\alpha_i} \in \mathbb{R}^{|\mathcal{A}_l| \times |\mathcal{A}_f|}$, i.e. given a leader action a_l and follower action a_f, a follower of type α_i would receive utility $M_{\alpha_i}[a_l, a_f]$. We assume that followers are perfectly rational and pick their action in order to maximize their utility in expectation over the randomness in the leader's mixed strategy, i.e., follower f_t's *best-response* to leader mixed strategy $\mathbf{x}_t$ is

$$b_{f_t}(\mathbf{x}_t) \in \arg\max_{a_f \in \mathcal{A}_f} \sum_{a_l \in \mathcal{A}_l} \mathbf{x}[a_l] \cdot M_{f_t}[a_l, a_f].$$

We assume that the set of all possible follower types is known to the leader, but that the follower's type at round t is not known to the leader until *after* round t is over.

At each time $t \in [T]$, nature selects a *context* $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ and reveals it to the leader. In line with the literature on linear contextual bandits, we assume that there is an (unknown) linear mapping from contexts and joint actions to expected leader utility $u_l : \mathcal{Z} \times \mathcal{A}_l \times \mathcal{A}_f \to \mathbb{R}$, given by $u_l(\mathbf{z}, a_l, a_f) = \langle \mathbf{z}, \boldsymbol{\theta}(a_l, a_f) \rangle$ for some $\boldsymbol{\theta}(a_l, a_f) = \boldsymbol{\theta}^{(a_l, a_f)} \in \mathbb{R}^d$ that is known to the leader. We assume that $u_l(\mathbf{z}, a_l, a_f) \in [-1, 1]$ for all $z \in \mathcal{Z}$, $a_l \in \mathcal{A}_l$, and $a_f \in \mathcal{A}_f$. We use the shorthand

$$u_l(\mathbf{z}, \mathbf{x}, b_f(\mathbf{z})) = \sum_{a_l \in \mathcal{A}_l} x[a_l] \cdot u_l(\mathbf{z}, a_l, b_f(\mathbf{z}))$$

to denote the leader's expected utility of playing mixed strategy $\mathbf{x}$ under context $\mathbf{z}$ against follower f. Given context $\mathbf{z}_t \in \mathcal{Z}$, the leader plays mixed strategy $\mathbf{x}_t$ and the follower best-responds by playing $b_{f_t}(\mathbf{x}_t)$. After each round, the leader receives noisy utility $u_{l,t}(\mathbf{z}_t, a_{l,t}, b_{f_t}(\mathbf{x}_t)) = u_l(\mathbf{z}_t, a_{l,t}, b_{f_t}(\mathbf{x}_t)) + \varepsilon_t$, where $a_{l,t} \sim \mathbf{x}_t$ and $\varepsilon_t \in \mathbb{R}$ is zero-mean sub-Gaussian random noise with variance η^2, and observes the follower type f_t. We measure the leader's performance via the notion of *contextual Stackelberg regret*.

Definition 2.1 (Contextual Stackleberg Regret, I)**.** *Given a sequence of followers $f_1, \ldots, f_T$ and a sequence of contexts $\mathbf{z}_1, \ldots, \mathbf{z}_T$, the leader's contextual Stackelberg regret is*

$$R(T) := \sum_{t=1}^{T} u_l(\mathbf{z}_t, \pi^*(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t))) - u_l(\mathbf{z}_t, \mathbf{x}_t, b_{f_t}(\mathbf{x}_t)),$$

where $\pi^* : \mathcal{Z} \to \Delta(\mathcal{A}_l)$ is the optimal policy, given knowledge of $f_1, \ldots, f_T$ and $\boldsymbol{\theta}(a_f, a_l)$ for all $a_f \in \mathcal{A}_f$ and $a_l \in \mathcal{A}_l$.

If an algorithm achieves regret $R(T) = o(T)$ (i.e. regret grows *sublinearly* with T), we say that it is a *no-regret* algorithm.

3 Online learning contextual Stackelberg games

In Section 3.1, we show that it is impossible for the leader to obtain sublinear regret when both the sequence of followers and contexts are chosen *adversarially*. Motivated by this observation, we consider two relaxations of this setting: one in which the sequence of followers are chosen *stochastically* (Section 3.2), and one in which the contexts are chosen stochastically (Section 3.3).

3.1 Impossibility result

We proceed via a reduction to the online linear thresholding problem, for which it is known that no algorithm can obtain no regret. In particular, we show that if there exists a no-regret algorithm for the contextual Stackelberg game problem, then it could be used to construct a no-regret algorithm for the online linear thresholding problem, which is a contradiction.

Online linear thresholding problem The online linear thresholding problem is as follows: At $t = 0$, an adversary chooses a *cutoff* $s \in [0, 1]$ and a sequence of points $\omega_1, \ldots, \omega_T \in [0, 1]$, possibly using knowledge of the learner's algorithm. A point ω_t is assigned label $y_t = 1$ if $\omega_t > s$. Otherwise the label is $y_t = -1$. For $t = 1, \ldots, T$, the learner receives the point $\omega_t \in [0, 1]$ and makes a *guess* $\hat{y}_t \in \{-1, 1\}$. We allow the learner to randomize by playing a mixed strategy $\mathbf{x}_t$ at time t, where $\mathbf{x}_t := [\mathbb{P}(\hat{y}_t = 1)\ \mathbb{P}(\hat{y}_t = 0)]^\top$. Note that the learner's optimal policy is $\pi^*_{\texttt{OLT}}(w_t) = [1\ 0]^\top$ if $w_t > s$ and $\pi^*_{\texttt{OLT}}(w_t) = [0\ 1]^\top$ if $w_t \leq s$, which achieves perfect classification on any point $w \in \mathbb{R}$. We make use of the following well-known impossibility result (see e.g. [9]).

Lemma 3.1. *Any algorithm suffers regret $R_{\texttt{OLT}}(T) = \Omega(T)$ in the online linear thresholding problem (where the expectation is taken over the algorithm's internal randomness) when $(s, \{w_t\}_{t=1}^T)$ are chosen by an adversary.*

We are now ready to state our impossibility result for learning in contextual Stackelberg games with adversarially-chosen contexts and followers.

