Minimax Rates in Contextual Partial Monitoring

Anonymous Author(s)
Affiliation
Address
email

Abstract

We generalize the finite partial monitoring problem to the contextual setting. Partial monitoring allows learning even when the loss of the chosen action is not observed. In the noncontextual problem, the minimax regret is known to be $O(T^{2/3})$ if a global observability condition is satisfied and improves to $O(\sqrt{T})$ under a stronger local observability condition. Perhaps surprisingly, we show that the same characterization does not hold in the contextual case and a stronger notion of *pairwise observability* is necessary for $O(\sqrt{T})$ minimax regret. In particular, we provide a lower bound of $O(T^{2/3})$ for any non-pairwise observable game, including locally observable games, in the contextual setting. We propose two algorithms for adversarial environments. The first requires a finite policy class but allows for arbitrary contexts and can be tuned to obtain the optimal $O(\sqrt{T})$ rate in pairwise observable settings or the optimal $O(T^{2/3})$ rate otherwise. The second allows for arbitrary policy classes with an empirical risk minimization oracle but requires i.i.d. contexts; we also show an $O(T^{2/3})$ upper bound and an efficient implementation using only a constant number of oracle calls per round.

1 Introduction

In online learning, we model the world as a sequential game of T rounds between the learner and a possibly adversarial environment. In particular, this paper studies the finite partial monitoring setting proposed in [14], where, for each round t, the learner chooses an action $I_t \in \underline{\mathrm{N}} := \{1, \dots, N\}$, the environment chooses a response $j_t \in \underline{\mathrm{M}} := \{1, \dots, M\}$, and the learner incurs loss L_{I_t,j_t}, the I_t, j_t entry of a fixed and known loss matrix $L \in [0,1]^{N \times M}$. However, the learner does not observe j_t or L_{I_t,j_t}, but rather H_{I_t,j_t}, the corresponding entry from the fixed and known feedback matrix $H \in [0,1]^{N \times M}$. Intuitively, one should imagine that each row of H has only a few distinct elements and that observing H_{I_t,j_t} only allows the learner to determine j_t up to some subset of $\underline{\mathrm{M}}$. In particular, the loss incurred by the algorithm is not observed, making partial monitoring more difficult that bandit feedback.

For example, the partial monitoring game where each row of H has distinct elements (i.e. $1, \dots, M$) is equivalent to full information, as we can infer j_t and therefore the full loss vector. The *Revealing action* game is more interesting. Define $L = \begin{bmatrix} 0 & a \\ a & 0 \\ c & c \end{bmatrix}$ and $H = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 2 \end{bmatrix}$, which encodes the game where the learner tries to match the play of the adversary and incurs loss a if incorrect. However, the learner only obtains useful feedback if she plays $I = 3$ at a fixed cost c, in which case j_t can be deduced. Partial monitoring allows games where the learner must choose between a low cost but uninformative action and a high cost but informative action.

Partial monitoring is more than a technical challenge and can be used to model practical scenarios. For example, consider dynamic pricing: at each round, the learner sets a price p_t and a buyer with valuation v_t arrives. If $p_t \leq v_t$, the good is purchased and the cost to the learner is $v_t - p_t$, the lost potential revenue from not setting a higher price. However, the learner only observes whether $v_t < p_t$ and not the loss (see [6] for details). More generally, partial monitoring can model scenarios where the learner only sees a quantized version of the true loss, such as: (i) surveys where a numerical values are binned, which have a long history as "partial identification" in the econometrics literature, (ii) recommendation engines, where only coarse feedback from the recommendation (e.g. a like/dislike) is provided, but the learner wishes to find the best recommendations and not just maximize the number of likes, and (iii) robust algorithms, where we only estimate parameters to some confidence region but still want good performance.

Perhaps the biggest obstacle to more widespread adoption of partial monitoring as a modeling tool has been its inability to use context. The learner often has access to a context vector and hopes to choose more informed actions because of it. For example, the noncontextual problem for recommendation engines is tantamount to learning the single best item across all users, but is devoid of personalization.

With this motivation in mind, we propose the contextual partial monitoring problem. In addition to L and H, the learner is also provided with a policy class Π and, at every round, a context $x_t \in \mathcal{X}$. A policy $\pi : \mathcal{X} \to \underline{\mathrm{N}}$ maps a context to an action. We only consider deterministic policies, but our results extend to randomized policies in expectation. When it is clear from context, we will also use $\pi(x)$ to represent $e_{\pi(x)}$, the unit vector corresponding to the choice of the policy. The goal of the learner is to minimize the contextual regret, the excess cumulative loss when compared to the best policy in Π:

$$\mathcal{R}_T := \sum_{t=1}^{T} L_{I_t, j_t} - \min_{\pi \in \Pi} \sum_{t=1}^{T} L_{\pi(x_t), j_t}. \tag{1}$$

We will present algorithms for two typical models of the policy class. The first assumes that Π is finite, which allows individual weights on every policy. The second assumes access to an empirical risk minimization (ERM) value oracle, which, given some list of contexts $x_1, \ldots, x_t$ and losses $\ell_1, \ldots, \ell_t$, returns $\min_{\pi \in \Pi} \sum_{s=1}^{t} \pi(x_s)^\top \ell_s$. Note that only the value of the optimal policy is required. We will present algorithms for both settings.

1.1 Related Work

To the best of our knowledge, there are no results on the contextual partial monitoring problem with arbitrary policy classes. The closest work to ours, by Bartók and Szepesvári [4], considered actions chosen by $f(x_t)$, where f is some some fixed but unknown function. The algorithms construct explicit confidence regions for f and the play optimistically.