Theorem 3.2. *If an adversary can choose both the sequence of contexts $\mathbf{z}_1, \ldots, \mathbf{z}_T$ and the sequence of followers $f_1, \ldots, f_T$, no algorithm can achieve better than $\Omega(T)$ contextual Stackelberg regret in expectation over the internal randomness of the algorithm.*

Proof Sketch. See Appendix B for the full proof. At a high level, the reduction to online linear thresholding proceeds by creating an instance of the contextual Stackelberg game problem such that the sequence of contexts $\mathbf{z}_1, \ldots, \mathbf{z}_T$ (roughly) correspond to the sequence of points $\omega_1, \ldots, \omega_T$ encountered, and the sequence of follower types $f_1, \ldots, f_T$ correspond to the sequence of labels $y_1, \ldots, y_T$. We then show that a no-regret algorithm in the online thresholding problem can be obtained by running an algorithm which minimizes contextual Stackelberg regret on the constructed contextual Stackelberg game instance. However this is a contradiction, since by Lemma 3.1 the online thresholding problem is not online learnable by any algorithm.

3.2 Stochastic follower types

In this setting we allow the sequence of contexts to be chosen by an adversary, but we restrict the sequence of followers to be drawn i.i.d. from some (unknown) distribution over follower types $\mathcal{F}$. We allow the adversary to have knowledge of $\mathcal{F}$, but not the realized draws $f_1, \ldots, f_T$, when picking the sequence of contexts. Under this relaxation, our measure of algorithm performance is expected contextual Stackelberg regret, where the expectation is now also taken over the randomness in the follower type distribution.

Definition 3.3 (Contextual Stackleberg Regret, II)**.** *Given a population of followers $\mathcal{F}$ and a sequence of contexts $\mathbf{z}_1, \ldots, \mathbf{z}_T$, the leader's expected contextual Stackelberg regret is*

$$\mathbb{E}[R(T)] := \mathbb{E}_{f_1, \ldots f_T \sim \mathcal{F}} \left[\sum_{t=1}^{T} u_l(\mathbf{z}_t, \pi^*(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t))) - u_l(\mathbf{z}_t, \mathbf{x}_t, b_{f_t}(\mathbf{x}_t)) \right]$$

where $\pi^* : \mathcal{Z} \rightarrow \Delta(\mathcal{A}_l)$ is the optimal policy given knowledge of $\mathcal{F}$ and $\boldsymbol{\theta}(a_f, a_l)$, $\forall a_f \in \mathcal{A}_f, a_l \in \mathcal{A}_l$.

As we show in Appendix C, a relatively simple closed-form characterization of the leader's optimal policy exists if the distribution $\mathcal{F}$ is known. When $\mathcal{F}$ is unknown, we show that the leader can obtain $\tilde{O}(\sqrt{T})$ regret by estimating the distribution over follower types in an online fashion, and acting optimally w.r.t. their estimate.

ALGORITHM 1: Learning in contextual Stackelberg games with stochastic follower types.

Set $\widehat{\mathbb{P}}_1(f = \alpha_i) = \frac{1}{K}, \forall i \in [K]$
for $t = 1, \ldots, T$ **do**
  Observe context $\mathbf{z}_t$, commit to mixed strategy
   $\mathbf{x}_t = \pi_t(\mathbf{z}_t) = \arg\max_{\mathbf{x} \in \mathcal{E}} \sum_{i=1}^{K} \widehat{\mathbb{P}}_t(f = \alpha_i) u_l(\mathbf{z}_t, \mathbf{x}, b_{\alpha_i}(\mathbf{x}))$.
  Receive utility $u_l(\mathbf{z}_t, a_{l,t}, b_{f_t}(\mathbf{x}_t))$, where $a_{l,t} \sim \mathbf{x}_t$, and observe follower type f_t.
  Set $\widehat{\mathbb{P}}_{t+1}(f = \alpha_i) = \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}\{f_s = \alpha_i\}$.
end

ALGORITHM 2: Learning in contextual Stackelberg games with stochastic contexts.

Consider $\Pi := \{\pi^{(\boldsymbol{\omega})}\}_{\boldsymbol{\omega} \in \Omega}$
Let $\mathbf{q}_1[\pi^{(\boldsymbol{\omega})}] := 1$, $\mathbf{p}_1[\pi^{(\boldsymbol{\omega})}] := \frac{1}{|\Pi|}$ for all $\pi^{(\boldsymbol{\omega})} \in \Pi$
for $t = 1, \ldots, T$ **do**
  Sample $\pi_t \sim \mathbf{p}_t$, $a_{l,t} \sim \pi_t(\mathbf{z}_t)$.
  Receive utility $u_{l,t}(\mathbf{z}_t, a_{l,t}, b_{f_t}(\pi_t(\mathbf{z}_t)))$ and observe follower type f_t.
  For each policy $\pi^{(\boldsymbol{\omega})} \in \Pi$, compute $\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}] := -u_l(\mathbf{z}_t, \pi^{(\boldsymbol{\omega})}(\mathbf{z}_t), b_{f_t}(\pi^{(\boldsymbol{\omega})}(\mathbf{z}_t)))$.
  Set $\mathbf{q}_{t+1}[\pi^{(\boldsymbol{\omega})}] = \exp\left(-\eta \sum_{s=1}^{t} \boldsymbol{\ell}_s[\pi^{(\boldsymbol{\omega})}]\right)$, $\mathbf{p}_{t+1}[\pi^{(\boldsymbol{\omega})}] = \mathbf{q}_{t+1}[\pi^{(\boldsymbol{\omega})}] / \sum_{\pi^{(\boldsymbol{\omega}')} \in \Pi} \mathbf{q}_{t+1}[\pi^{(\boldsymbol{\omega}')}]$.
end

Theorem 3.4. *Algorithm 1 obtains expected contextual Stackelberg regret* $\mathbb{E}[R(T)] \leq \mathcal{O}(\sqrt{K^2 T \log(T)})$*, where the set* $\mathcal{E}$ *is defined as in Lemma C.3 and the expectation is taken over both the follower population and the randomness in the leader's mixed strategies as in Definition 3.3.*

Proof Sketch. See Appendix C for the full proof. At a high level, our results in this section rely on showing that the leader can learn an accurate estimate of $\mathcal{F}$ sufficiently quickly, and generalizing some of the results and ideas from Balcan et al. [4] to incorporate side information. (See Appendix A for more details on how our work is related to theirs.)