The noncontextual partial monitoring problem, proposed by Piccolboni and Schindelhauer [14], has been well studied. The *global observability* condition was established as a necessary and sufficient for sublinear regret [8] by a $O(T^{2/3})$ upper bound and matching $\Omega(T^{2/3})$ lower bound. A faster rate of $O(\sqrt{T})$ is possible when a stronger *local observability* holds [5] which matches a $\Omega(T^{2/3})$ lower bound for all games without local observability. These results were later extended to stochastic adversaries [10] with the same classification, and the effect of degenerate actions was recently resolved [11]. See the work of Bartók et al. [6] for more discussion.

Partial monitoring is an extension of a large body of work on games with incomplete information, including bandit feedback [7], semi-bandit feedback [3], graph feedback [2], and many others. Finally, we note that the problem of partial monitoring has a very long history in the econometrics literature under the name partial identification; see e.g. [13, 12] and references therein.

Relaxation based algorithms, first proposed by Rakhlin et al. [16], have recently been extended to the contextual bandit setting [15, 17]. More generally, algorithms for contextual settings that leveraging empirical risk minimization oracles has been applied to other online learning algorithms, such as follow the perturbed leader [9].

1.2 Our Contributions

We propose the contextual finite partial monitoring problem and characterize the possible minimax rates. In the noncontextual case, the *local observability* condition, which essentially requires that similar losses can be distinguished solely from the feedback from playing those losses, was shown to be necessary and sufficient to obtain $O(\sqrt{T})$ regret. To the contrary, we show that in the contextual case, a much stronger *pairwise observability* condition is needed.

In Section 3, we provide two exponential weights algorithms that dynamically change the point of reference of the loss estimates and are able to capture $O(\sqrt{T})$ regret in the pairwise observability setting and $O(T^{2/3})$ regret in the globally observable setting. These results hold for adversarial contexts and actions but require a finite policy class.

Section 4 provides the first relaxation based algorithm in the partial monitoring setting and shows a $O(T^{2/3})$ regret in the finite policy class case. This setting requires access to unlabeled samples of the contexts but allows any policy class with an ERM value oracle. We also provide an efficient implementation requiring only $O(N)$ oracle calls per round.

We turn to lower bounds in Section 5 and show that, without pairwise observability, no algorithm can achieve better that $\Omega(T^{2/3})$ expected regret. The main implication is that the algorithms from Section 3 capture the correct structural dependence on the regret. Additionally, to the best of our knowledge, this is the first time relaxation based algorithms obtain the minimax regret in an adversarial partial information setting, since the $\Omega(T^{2/3})$ lower matches the upper bound of Section 4. There are no known tight bounds for the bandit setting.

Notation Collections over time are denoted by $a_{1:t} := a_1, \ldots, a_t$. The ith standard basis vector is denoted e_i, the all ones vector denoted $\mathbf{1}$, and the ith element of vector v denoted $v(i) := v^\top e_i$. Functions are applied to vectors elementwise; in particular, $\mathrm{sgn}(v)$ is a vector with ith element equal to the sign function of $v(i)$. In general, adversary distribution are denoted by p_t and player distributions by q_t. Finally, $\mathbb{1}\{\cdot\}$ is the indicator function.

2 Partial Monitoring with Context

Obtaining low regret in a partial monitoring game requires careful study of loss structure, the feedback structure, and how the relate. For clarity, we break up the discussion and the definitions in this manner.

The Loss Structure Define the loss vector $\ell_i = \boldsymbol{e}_i^\top L^\top$ to be the transpose of the row corresponding to playing action i. The cell corresponding to action i is $C_i := \{p \in \triangle_M : \ell_i^\top p \le \ell_j^\top p \ \forall j\}$, the set of stochastic adversary strategies for which action i is optimal. The cells $C_1, \ldots, C_N$ induce a partition of $\triangle_M$. Action i is degenerate if $C_i \subset C_j$ for some j. Two non-degenerate actions i and j are neighbors if $C_i \cap C_j$ is an $M-2$ dimensional polytope (e.g. there is a set of $p \in \triangle_M$ where i and j are both optimal) and denote the neighbor set of (i,j) as $N_{i,j} := \{k \in \underline{\mathrm{N}} : C_i \cap C_j \subseteq C_k\}$, which includes i, j, and any action k with $C_k \subseteq C_i \cap C_j$. We may not simply ignore actions like k because action k could provide information that action i or j cannot. Finally, the set of all pairs of neighboring actions is denoted $\mathcal{N}$.

The Feedback Structure We enumerate the distinct values of the ith row of H by $\sigma_1, \ldots, \sigma_{s_i}$; when we play i, we observe feedback σ and can conclude that j_t must have been such that $H_{I_t, j_t} = \sigma$. Define the signaling matrix $(S_k)_{(i,j)} = \mathbb{1}\{H_{(k,j)} = \sigma_i\}$ such that the jth row of S_k indicated all choices j_t that could have produced σ_i. Since the exact values of the σ do not matter, we may assume that the feedback received is $Y_t = S_{I_t} e_{j_t}$.

Their Interaction Estimating the loss vectors ℓ_i is impossible for many easy games; for example, in the revealing action game, it is impossible to learn A, but we may still determine which is the optimal action. In fact, estimating the pairwise loss differences is sufficient for low regret. The three following definitions, presented in decreasing generality, encapsulate this notion.

Definition 1. *A partial monitoring game is* globally observable *if, for all pairs i, j, there exists a collection of actions $V_{i,j} \subseteq \underline{N}$ and observer vectors $\{v_{i,j,k} \in \mathbb{R}^{s_k} | k \in V_{i,j}\}$, such that*

$$\ell_i - \ell_j = \sum_{k \in V_{i,j}} S_k^\top v_{i,j,k}. \tag{2}$$

Throughout, we use $V_\infty = \max_{i,j,k} \|v_{i,j,k}\|_\infty$. Having $V_{i,j} = \underline{N}$ is sufficient for global observability.

If, in addition, we make take $V_{i,j} = N_{i,j}$ for all neighbor pairs $(i,j) \in \mathcal{N}$, then the game is said to be locally observable. *Finally, if we may take $V_{i,j} = \{i, j\}$ for all non-degenerate i and j, then the game is* pairwise observable.

Intuitively, global observability means that we can construct unbiased estimates of the loss differences from the feedback by exploiting the equality $(\ell_i - \ell_j)^\top e_{j_t} = \sum_{k \in V_{i,j}} v_{i,j,k}^\top S_k e_{j_t}$; the left hand side is the actual loss difference between action i and j and the right hand side can be estimated from the feedback $Y_t = S_{I_t} e_{j_t}$. Each round, we will choose a base arm b_t and estimate the column vector

$$\Delta_t = (L - \mathbf{1}\ell_{b_t}^\top) e_{j_t} \tag{3}$$

so that $e_{I_t}^\top \Delta_t = L_{I_t, j_t} - L_{b_t, j_t}$. Out general strategy will be to carefully select b_t every round and use an unbiased estimate of Δ_t as a proxy for the true losses.

3 Exponential Weights Algorithms

This section extends the exponential weights algorithm to the contextual partial monitoring setting and provides upper bounds in the globally observable and pairwise observable settings. The estimator $\hat{\Delta}_t(i) := \sum_{k \in V_{i,b_t}} \mathbb{1}\{k = I_t\} \frac{v_{i,b_t,k}^\top S_k e_{j_t}}{q_t(k)}$ uses importance sampling and is an unbiased estimator for $\Delta_t(i)$. Written in terms of the feedback $Y_t = S_{I_t} e_{j_t}$,

$$\hat{\Delta}_t = \left(\mathbb{1}\{I_t \in V_{1,b_t}\} \frac{v_{1,b_t,I_t}^\top Y_t}{q_t(I_t)}, \ldots, \mathbb{1}\{I_t \in V_{N,b_t}\} \frac{v_{N,b_t,I_t}^\top Y_t}{q_t(I_t)} \right)^\top, \tag{4}$$

and we can show its unbiasedness by calculating

$$\hat{\Delta}_t(i) = \mathbb{E}\left[\mathbb{1}\{I_t \in V_{i,b_t}\} \frac{v_{i,b_t,I_t}^\top S_{I_t} e_{j_t}}{q_t(I_t)} \right] = \mathbb{E}\left[\sum_{k \in V_{i,b_t}} q_t(k) \frac{v_{i,b_t,k}^\top S_k e_{j_t}}{q_t(k)} \right] = (\ell_i - \ell_{b_t})^\top e_{j_t} = \Delta_t(i).$$

Our algorithm, EXP4.PM, is presented in Algorithm 1. At a high level, it is EXP4 using $\hat{\Delta}_t$ for loss estimates with γ uniform exploration and a recentering step that moves the base action to the arm with the highest weight. A similar idea was concurrently proposed for the noncontextual setting by Lattimore and Szepesvari [11].

Algorithm 1 Recentered EXP4.PM

> **Input:** η, γ, T, L, H, and Π
> Calculate observer vectors $v_{i,j,k}$
> Initialize $w_1 = 1/K$
> **for all** $t = 1, \ldots, T$ **do**
>   Receive context x_t
> $q_t \leftarrow (1 - N\gamma) \sum_{k=1}^K \pi_k(x_t) w_t(k) + \gamma \mathbf{1}$
>   Play $I_t \sim q_t$, observe $Y_t = S_{I_t} e_{j_t}$
> $b_t \leftarrow \arg\max_i q_t(i)$
>   Calculate $\hat{\Delta}_t$ from (4)
> $w_{t+1}(k) \leftarrow w_t(k) e^{-\eta \pi_k(x_t)^\top \hat{\Delta}_t}$
> **end for**

As the following theorem shows, we can always tune η and γ to guarantee a $O(T^{2/3})$ regret. If pairwise observability holds, we can obtain a faster rate by using $\gamma = 0$ and by playing the subgame with the degenerate actions removed. This subgame does not have higher regret (there is always a non-degenerate action with loss no higher) and pairwise observability ensures that we construct unbiased estimates of Δ_t from the plays of non-degenerate actions only.

Theorem 1. *For any globally observable game, arbitrary sequence of contexts $x_1, \ldots, x_T$ and adversary actions $j_1, \ldots, j_T$, Algorithm 1 with $\eta = N^{-\frac{1}{3}} \left(\frac{\log(K)}{V_\infty T} \right)^{\frac{2}{3}}$ and $\gamma = \left(\frac{V_\infty^2 \log(K)}{N^2 T} \right)^{\frac{1}{3}}$ yields an expected regret with the bound*

$$\mathbb{E}[\mathcal{R}_T] \leq 3 \left(N V_\infty^2 \log(K) \right)^{\frac{1}{3}} T^{\frac{2}{3}}.$$

If pairwise observability holds, the same algorithm with degenerate actions removed and parameters $\gamma = 0$ and $\eta = \sqrt{\frac{\log(K)}{T V_\infty^2 (N+3)}}$ observes

$$\mathbb{E}[\mathcal{R}_T] \leq 2 V_\infty \sqrt{T(N+3)\log(K)}.$$

The observability and optimal choice of γ and η is determined a priori by L and H, and one can use the standard doubling trick if T is unknown to obtain the same regret rates with a worse constant.