3.3 Stochastic contexts

We now consider a setting in which the sequence of contexts are drawn i.i.d. from some unknown distribution $\mathcal{P}$ and the sequence of followers is chosen by an adversary with knowledge of the leader's algorithm and of $\mathcal{P}$ (but not $\mathbf{z}_1, \ldots, \mathbf{z}_T$). As was the case in the previous section, we update our definition of contextual Stackelberg regret in order to reflect the additional stochasticity under this setting.

Definition 3.5 (Contextual Stackleberg Regret, III)**.** *Given a sequence of followers* $f_1, \ldots, f_T$ *and a distribution over contexts* $\mathcal{P}$*, the leader's expected contextual Stackelberg regret is*

$$\mathbb{E}[R(T)] := \mathbb{E}_{\mathbf{z}_1, \ldots \mathbf{z}_T \sim \mathcal{P}} \left[\sum_{t=1}^{T} u_l(\mathbf{z}_t, \pi^*(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t))) - u_l(\mathbf{z}_t, \mathbf{x}_t, b_{f_t}(\mathbf{x}_t)) \right]$$

where $\pi^* : \mathcal{Z} \to \Delta(\mathcal{A}_l)$ *is the optimal-in-hindsight policy.*

Our key insight in this section is that when the sequence of contexts is generated stochastically, it suffices to consider only a finite number of policies in order to find one which is optimal. Therefore, the leader can play an off-the-shelf online learning algorithm (e.g. Hedge) over this finite set of policies to achieve sublinear regret. This intuition is formalized in Algorithm 2 and the following theorem.

Theorem 3.6. *Algorithm 2 obtains expected contextual Stackelberg regret* $\mathbb{E}[R(T)] \leq \mathcal{O}(\sqrt{TK \log(T)})$ *when* $\eta = \sqrt{\frac{\log |\Pi|}{T}}$*, where* $\Omega := \{\boldsymbol{\omega} : \boldsymbol{\omega} \in \Delta^K, T \cdot \boldsymbol{\omega}[i] \in \mathbb{N}, \forall i \in [K]\}$ *and the expectation is taken over both the distribution over contexts and the randomness in the leader's mixed strategies as in Definition 3.5.*

Proof Sketch. See Appendix D for the full proof. The first part of the analysis leverages the geometry of the leader's optimal policy to show that it suffices to consider a discrete set of policies. The second part of the analysis follows the standard analysis of Hedge.

References

[1] Bo An, Fernando Ordóñez, Milind Tambe, Eric Shieh, Rong Yang, Craig Baldwin, Joseph DiRenzo III, Kathryn Moretti, Ben Maule, and Garrett Meyer. A deployed quantal response-based patrol planning system for the us coast guard. *Interfaces*, 43(5):400–420, 2013.

[2] Bo An, Milind Tambe, and Arunesh Sinha. Stackelberg security games (ssg) basics and application overview. *Improving Homeland Security Decisions*, page 485, 2017.

[3] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On the convergence of no-regret learning dynamics in time-varying games. *arXiv preprint arXiv:2301.11241*, 2023.

[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] Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. 2014.

[6] Matthew Brown, Arunesh Sinha, Aaron Schlenker, and Milind Tambe. One size does not fit all: A game-theoretic approach for dynamically and effectively screening for threats. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[7] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

[8] Fei Fang, Peter Stone, and Milind Tambe. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *IJCAI*, pages 2589–2595, 2015.

[9] Nika Haghtalab. Lecture 12: Introduction to online learning 2. *CS6781: Theoretical Foundations of Machine Learning course notes*, 2020.

[10] Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 917–918, 2022.

[11] Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *arXiv preprint arXiv:2306.02704*, 2023.

[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, Ioannis Anagnostides, Gabriele Farina, Mikhail Khodak, Zhiwei Steven Wu, and Tuomas Sandholm. Meta-learning in games. *arXiv preprint arXiv:2209.14110*, 2022.

[14] Indiana Intelligence Fusion Center Iifc. 8 signs of terrorism, Jul 2022. URL `https://www.in.gov/iifc/8-signs-of-terrorism/`.

[15] Manish Jain, Jason Tsai, James Pita, Christopher Kiekintveld, Shyamsunder Rathi, Milind Tambe, and Fernando Ordónez. Software assistants for randomized patrol planning for the lax airport police and the federal air marshal service. *Interfaces*, 40(4):267–290, 2010.

[16] Debarun Kar, Thanh H Nguyen, Fei Fang, Matthew Brown, Arunesh Sinha, Milind Tambe, and Albert Xin Jiang. Trends and applications in stackelberg security games. *Handbook of dynamic game theory*, pages 1–47, 2017.

[17] Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pages 250–262. Springer, 2009.

[18] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2149–2156, 2019.

[19] Pier Giuseppe Sessa, Ilija Bogunovic, Andreas Krause, and Maryam Kamgarpour. Contextual games: Multi-agent learning with side information. *Advances in Neural Information Processing Systems*, 33:21912–21922, 2020.

[20] Arunesh Sinha, Debarun Kar, and Milind Tambe. Learning adversary behavior in security games: A pac model perspective. *arXiv preprint arXiv:1511.00043*, 2015.

[21] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security games: Looking beyond a decade of success. IJCAI, 2018.

[22] Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[23] Jin Zhang and Qian Zhang. Stackelberg game for utility-based cooperative cognitiveradio networks. In *Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, pages 23–32, 2009.

[24] Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning*, pages 26772–26808. PMLR, 2022.

A Related work

See [21, 16, 2] for an overview of the literature on applications of Stackelberg security games. From a technical point of view, our results are most related to two lines of work: learning in Stackelberg games, and dealing with various forms of additional information in repeated game settings.

Learning in Stackelberg games Conitzer and Sandholm [7] were the first to provide algorithms for recovering the leader's optimal mixed strategy in Stackelberg games, when the follower's payoff matrix is *known* to the leader. Letchford et al. [17] was the first to consider learning the leader's optimal mixed strategy in the repeated Stackelberg game setting against a perfectly rational follower. Peng et al. [18] study the same setting, providing improved rates and formally showing that the problem is NP-Hard. Importantly, both of these papers impose a "minimum volume constraint" on the leader's mixed strategy space with respect to each of the follower's pure strategies, meaning they only consider a subset of all possible Stackelberg games. Learning algorithms to recover the leader's optimal mixed strategy have also been studied in Stackelberg security games [4, 5, 20]. In particular, our work builds off of several results established for (non-contextual) Stackelberg games in Balcan et al. [4]. More recent work on learning in Stackelberg games considers the effects of non-myopic followers [11] and calibration [10].