The proof of Theorem 1 is mostly identical to the standard EXP4 proof. Recall that the EXP4 importance weighted estimate, $\hat{\ell}_t = e_{I_t}\ell_t(I_t)/q_t(I_t)$, only has support on the I_t entry, which allows the variance term in the analysis to be easily bounded by $\mathbb{E}\left[\sum_k w_t(x)(\hat{\ell}_t\pi_k(x_t))^2\right] \leq \mathbb{E}\left[V_\infty q_t^\top \hat{\ell}_t q_t(I_t)^{-2}\right]$ since $q_t^\top \hat{\ell}_t = \ell_t(I_t) \leq 1$. In contrast, the importance weighted estimates $\hat{\Delta}_t$ may be non-zero for any entry. Without pairwise observability, $q_t^\top \hat{\Delta}_t$ could have magnitude $\max_i 1/q_t(i)$ which we control by setting $\gamma > 0$. With pairwise observability, we may choose $\hat{\Delta}_t$ to be supported on e_{I_t} and e_{b_t} only, allowing us to control $q_t^\top \hat{\Delta}_t$ by choosing b_t such that $q_t(b_t)$ is not too small. The full proof is in Appendix A.

4 A Relaxation Algorithm

We now turn our focus to policy classes where the only assumption made is access to an ERM value oracle, thereby extending the relaxation framework for contextual bandits [15] to the contextual partial monitoring setting. We first review the necessary details of the relaxation framework before describing an efficient (in terms of the number of oracle calls) algorithm with $O(T^{2/3})$ regret, which will match the lower bound of Section 5.

4.1 A Sparser Offset Loss Estimate

The regret analysis for relaxation algorithms requires careful control of the sparsity of the offset loss estimates. Throughout this section, fix a base arm b. Define $V(i) = [v_{1,b,i}, \ldots, v_{N,b,i}]$, which implies that $\Delta_t = \sum_k V(k)^\top S_k e_{j_t}$. Instead of constructing an unbiased estimate for Δ_t from importance weighting, we will instead borrow a trick from [17] and correct for bias by multiplying $V(I_t)^\top Y_t$ by a Bernoulli random variable with expectation $\propto 1/q_t(I_t)$, as described by the following lemma.

Lemma 1. *Assume that $q_t(i) \geq \gamma$ for all i and define $\hat{Z}_t = V_\infty \gamma^{-1} \operatorname{diag}(\hat{B}_{1,t}, \ldots, \hat{B}_{N,t})$ for $\hat{B}_{i,t} \sim \operatorname{Bernoulli}\left(\frac{\gamma}{V_\infty}\frac{|e_i^\top V(I_t)^\top Y_t|}{q_t(I_t)}\right)$. Then, the following offset loss estimate is unbiased:*

$$\hat{\Delta}_t := \hat{Z}_t \operatorname{sgn}(V(I_t)^\top Y_t). \tag{5}$$

Proof. First, the probability that $\hat{B}_{i,t} = 1$ is well defined: $q_t(i) \geq \gamma$ and, since $Y_t = S_k e_{j_t}$ is a unit vector, $|V(i)^\top Y_t| \leq V_\infty$. We can directly verify that

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_t(i)] &= \mathbb{E}\left[\operatorname{sgn}\left(e_i^\top V(I_t)^\top Y_t\right) V_\infty \gamma^{-1} \mathbb{E}\left[\hat{B}_{i,t}\middle| I_t\right]\right] \\ &= \mathbb{E}\left[\operatorname{sgn}\left(e_i^\top V(I_t)^\top S_{I_t} e_{j_t}\right) \frac{\left|e_i^\top V(I_t)^\top S_{I_t} e_{j_t}\right|}{q_t(I_t)}\right] = \Delta_t(i).\end{aligned}$$

□

4.2 Relaxations

The relaxation framework allows one to simultaneous derive algorithms and upper bounds on regret for sequential learning problems. We will keep our description short and refer the reader to [16] and [15] for elaboration on the general sequential prediction and bandit feedback settings, respectively.

During round t, the algorithm collects history $(x_t, I_t, q_t, Y_t, \hat{Z}_t)$. Let $\mathcal{H}^t = \mathcal{H}^{t-1} \cap (x_t, I_t, q_t, Y_t, \hat{Z}_t)$ with $\mathcal{H}^0 = \emptyset$.

Definition 2. *A relaxation $\mathbf{Rel}(\cdot)$ is a function from $\mathcal{H}^t$ to $\mathbb{R}$ for all $t = 1, \ldots, T$. It is admissible if,*

1. *for all $x_{1:T}$, $j_{1:T}$, and $q_{1:T}$,*

$$\mathbb{E}_{I_{1:T}\sim q_{1:T}, \hat{Z}_{1:T}}[\mathbf{Rel}(\mathcal{H}^T)] \geq -\inf_{\pi\in\Pi} \sum_{t=1}^{T} \pi(x_t)^\top \Delta_t, \text{ and} \tag{6}$$

2. for all $t = 1, \dots, T$ and any history $\mathcal{H}^{t-1}$,

$$\mathbb{E}_{x_t}\left[\inf_{q_t}\sup_{j_t}\mathbb{E}_{I_t\sim q_t,\hat{Z}_t}\left[e_{I_t}^\top\Delta_t + \mathbf{Rel}(\mathcal{H}^{t-1}\cup\mathcal{H}^t)\right]\right] \le \mathbf{Rel}(\mathcal{H}^{t-1}). \tag{7}$$

Furthermore, any strategy q_t which satisfies (7) *is called admissible.*

The first condition ensures that $\mathbf{Rel}(\mathcal{H}^T)$ is an upper bound on the offset loss of the comparator, and the second condition ensures that, under the player strategy q_t, the relaxation remains an upper bound against all j_t. The main utility of $\mathbf{Rel}$ is that it produces a bound on the regret, even though it is defined it terms of the offset losses.