Additional information in repeated games Sessa et al. [19] study a repeated game setting in which the players receive additional information (or *context*) at each round, much like in our setting. However, their focus is on repeated normal-form games, which require different tools and techniques to solve than the Stackelberg game setting we consider. Other work has also considered repeated games which change over time in different ways. In particular, [13] study a meta-learning setting in which the game being played changes after a fixed number of rounds, and [24, 3] study learning dynamics in time-varying game settings.

B Appendix for Section 3.1: Impossibility result

Theorem B.1. *If an adversary can choose both the sequence of contexts $\mathbf{z}_1, \ldots, \mathbf{z}_T$ and the sequence of followers $f_1, \ldots, f_T$, no algorithm can achieve better than $\Omega(T)$ contextual Stackelberg regret in expectation over the internal randomness of the algorithm.*

Proof. Let ALG denote any algorithm which achieves $o(T)$ contextual Stackelberg regret under adversarially-chosen contexts and follower types. Note that at every time-step, ALG takes as input a context $\mathbf{z}_t$ and produces a mixed strategy $\mathbf{x}_t$. We reduce to the problem of online linear thresholding.

Consider the following family of contextual Stackelberg game instances with two follower types α_1 and α_2: $\mathcal{A}_l = \mathcal{A}_f = \{a_1, a_2\}$, $\mathcal{Z} = [0,1] \times \{1\}$, and $b_{\alpha_1}(\mathbf{x}) = a_1$ and $b_{\alpha_2}(\mathbf{x}) = a_2$ for all $\mathbf{x} \in \Delta(\mathcal{A}_l)$. Furthermore, suppose that $\boldsymbol{\theta}(a_l, a_f)[j] = 0$ for all a_l, a_f and $j \leq m$, and $\boldsymbol{\theta}(a_l, a_f)[m+1] = \mathbb{1}\{a_l = a_f\}$ for all a_l, a_f. Since each follower type's best-response does not depend on the mixed strategy played by the leader, we use the shorthand $b_{f_t} := b_{f_t}(\mathbf{x})$.

The reduction proceeds as follows: given input $\mathbf{w}_t \in \mathbb{R}^m$, we give the context $\mathbf{z}_t := [\mathbf{w}_t, 1]^\top$ as input to ALG and receive mixed strategy $\mathbf{x}_t \in \mathbb{R}^2$. We sample $a_{l,t}$ according to $\mathbf{x}_t$ and let $\hat{y}_t = 1$ if $a_{l,t} = a_1$ and $\hat{y}_t = 2$ if $a_{l,t} = a_2$. We receive utility $\mathbb{1}\{\hat{y}_t = y_t\}$ and feedback y_t from the environment. We then set $f_t = \alpha_1$ if $y_t = 1$ and $f_t = \alpha_2$ if $y_t = -1$, which determines the utility $u_l(\mathbf{z}_t, a_{l,t}, b_{f_t}(\mathbf{z}_t))$ received by ALG. Finally, we give the follower type f_t as input to ALG. Since ALG is a no-regret learning algorithm for the contextual Stackelberg game setting, by Definition 2.1 we know that

$$R(T) = \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] u_l(\mathbf{z}_t, a_l, b_{f_t}(\mathbf{z}_t)) - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] u_l(\mathbf{z}_t, a_l, b_{f_t}(\mathbf{x}_t)) = o(T).$$

Next, we show that this implies that $R_{\texttt{OLT}}(T) = o(T)$, where

$$\begin{aligned} R_{\texttt{OLT}}(T) &:= T - \sum_{t=1}^{T} \mathbb{P}(\hat{y}_t = y_t) \\ &= \sum_{t=1}^{T} \pi^*_{\texttt{OLT}}(w_t)[i_t] - \mathbb{P}(\hat{y}_t = y_t) \\ &= \sum_{t=1}^{T} \pi^*_{\texttt{OLT}}(w_t)[i_t] - \mathbf{x}_t[i_t], \end{aligned}$$

where $\mathbf{x}_t = [\mathbb{P}(\hat{y}_t = 1)\ \mathbb{P}(\hat{y}_t = 0)]^\top$ is the mixed strategy played by the learner at time t and $i_t = 1$ if $w_t > s$ and $i_t = 2$ otherwise.

$$\begin{aligned} R(T) &:= \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] u_l(\mathbf{z}_t, a_l, b_{f_t}(\mathbf{z}_t)) - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] u_l(\mathbf{z}_t, a_l, b_{f_t}(\mathbf{x}_t)) \\ &= \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] \langle \mathbf{z}_t, \boldsymbol{\theta}(a_l, b_{f_t}(\pi^*(\mathbf{z}_t))) \rangle - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] \langle \mathbf{z}_t, \boldsymbol{\theta}(a_l, b_{f_t}(\mathbf{x}_t)) \rangle \\ &= \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] \cdot \boldsymbol{\theta}(a_l, b_{f_t}(\pi^*(\mathbf{z}_t)))[m+1] - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] \cdot \boldsymbol{\theta}(a_l, b_{f_t}(\mathbf{x}_t))[m+1] \\ &= \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] \cdot \mathbb{1}\{a_l = b_{f_t}(\pi^*(\mathbf{z}_t))\} - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] \cdot \mathbb{1}\{a_l = b_{f_t}(\mathbf{x}_t)\} \\ &= \sum_{t=1}^{T} \sum_{a_l \in \mathcal{A}_l} \pi^*(\mathbf{z}_t)[a_l] \cdot \mathbb{1}\{a_l = b_{f_t}\} - \sum_{a_l \in \mathcal{A}_l} \mathbf{x}_t[a_l] \cdot \mathbb{1}\{a_l = b_{f_t}\} \\ &= \sum_{t=1}^{T} \pi^*(\mathbf{z}_t)[b_{f_t}] - \mathbf{x}_t[b_{f_t}] = \sum_{t=1}^{T} \mathbb{P}(\pi^*(\mathbf{z}_t) = a_{y_t}) - \mathbb{P}(\pi_t(\mathbf{z}_t) = a_{y_t}) \\ &= \sum_{t=1}^{T} \mathbb{P}(\pi^*(\mathbf{z}_t) = a_{y_t}) - \mathbb{P}(\hat{y}_t(\mathbf{w}_t) = y_t) = T - \sum_{t=1}^{T} \mathbb{P}(\hat{y}_t = y_t) \end{aligned}$$