Lemma 2. *If $\mathbf{Rel}(\cdot)$ is an admissible relaxation, then for any $j_{1:T}$, we have*

$$\mathbb{E}[\mathcal{R}_T] \le \mathbf{Rel}(\emptyset).$$

Proof. Applying Lemma 1 from Rakhlin and Sridharan [15] with $c_t = \Delta_t$,

$$\begin{aligned}\mathbf{Rel}(\emptyset) &\ge \mathbb{E}[\sum_{t=1}^T q_t^\top\Delta_t - \inf_{\pi\in\Pi}\sum_{t=1}^T \pi(x_t)^\top\Delta_t]\\ &= \mathbb{E}[\sum_{t=1}^T q_t^\top(L - \mathbf{1}\ell_b^\top)e_{j_t} - \inf_{\pi\in\Pi}\sum_{t=1}^T\pi(x_t)^\top(L-\mathbf{1}\ell_b^\top)e_{j_t}] = \mathbb{E}[\mathcal{R}_T].\end{aligned}$$

□

4.3 An Admissible Relaxation

The relaxation framework provides at automatic regret bound for any admissible relaxation; the challenge typically is in proving admissibility. The relaxation we use is is modeled on the relaxation for the adversarial contextual bandit settings [15, 17] and uses sequential Rademacher averages and a random rollout.

Notation Let $\boldsymbol{\epsilon}_{t,i}$ denote the matrix with zero elements except for a Rademacher random variable (equal probability $\{1,-1\}$) in the i,i entry, $\boldsymbol{\epsilon}_t$ denote their collection, and $\boldsymbol{Z}_t$ denote the random vector with i.i.d. coordinates equal to $V_\infty\gamma^{-1}$ with probability γ and equal to 0 otherwise. We will use $\mathcal{D}_1 = \{-V_\infty\gamma^{-1}, 0, V_\infty\gamma^{-1}\}$ and $\mathcal{D} = \{x\in\mathbb{R}^N : x^\top e_i \in \mathcal{D}_1 \forall i\}$. We also define $\triangle'_{\mathcal{D}}$ to be distributions such that, if $X \sim p \in \triangle'_{\mathcal{D}}$, then $p(X^\top e_i = V_\infty\gamma^{-1}) \le N\gamma$ and $p(X^\top e_i = -V_\infty\gamma^{-1}) \le N\gamma$ for all i. Finally, we write $D_i = \mathrm{diag}(e_i)$ so that $u^\top v = \sum_{i=1}^N u^\top D_i v$.

The relaxation at round t is a function of the past data, encapsulated in $\mathcal{H}^{t-1}$, and some randomly drawn data representing the uncertainty in the future $\rho_t = \bigcup_{s\in t+1:T}\{x_s, \boldsymbol{\epsilon}_s, \boldsymbol{Z}_s\}$, where x_s for $s > t$ are i.i.d. samples from the context distribution. Define

$$R_i\left(\mathcal{H}^t, \rho_t\right) = \sup_{\pi\in\Pi} \underbrace{-\sum_{s=1}^t \pi(x_s)^\top D_i\hat{\Delta}_s}_{\text{past data}} \underbrace{-\sum_{s=t+1}^T 2\pi(x_s)^\top\boldsymbol{\epsilon}_{s,i}Z_s}_{\text{future uncertainty}} + (T-t)\gamma, \tag{8}$$

which is best fit of Π to the ith coordinate of the past and future data.

Theorem 2. *The relaxation*

$$\mathbf{Rel}(\mathcal{H}^{t-1}) = \mathop{\mathbb{E}}_{\rho_t}\left[\sum_{i=1}^N R_i\left(\mathcal{H}^{t-1},\rho_t\right)\right] \tag{9}$$

and the strategy that samples $\rho_t = \bigcup_{s\in t+1:T}\{x_s, \boldsymbol{\epsilon}_s, \boldsymbol{Z}_s\}$ and plays $q_t = (1-N\gamma)q_t^* + \gamma\mathbf{1}$ for

$$q_t^* = \mathop{\arg\min}_{q\in\triangle_N}\sup_{p_t\in\triangle'_{\mathcal{D}}}\mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[q^\top\hat{\Delta}_t + \sum_{i=1}^N R_i\left(\mathcal{H}^{t-1},\rho_t\right)\right] \tag{10}$$

are admissible.

Algorithm 2 Computing q_t^*

Input: History $\mathcal{H}^{t-1}$, random rollout ρ_t

Calculate $a_i := \frac{\gamma}{V_\infty}\min\left\{\psi^0 - \psi_i^+, \psi_i^- - \psi^0\right\}$ and $b_i := \frac{\gamma}{V_\infty}\max\left\{\psi^0 - \psi_i^+, \psi_i^- - \psi^0\right\}$, where

$$\psi_i^+ = \sup_{\pi\in\Pi} -\pi(x_t)^\top \frac{V_\infty e_i}{\gamma} + A_i(\pi), \psi_i^- = \sup_{\pi\in\Pi} \pi(x_t)^\top \frac{V_\infty e_i}{\gamma} + A_i(\pi), \text{ and } \psi_i^0 = \sup_{\pi\in\Pi} A_i(\pi)$$

for $A_i(\pi) = -\sum_{s=1}^{t-1}\pi(x_s)^\top D_i \hat{\Delta}_s - \sum_{s=t+1}^{T} 2\pi(x_s)^\top \boldsymbol{\epsilon}_{s,i}\boldsymbol{Z}_s$.