where the first equality is due to the definition of leader utility, the second follows because $\boldsymbol{\theta}(a_l, a_f)[j] = 0$ for all $j \leq m$, the third is from the fact that $\boldsymbol{\theta}(a_l, a_f)[m+1] = \mathbb{1}\{a_l = a_f\}$ for all a_l, a_f, the fourth equality holds because the follower best-responses do not depend on the leader's mixed strategy, and the last equality follows from the fact that the following policy achieves perfect performance in the contextual Stackelberg game setting:

$$\pi^*(\mathbf{z}_t) = \begin{cases} a_{l,1} & \text{if } \mathbf{w}_t > s \\ a_{l,2} & \text{if } \mathbf{w}_t \leq s. \end{cases}$$

□

C Appendix for Section 3.2: Stochastic follower types

Observe that for any context $\mathbf{z}$, $\pi^*(\mathbf{z})$ takes the following closed form:

$$\begin{aligned} \pi^*(\mathbf{z}) &= \arg\max_{\mathbf{x} \in \Delta(\mathcal{A})} \mathbb{E}_{f \sim \mathcal{F}} \left[\sum_{a_l \in \mathcal{A}_l} \mathbf{x}[a_l] \cdot \langle \mathbf{z}, \boldsymbol{\theta}(a_l, b_f(\mathbf{x})) \rangle \right] \\ &= \arg\max_{\mathbf{x} \in \Delta(\mathcal{A})} \sum_{i=1}^{K} \mathbb{P}(f = \alpha_i) \sum_{a_l \in \mathcal{A}_l} \mathbf{x}[a_l] \cdot \langle \mathbf{z}, \boldsymbol{\theta}(a_l, b_{\alpha_i}(\mathbf{x})) \rangle \end{aligned}$$

The solution to the above optimization may be obtained by first solving

$$\begin{aligned}\mathbf{x}_{a_{f,1},\ldots,a_{f,K}}(\mathbf{z}) = \arg\max_{\mathbf{x}\in\Delta(\mathcal{A})} \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) \sum_{a_l\in\mathcal{A}_l} \mathbf{x}[a_l]\cdot\langle \mathbf{z}, \boldsymbol{\theta}(a_l, a_{f,i})\rangle \\ \text{s.t. } b_{\alpha_1}(\mathbf{x}) = a_{f,1}, b_{\alpha_2}(\mathbf{x}) = a_{f,2}, \ldots, b_{\alpha_K}(\mathbf{x}) = a_{f,K}\end{aligned} \tag{1}$$

for all $|\mathcal{A}_f|^K$ possible combinations of $a_{f,1},\ldots,a_{f,K}\in\mathcal{A}_f$ and then setting

$$\pi^*(\mathbf{z}) = \arg\max_{a_{f,1}\in\mathcal{A}_f,\ldots,a_{f,K}\in\mathcal{A}_f} \mathbf{x}_{a_{f,1},\ldots,a_{f,K}}(\mathbf{z}). \tag{2}$$

Borrowing notation from Balcan et al. [4], we introduce the notion of a *best-response region*.

Definition C.1 (Follower Best-Response Region)**.** *For every follower type $\alpha_i : i\in[K]$ and follower action $a_f\in\mathcal{A}_f$, let $\mathcal{P}(\alpha_i, a_f)$ denote the set of all leader mixed strategies such that a follower of type α_i best-responds by playing action a_f, i.e.,*

$$\mathcal{P}(\alpha_i, a_f) = \{\mathbf{x}\in\Delta(\mathcal{A}_l) : b_{\alpha_i}(\mathbf{x}) = a_f\}.$$

As in Balcan et al. [4], $\mathcal{P}(\alpha_i, a_f)$ is a (possibly empty) convex and bounded, but not necessarily closed, polytope for all $i\in[K]$ and $a_f\in\mathcal{A}_f$.

Definition C.2 (Best-Response Region)**.** *For a given best-response function $\sigma:\{\alpha_1,\ldots,\alpha_K\}\to\mathcal{A}_f$, let $\mathcal{P}_\sigma$ denote the set of all valid leader mixed strategies such that for all $i\in[K]$, a follower of type α_i plays action $\sigma(\alpha_i)$. In other words, $\mathcal{P}_\sigma = \cap_{i\in[K]}\mathcal{P}(\alpha_i,\sigma(\alpha_i))$.*

Note that there are at most $|\mathcal{A}_f|^K$ different best-response functions (and hence, best-response regions). As in Balcan et al. [4], we consider the set of (approximate) extreme points $\mathcal{E}$ of all best-response regions. Formally, for a given $\delta>0$, $\mathcal{E}$ is the set of leader mixed strategies such that for all σ and any $\mathbf{x}\in\Delta(\mathcal{A}_l)$ that is an extreme point of $\text{cl}(\mathcal{P}_\sigma)$, $\mathbf{x}\in\mathcal{E}$ if $\mathbf{x}\in\mathcal{P}_\sigma$, otherwise there is some $\mathbf{x}'\in\mathcal{E}$ such that $\mathbf{x}'\in\mathcal{P}_\sigma$ and $\|\mathbf{x}'-\mathbf{x}\|_1\le\delta$. The following lemma is a generalization of Lemma 4.3 in Balcan et al. [4] to the contextual Stackelberg game setting, and its proof uses similar techniques from convex analysis.