Return the closest $q \in \triangle_N$ to $[a_1, b_1] \times \ldots \times [a_N, b_N]$ in $\|\cdot\|_1$ norm.

The optimization objective in (10) has $\hat{\Delta}_t$ appearing in two places: the $q^\top\hat{\Delta}_t$ and $\mathcal{H}^t$, and hence the p_t optimization accounts for the worst case adversary action considering the loss introduced at round t as well as the potential future losses (from the $R_i(\mathcal{H}^t, \rho_t)$ terms). The algorithm picks q_t to mitigate the regret caused by the worst case p_t.

Since $\mathbf{Rel}(\mathcal{H}^t)$ is admissible, we may apply Lemma 2 to $\mathbf{Rel}(\emptyset)$ and obtain a regret bound. The random variable $\boldsymbol{Z}_t$ has maximum magnitude γ^{-1}, and hence we can optimize over γ to obtain a sub-linear regret, as stated in the following corollary.

Corollary 1. *The algorithm that plays the q_t defined in Theorem 2 has the regret bound*

$$\mathbb{E}[\mathcal{R}_t] \le \mathop{\mathbb{E}}_{\boldsymbol{Z}_{1:T},\boldsymbol{\epsilon}_{1:T}}\left[\sum_{i=1}^{N}\sup_{\pi\in\Pi}\sum_{t=1}^{T}\pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t\right] + TN\gamma.$$

If the policy class is finite, choosing $\gamma = (V_\infty \log(|\Pi|)/(2TN))^{\frac{1}{3}}$ produces

$$\mathbb{E}[\mathcal{R}_T] \le N\left(4V_\infty\log(|\Pi|)\right)^{\frac{1}{3}} T^{\frac{2}{3}}. \tag{11}$$

With details in Appendix B, the second claim follows from optimizing γ over the bound $\mathbb{E}_{\boldsymbol{Z}_{1:T},\boldsymbol{\epsilon}_{1:T}}\left[\sum_{i=1}^{N}\sup_{\pi\in\Pi}\sum_{t=1}^{T}\pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t\right] \le N\sqrt{2T\gamma^{-1}V_\infty\log(|\Pi|)}$.

4.4 Computation

At first glance, the relaxation algorithm, which samples a random rollout ρ_t then computes $\arg\min_{q\in\triangle_N}\sup_{p_t\in\triangle'_\mathcal{D}}\mathbb{E}_{\hat{\Delta}_t\sim p_t}\left[q^\top\hat{\Delta}_t + \sum_{i=1}^{N} R_i\left(\mathcal{H}^t, \rho_t\right)\right]$, does not see easy to compute. Fortunately, it is possible to exploit the structure of $\triangle'_\mathcal{D}$ and compute q_t^* using only $3N$ oracle calls per round. We give pseudocode in Algorithm 2 and specify every oracle query. The objective with the $\sup_{p_t}$ resolved is a convex function in q_t which has a local minimum for any q in the rectangle $[a, b] := [a_1, b_1] \times \ldots \times [a_N, b_N]$ and slope with constant magnitude in all coordinates outside of $[a, b]$. Hence, any $q \in \triangle_N$ with minimum $\|\cdot\|_1$ distance to $[a, b]$ will be optimal. Consider the case when $\sum_i a_i \le 1 \le \sum_i b_i$. With $q_i(x) = a_i\mathbb{1}\{x \le a_i\} + x\mathbb{1}\{a_i < x < b_i\} + b_i\mathbb{1}\{x \ge b_i\}$, we return an x such that $\sum_i q_i(x) = 1$; this is doable is $O(N)$ time because $q_i(x)$ is a increasing, piecewise linear function where the slope changes every time x passes some a_i or b_i; start $x_0 = a$ and increase x until $\sum_i q_i(x) = 1$. In the case where $\sum_i b_i \le 1$, one can perform regular water-filling starting b, and if $\sum_i a_i \ge 1$, one can perform water-draining starting from a. See Appendix C for details and proofs.

Lemma 3. *Algorithm 2 correctly calculates $q_t^*(\rho)$, has complexity $O(N)$, and requires only $3N$ oracle calls.*

5 Lower Bound

The previous algorithms only delivered fast $O(\sqrt{T})$ rates if pairwise observability holds. This section shows our upper bounds are tight: pairwise observability is necessary for the fast rate.

Theorem 3. *Consider a contextual partial monitoring game that is not pairwise observable. Then there exists a policy class and stochastic adversary such that any algorithm will incur expected regret* *w.r.t. the policy class and all fixed actions of at least*

$$\mathbb{E}[\mathcal{R}_T] \geq CT^{2/3},$$

where C is some constant depending on L and H only.

Such a lower bound cannot hold for arbitrary policy classes; if this were true, then picking the policy class of constant actions would contradict the lower bound for local observability.