Lemma C.3. *Let $\pi^*:\mathcal{Z}\to\Delta(\mathcal{A}_l)$ be defined as in Equation (2) and $\mathcal{E}$ be defined as above. For any distribution over followers $\mathcal{F}$ and contexts $\mathbf{z}_1,\ldots,\mathbf{z}_T$, there exists a policy $\pi^{\mathcal{E}}:\mathcal{Z}\to\mathcal{E}$ such that*

$$\mathbb{E}_{f_1,\ldots f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T} u_l(\mathbf{z}_t, \pi^{\mathcal{E}}(\mathbf{z}_t), b_{f_t}(\pi^{\mathcal{E}}(\mathbf{z}_t)))\right] \ge \mathbb{E}_{f_1,\ldots f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T} u_l(\mathbf{z}_t, \pi^*(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t)))\right] - 2\delta T.$$

Proof. Observe that fixing $a_{f,1},\ldots,a_{f,K}$ as in Optimization (1) fixes a mapping σ and thus a best-response region $\mathcal{P}_\sigma$. Since each $\mathcal{P}_\sigma$ is a convex polytope, the optimal solution of Optimization (1) will be an extreme point of the induced best-response region. Therefore for any $\mathbf{z}\in\mathcal{Z}$, $\pi^*(\mathbf{z})$ will be an extreme point of $\mathcal{P}_\sigma$ for some σ, although $\mathbb{E}_{f\sim\mathcal{F}}[u_l(\mathbf{z},\pi^*(\mathbf{z}), b_f(\pi^*(\mathbf{z})))]$ may not necessarily be attained due to follower tie-breaking rules. Overloading notation, let $\mathcal{P}_{\pi^*(\mathbf{z}_t)}$ denote the best-response region corresponding to $\pi^*(\mathbf{z}_t)$, i.e., $\pi^*(\mathbf{z}_t)\in\mathcal{P}_{\pi^*(\mathbf{z}_t)}$. Since for a given context $\mathbf{z}\in\mathcal{Z}$ the leader's utility is a linear function of $\mathbf{x}$ over the convex polytope $\mathcal{P}_{\pi^*(\mathbf{z})}$, there exists a point $\mathbf{x}_t(\mathbf{z}_t)\in\text{cl}(\mathcal{P}_{\pi^*(\mathbf{z}_t)})$ such that $\mathbb{E}_{f\sim\mathcal{F}}[u_l(\mathbf{z}_t,\mathbf{x}_t(\mathbf{z}_t), b_f(\pi^*(\mathbf{z}_t)))] \ge \mathbb{E}_{f\sim\mathcal{F}}[u_l(\mathbf{z}_t,\pi^*(\mathbf{z}_t), b_f(\pi^*(\mathbf{z}_t)))]$. Let $\mathbf{x}'_t(\mathbf{z}_t)$ denote the corresponding point in $\mathcal{E}$ such that $\|\mathbf{x}'_t(\mathbf{z}_t)-\mathbf{x}_t(\mathbf{z}_t)\|_1\le\delta$. Since $u_l\in[-1,1]$ and is linear in $\mathbf{x}$ for a fixed context and follower best-response,

$$\begin{aligned}\mathbb{E}_{f_t\sim\mathcal{F}}[u_l(\mathbf{z}_t,\mathbf{x}'_t(\mathbf{z}_t), b_{f_t}(\mathbf{x}'_t(\mathbf{z}_t)))] &= \mathbb{E}_{f_t\sim\mathcal{F}}[u_l(\mathbf{z}_t,\mathbf{x}'_t(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t)))]\\ &\ge \mathbb{E}_{f_t\sim\mathcal{F}}[u_l(\mathbf{z}_t,\mathbf{x}_t(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t)))] - 2\delta\\ &\ge \mathbb{E}_{f_t\sim\mathcal{F}}[u_l(\mathbf{z}_t,\pi^*(\mathbf{z}_t), b_{f_t}(\pi^*(\mathbf{z}_t)))] - 2\delta\end{aligned}$$

Summing over T, we obtain the desired result for the policy $\pi^{\mathcal{E}}(\mathbf{z}) = \mathbf{x}'_t(\mathbf{z})$. □

Lemma C.3 implies that the leader pays at most a cost of $2\delta T$ by restricting himself to policies which map to mixed strategies in $\mathcal{E}$. For small enough choice of δ (e.g., $\delta=\mathcal{O}(\frac{1}{\sqrt{T}})$), this additional regret is negligible.

In order to prove performance guarantees for Algorithm 1 (where $\mathcal{F}$ is unknown), we make use of the following lemma:

Lemma C.4. *If*

$$\pi^{(\mathbb{P})}(\mathbf{z}) := \arg\max_{\mathbf{x}\in\mathcal{E}} \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\mathbf{x},b_{\alpha_i}(\mathbf{x}))$$

and

$$\pi^{(\mathbb{P}')}(\mathbf{z}) := \arg\max_{\mathbf{x}\in\mathcal{E}} \sum_{i=1}^{K} \mathbb{P}'(f=\alpha_i) u_l(\mathbf{z},\mathbf{x},b_{\alpha_i}(\mathbf{x})),$$

then

$$\sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi(\mathbf{z}))) \geq \sum_{i=1}^{K} \mathbb{P}'(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P})}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P})}(\mathbf{z}))) - \|\mathbb{P}-\mathbb{P}'\|_1.$$

Proof. By the definition of $\pi(\mathbf{z})^{(\mathbb{P}')}$,

$$\sum_{i=1}^{K} \mathbb{P}'(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P})}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P})}(\mathbf{z}))) \leq \sum_{i=1}^{K} \mathbb{P}'(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z})))$$

Adding and subtracting $\mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z})))$, we see that

$$\begin{aligned}
\sum_{i=1}^{K} \mathbb{P}'(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P})}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P})}(\mathbf{z}))) &\leq \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z}))) \\
&+ \sum_{i=1}^{K} (\mathbb{P}'(f=\alpha_i) - \mathbb{P}(f=\alpha_i)) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z}))) \\
&\leq \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z}))) \\
&+ \sum_{i=1}^{K} |(\mathbb{P}'(f=\alpha_i) - \mathbb{P}(f=\alpha_i)) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi(\mathbf{z})))| \\
&\leq \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z}))) \\
&+ \sum_{i=1}^{K} |(\mathbb{P}'(f=\alpha_i) - \mathbb{P}(f=\alpha_i))| \\
&= \sum_{i=1}^{K} \mathbb{P}(f=\alpha_i) u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z}))) + \|\mathbb{P}-\mathbb{P}'\|_1
\end{aligned}$$

where the third inequality uses the fact that $|u_l(\mathbf{z},\pi^{(\mathbb{P}')}(\mathbf{z}),b_{\alpha_i}(\pi^{(\mathbb{P}')}(\mathbf{z})))| \leq 1$. Rearranging terms obtains the desired result. □

Theorem C.5. *Algorithm 1 obtains expected contextual Stackelberg regret*

$$\mathbb{E}[R(T)] \leq \mathcal{O}(\sqrt{K^2 T \log(T)}),$$

where the expectation is taken over both the follower population and the randomness in the leader's mixed strategies as in Definition 3.3.