The complete proof is presented in Appendix D, but the high level ideas are given here. The proof explicitly constructs a hard example. Let $x_t \sim \mathrm{Uniform}([0,1])$ (fortunately, we do not need to turn the distribution of x_t). Assume action 1 and 2 are not pairwise observable, and hence $\ell_1 - \ell_2 \notin \mathrm{Im}\,(S_1^\top) \oplus \mathrm{Im}\,(S_2^\top)$. This implies that there some $v \in \ker(S_1) \cap \ker(S_2)$ with $(\ell_1 - \ell_2)^\top v = 1$ (since we can scale v) because the kernel of a matrix X is the orthogonal complement of $\mathrm{Im}\,(X^\top)$. Also, $\mathbf{1} \in \boldsymbol{S}_1^\top$, so $v^\top \mathbf{1} = 0$.

Fix some $q_1 \in C_1$ and $q_2 \in \boldsymbol{C}_2$ and define $P_1(\epsilon) = (q_1 - \epsilon v)\mathbb{1}\{x \leq \frac{1}{2}\} + (q_2 - \epsilon v)\mathbb{1}\{x > \frac{1}{2}\}$ and $P_2(\epsilon) = P_1(-\epsilon)$. Under P_1, action 1 incurs slightly less loss and action 2 incurs slightly more. The key is that, for any $i, j \in \{1,2\}$, $S_i(q_j - \epsilon v) = S_i(q_j + \epsilon v)$, so the distribution of feedback symbols observed by the algorithm is exactly the same if actions 1 or 2 is played.

We define $\pi_1(x) = e_1\mathbb{1}\{x \leq {}^1\!/_2 + \beta_1\} + e_2\mathbb{1}\{x > {}^1\!/_2 + \beta_1\}$ and $\pi_2(x) = e_1\mathbb{1}\{x \leq {}^1\!/_2 - \beta_2\} - e_2\mathbb{1}\{x > {}^1\!/_2 - \beta_2\}$. Policy π_i has a bias towards action i (playing it 2β more), and hence π_i does better on P_i by $O(\epsilon(\beta + 1 + \beta_2))$. We show that β_1 and β_2 can be tuned to a problem dependent constant such that either policy still outperforms all actions by a constant. The feedback structure ensures that any algorithm receives the same feedback distribution when following π_1 or π_2, and hence the algorithm must play other actions to determine which policy is better. By our construction, doing so will add constant regret.

The strategies P_i are $O(\epsilon^2)$ apart in KL-divergence, which allows us to show that the strategies are hard to separate given the feedback from any action. Hence, the learner must balance playing suboptimal actions with learning which π_i is better. Setting $\epsilon = T^{-1/3}$ produces the lower bound.

6 Conclusion

This paper characterized the minimax regret for contextual partial monitoring for finite policy classes. We showed that pairwise observability is necessary and sufficient for the fast $O(\sqrt{T})$ rate. This result is surprising, since the noncontextual setting needs the significantly less strong notion of local observability for the $O(\sqrt{T})$ rate. Our lower bound implies that the relaxation algorithm is optimal for the local and global observability settings; this is the first known adversarial partial information setting where a relaxation algorithm obtains the minimax regret.

A few open problems remain. First, how does the complexity of the context affect the rate? Consider a game that is locally but not pairwise observable. If x_t is a constant (or if the policy class ignores it), then the contextual case reduces to the noncontextual case and a fast rate is achievable. However, if $x_t \sim \mathrm{Uniform}([0,1])$, then we showed that the fast rate is impossible. Can one obtain lower and upper bounds in terms of the complexity of the context and interpolate between these two regimes? Second, is it possible to obtain $O(\sqrt{T})$ rates with a relaxation algorithm? The particular form of the optimal q_t in Algorithm 2 suggests a way forward by using properties of Π to control a from below and thereby controlling the importance weights without needing uniform exploration.

References

[1] Jacob D Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. In *COLT*, 2009.

[2] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, pages 23–35, 2015.

[3] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[4] Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pages 305–319. Springer, 2012.

[5] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *COLT*, volume 2011, pages 133–154, 2011.

[6] Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[7] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[8] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

[9] Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient learning and auction design. *arXiv preprint arXiv:1611.01688*, 2016.

[10] Dean Foster and Alexander Rakhlin. No internal regret via neighborhood watch. In *Artificial Intelligence and Statistics*, pages 382–390, 2012.

[11] T. Lattimore and C. Szepesvari. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. *ArXiv e-prints*, May 2018.

[12] Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

[13] Jacob Marschak and William H Andrews. Random simultaneous equations and the theory of production. *Econometrica*, pages 143–205, 1944.

[14] Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory*, pages 208–223. Springer, 2001.

[15] Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2016.

[16] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and localize: From value to algorithms. *CoRR*, abs/1204.0870, 2012.

[17] Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3135–3143, 2016.

[18] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

A Proof of Theorem 1

For this section, $\mathcal{F}_t$ denotes the filtration generated by I_t and x_t, and $\mathbb{E}_t[\cdot] := \mathbb{E}_{I_t}[\cdot|\mathcal{F}]$ is the conditional expectation over the player's actions. Our analysis will borrow the following theorem:

Theorem 4 ([1]Theorem 2.1)**.** *The exponential weights algorithm using loss ℓ_t, which plays $q_t(i) \propto \exp(-\eta \sum_{s=1}^{t-1} \ell_t(i))$ has regret*

$$\sum_{t=1}^{T} (\ell_t)^\top (q_t - u) \leq \eta \sum_{t=1}^{T} (\|\ell_t\|_{q_t}^*)^2 + \frac{\log(K)}{\eta}, \tag{12}$$

where $\|\ell_t\|_{q_t}^*$ is the local norm defined by $\|\ell_t\|_{q_t}^* = \sqrt{(\ell_t)^\top \operatorname{diag}(q_t)\ell_t}$ and $u \in \triangle_N$ is any fixed distribution of actions.