Proof. By Lemma C.3,

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq \mathbb{E}_{f_1,\ldots,f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T} u_l(\mathbf{z}_t,\pi^{\mathcal{E}}(\mathbf{z}_t),b_{f_t}(\pi^{\mathcal{E}}(\mathbf{z}_t))) - u_l(\mathbf{z}_t,\pi_t(\mathbf{z}_t),b_{f_t}(\pi_t(\mathbf{z}_t)))\right] + 2\sqrt{T} \\
&= \sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{P}(f=\alpha_i)(u_l(\mathbf{z}_t,\pi^{\mathcal{E}}(\mathbf{z}_t),b_{\alpha_i}(\pi^{\mathcal{E}}(\mathbf{z}_t))) - u_l(\mathbf{z}_t,\pi_t(\mathbf{z}_t),b_{\alpha_i}(\pi_t(\mathbf{z}_t)))) + 2\sqrt{T}
\end{aligned}$$

for $\delta \leq T^{-1/2}$, since $\mathbb{P}(f_t = \alpha_i) = \mathbb{P}(f = \alpha_i)$ for all $t \in [T]$. Since Lemma C.4 applies for any realization of $\mathbb{P}' = \widehat{\mathbb{P}}_t$ and $\pi^{(\mathbb{P}')} = \pi_t$, it also holds in expectation over $f_1, \ldots, f_T \sim \mathcal{F}$. Applying this result, we obtain

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq \mathbb{E}_{f_1,\ldots,f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\mathbb{P}(f=\alpha_i) - \widehat{\mathbb{P}}_t(f=\alpha_i))u_l(\mathbf{z}_t, \pi^{\mathcal{E}}(\mathbf{z}_t), b_{\alpha_i}(\pi^{\mathcal{E}}(\mathbf{z}_t))) + \|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] + 2\sqrt{T} \\
&\leq \mathbb{E}_{f_1,\ldots,f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}|(\mathbb{P}(f=\alpha_i) - \widehat{\mathbb{P}}_t(f=\alpha_i))| \cdot |u_l(\mathbf{z}_t, \pi^{\mathcal{E}}(\mathbf{z}_t), b_{\alpha_i}(\pi^{\mathcal{E}}(\mathbf{z}_t)))| + \|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] + 2\sqrt{T} \\
&\leq \mathbb{E}_{f_1,\ldots,f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}|(\mathbb{P}(f=\alpha_i) - \widehat{\mathbb{P}}_t(f=\alpha_i))| + \|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] + 2\sqrt{T} \\
&= 2\mathbb{E}_{f_1,\ldots,f_T\sim\mathcal{F}}\left[\sum_{t=1}^{T}\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] + 2\sqrt{T} = 2\sum_{t=1}^{T}\mathbb{E}_{f_1,\ldots,f_{t-1}\sim\mathcal{F}}\left[\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] + 2\sqrt{T}
\end{aligned}$$

Next, we upper-bound $\mathbb{E}_{f_1,\ldots,f_{t-1}\sim\mathcal{F}}\left[\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right]$ via Hoeffding's inequality. For $t = 1$, a trivial upper-bound on $\|\mathbb{P} - \widehat{\mathbb{P}}_1\|_1$ is K. For $t \geq 2$, note that $\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1$ may be rewritten as

$$\begin{aligned}
\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1 &:= \sum_{i=1}^{K}|\widehat{\mathbb{P}}_t(f=\alpha_i) - \mathbb{P}(f=\alpha_i)| \\
&= \sum_{i=1}^{K}|\widehat{\mathbb{P}}_t(f=\alpha_i) - \mathbb{E}[\mathbb{1}\{f=\alpha_i\}]| \\
&= \sum_{i=1}^{K}\frac{1}{t-1}\left|\sum_{s=1}^{t-1}\mathbb{1}\{f_s=\alpha_i\} - \mathbb{E}[\sum_{s=1}^{t-1}\mathbb{1}\{f_s=\alpha_i\}]\right|.
\end{aligned}$$

By Hoeffding's inequality,

$$\frac{1}{t-1}\left|\sum_{s=1}^{t-1}\mathbb{1}\{f_s=\alpha_i\} - \mathbb{E}[\sum_{s=1}^{t-1}\mathbb{1}\{f_s=\alpha_i\}]\right| \geq \sqrt{\frac{\log(2/\beta)}{2(t-1)}}$$

with probability at least $1 - \beta$. Therefore,

$$\begin{aligned}
\mathbb{E}_{f_1,\ldots,f_{t-1}\sim\mathcal{F}}\left|\widehat{\mathbb{P}}_t\{f=\alpha_i\} - \mathbb{P}\{f=\alpha_i\}\right| &\leq (1-\beta)\cdot\sqrt{\frac{\log(2/\beta)}{t-1}} + \beta\cdot 1 \\
&\leq \sqrt{\frac{\log(2/\beta)}{2(t-1)}} + \beta
\end{aligned}$$

for any $\beta \in (0, 1)$. Setting $\beta = T^{-1}$, we see that

$$\mathbb{E}_{f_1,\ldots,f_{t-1}\sim\mathcal{F}}\left[\|\mathbb{P} - \widehat{\mathbb{P}}_t\|_1\right] \leq \sqrt{\frac{K^2\log(2T)}{2(t-1)}} + \frac{K}{T}.$$

Plugging this expression into our upper-bound on $\mathbb{E}[R(T)]$, we obtain

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq 2\sqrt{T} + K + 1 + \sum_{t=2}^{T}\sqrt{\frac{K^2\log(2T)}{2(t-1)}} \\
&\leq 2\sqrt{T} + K + 1 + \sqrt{K^2\log(2T)}\sum_{t=2}^{T}\sqrt{t-1} - \sqrt{t-2} \\
&\leq 2\sqrt{T} + K + 1 + \sqrt{K^2T\log(2T)}.
\end{aligned}$$

□

D Appendix for Section 3.3: Stochastic contexts

Corollary D.1 (Optimal Policy). *The optimal-in-hindsight policy takes the following closed form:*

$$\pi^*(\mathbf{z}) = \arg\max_{\mathbf{x}\in\Delta(\mathcal{A}_l)} \sum_{a_l\in\mathcal{A}_l} \mathbf{x}[a_l]\cdot\langle \mathbf{z}, \sum_{i=1}^{K} \boldsymbol{\theta}(a_l, b_{\alpha_i}(\mathbf{x}))\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{f_t=\alpha_i\}\rangle$$

Proof.

$$\begin{aligned}\pi^*(\mathbf{z}) &:= \arg\max_{\mathbf{x}\in\Delta(\mathcal{A}_l)} \sum_{t=1}^{T}\sum_{a_l\in\mathcal{A}_l} \mathbf{x}[a_l]\cdot\langle \mathbf{z}, \boldsymbol{\theta}(a_l, b_{f_t}(\mathbf{x}))\rangle \\ &= \arg\max_{\mathbf{x}\in\Delta(\mathcal{A}_l)} \sum_{a_l\in\mathcal{A}_l} \mathbf{x}[a_l]\cdot\langle \mathbf{z}, \sum_{t=1}^{T}\boldsymbol{\theta}(a_l, b_{f_t}(\mathbf{x}))\rangle \\ &= \arg\max_{\mathbf{x}\in\Delta(\mathcal{A}_l)} \sum_{a_l\in\mathcal{A}_l} \mathbf{x}[a_l]\cdot\langle \mathbf{z}, \sum_{i=1}^{K} \boldsymbol{\theta}(a_l, b_{\alpha_i}(\mathbf{x}))\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{f_t=\alpha_i\}\rangle\end{aligned}$$

□

Theorem D.2. *Algorithm 2 obtains expected contextual Stackelberg regret $\mathbb{E}[R(T)] \leq \mathcal{O}(\sqrt{TK\log(T)})$, where $\Omega := \{\boldsymbol{\omega} : \boldsymbol{\omega}\in\Delta^K, T\cdot\boldsymbol{\omega}[i]\in\mathbb{N}, \forall i\in[K]\}$ and the expectation is taken over both the distribution over contexts and the randomness in the leader's mixed strategies as in Definition 3.5.*

Given Corollary D.1, the proof follows from the standard potential-based analysis of Hedge.

Proof. Let $\Phi_t := \sum_{\pi^{(\boldsymbol{\omega})}\in\Pi} \mathbf{q}_t[\pi^{(\boldsymbol{\omega})}]$. Observe that

$$\begin{aligned}\Phi_{t+1} &= \sum_{\pi^{(\boldsymbol{\omega})}\in\Pi} \mathbf{q}_t[\pi^{(\boldsymbol{\omega})}]\cdot\exp(-\eta\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}])\cdot\frac{\sum_{\pi^{(\boldsymbol{\omega}')}\in\Pi}\mathbf{q}_t[\pi^{(\boldsymbol{\omega}')}]}{\sum_{\pi^{(\boldsymbol{\omega}')}\in\Pi}\mathbf{q}_t[\pi^{(\boldsymbol{\omega}')}]} \\ &= \Phi_t\cdot\sum_{\pi^{(\boldsymbol{\omega})}\in\Pi}\mathbf{p}_t[\pi^{(\boldsymbol{\omega})}]\cdot\exp(-\eta\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}]) \\ &\leq \Phi_t\cdot\sum_{\pi^{(\boldsymbol{\omega})}\in\Pi}\mathbf{p}_t[\pi^{(\boldsymbol{\omega})}]\cdot(1-\eta\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}]+\eta^2\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}]^2)\end{aligned}$$

where the inequality follows from the fact that $e^{-x}\leq 1-x+x^2$, for $|x|\leq 1$. Distributing terms, we see that

$$\begin{aligned}\Phi_{t+1} &\leq \Phi_t\cdot(1-\eta\sum_{\pi^{(\boldsymbol{\omega})}\in\Pi}\mathbf{p}_t[\pi^{(\boldsymbol{\omega})}]\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}]+\eta^2\sum_{\pi^{(\boldsymbol{\omega})}\in\Pi}\mathbf{p}_t[\pi^{(\boldsymbol{\omega})}]\cdot\boldsymbol{\ell}_t[\pi^{(\boldsymbol{\omega})}]^2) \\ &\leq \Phi_t\cdot\exp(-\eta\cdot\langle\mathbf{p}_t,\boldsymbol{\ell}_t\rangle+\eta^2\cdot\langle\mathbf{p}_t,\boldsymbol{\ell}_t^2\rangle),\end{aligned}$$

where the second inequality follows from the fact that $1+x\leq e^x$, and $\boldsymbol{\ell}_t^2\in\mathbb{R}^{|\Pi|}$ is defined such that $\boldsymbol{\ell}_t^2[\pi]:=\boldsymbol{\ell}_t[\pi]^2$. Since $\Phi_T\geq\mathbf{q}_T(\pi)=\exp(-\eta\cdot\sum_{t=1}^{T}\boldsymbol{\ell}_t[\pi])$ for any $\pi\in\Pi$, we know that

$$\exp(-\eta\cdot\sum_{t=1}^{T}\boldsymbol{\ell}_t[\pi^*])\leq|\Pi|\cdot\exp(-\eta\cdot\sum_{t=1}^{T}\langle\mathbf{p}_t,\boldsymbol{\ell}_t\rangle+\eta^2\cdot\sum_{t=1}^{T}\langle\mathbf{p}_t,\boldsymbol{\ell}_t^2\rangle).$$

Taking the log on both sides, rearranging terms, and using the fact that losses are bounded between -1 and 1, we get

$$\sum_{t=1}^{T}\langle\mathbf{p}_t,\boldsymbol{\ell}_t\rangle-\sum_{t=1}^{T}\boldsymbol{\ell}_t(\pi^*)\leq\frac{\log|\Pi|}{\eta}+\eta T,$$

which is less than $2\sqrt{T\log|\Pi|}$ if $\eta=\sqrt{\frac{\log|\Pi|}{T}}$. The final result is obtained by observing that $|\Pi|\leq T^K$ for $T\geq 2$. □