Proof of Theorem 1. For this proof, define $\Pi(x_t)$ to be the matrix with columns $\pi_1(x_t), \ldots, \pi_K(x_t)$, which allows us to write the expected loss from a distribution of policies $w_t \in \triangle_K$ as $\Delta_t^\top \Pi(x_t) w_t$ and the strategy of Algorithm 1 as $q_t = (1 - N\gamma)\Pi(x_t)w_t + \gamma \mathbf{1}$.

Consider the exponential weights strategy that plays $w_t \propto \exp(-\eta \sum_{s=1}^{t-1} \Delta_s \Pi(x_s)^\top)$. The expectation of the guarantee from applying Theorem 4 with $\ell_t = \Pi(x_t)^\top \hat{\Delta}_t$ is

$$\begin{aligned}
\eta \sum_{t=1}^{T} \mathbb{E}\left[(\|\Pi(x_t)^\top \hat{\Delta}_t\|_{w_t}^*)^2\right] + \frac{\log(K)}{\eta} &\geq \mathbb{E}\left[\sum_{t=1}^{T} (\Pi(x_t)^\top \hat{\Delta}_t)^\top (w_t - u)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[(\Pi(x_t)^\top \hat{\Delta}_t)^\top (w_t - u)\right]\right] \\
&= \sum_{t=1}^{T} \Delta_t^\top \Pi(x_t) w_t - \sum_{t=1}^{T} \Delta_t^\top \Pi(x_t) u \\
&= \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top (L - \mathbf{1}\ell_{b_t}^\top)^\top \Pi(x_t) w_t - \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top (L - \mathbf{1}\ell_{b_t}^\top)^\top \Pi(x_t) u \\
&= \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top L^\top \Pi(x_t) w_t - \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top L^\top \Pi(x_t) u.
\end{aligned}$$

We can relate the last line to the loss of Algorithm 1 by the simple inequality

$$\sum_{t=1}^{T} e_{j_t}^\top L^\top \Pi(x_t) w_t - \sum_{t=1}^{T} e_{j_t}^\top L^\top q_t = \sum_{t=1}^{T} e_{j_t}^\top L^\top (N\gamma \Pi(x_t)^\top w_t - \gamma \mathbf{1}) \leq \gamma T N.$$

Combining the two inequalities above, we can show that

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &= \sum_{t=1}^{T} e_{j_t}^\top L^\top q_t - \min_u \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top L^\top \Pi(x_t) u \\
&\leq \sum_{t=1}^{T} e_{j_t}^\top L^\top \Pi(x_t) w_t - \min_u \sum_{t=1}^{T} \boldsymbol{e}_{j_t}^\top L^\top \Pi(x_t) u + \gamma T N \\
&\leq \eta \sum_{t=1}^{T} \mathbb{E}\left[(\|\Pi(x_t)^\top \hat{\Delta}_t\|_{w_t}^*)^2\right] + \frac{\log(K)}{\eta} + \gamma T N.
\end{aligned} \tag{13}$$

The analysis diverges depending on whether pairwise observability holds. First, assume that it does not hold. To easy notation, define the matrix

$$V_t(k) := (v_{1,b_t,k}, \ldots, v_{N,b_t,k})$$

so that $\Delta_t = \sum_{k=1}^N V_t(k)^\top S_k e_{j_t}$ and $\hat{\Delta}_t = V_t(I_t)Y_t/q_t(I_t)$. The bound on the observability vectors implies that $V_\infty \geq \max_i \|V(i)\|_\infty$. We can bound the first term in (13) by

$$
\begin{aligned}
\mathbb{E}_t\left[(\|\Pi(x_t)^\top \hat{\Delta}_t\|_{w_t}^*)^2\right] &= \mathbb{E}_t\left[\hat{\Delta}_t^\top \Pi(x_t)\operatorname{diag}(w_t)\Pi(x_t)^\top \hat{\Delta}_t\right] \\
&= \mathbb{E}_t\left[\sum_{k=1}^K w_t(k)(\hat{\Delta}_t^\top \pi_k(x_t))^2\right] \\
&= \mathbb{E}_t\left[\sum_{k=1}^K w_t(k)\left(\frac{V(I_t)^\top \pi_k(x_t)}{q_t(I_t)}\right)^2\right] \\
&\leq \mathbb{E}_t\left[\frac{V_\infty V(I_t)^\top \Pi(x_t) w_t}{q_t(I_t)^2}\right] \\
&\leq V_\infty^2 \gamma^{-1}.
\end{aligned}
$$

where, in the last inequality, we used $\Pi(x_t)w_t \in \triangle_N$ and $q_t \geq \gamma$. Combining with (13), we have

$$
\mathbb{E}[\mathcal{R}_t] \leq \frac{\eta}{\gamma} T V_\infty^2 + \frac{\log(K)}{\eta} + \gamma T N.
$$

Optimizing over the parameters and setting $\eta = N^{-\frac{1}{3}}\left(\frac{\log(K)}{V_\infty T}\right)^{\frac{2}{3}}$ and $\gamma = \left(\frac{V_\infty^2 \log(K)}{N^2 T}\right)^{\frac{1}{3}}$ yields

$$
\mathbb{E}[\mathcal{R}_T] \leq 3\left(N V_\infty^2 \log(K)\right)^{\frac{1}{3}} T^{\frac{2}{3}}.
$$

Now, consider the pairwise observability case where $\gamma = 0$. We may always choose $V_{i,j} = \{i, j\}$, and so the offset loss estimate will have support on I_t and b_t only. This implies that

$$
\hat{\Delta}_t(i) = \mathbb{1}\{I_t = i\}\frac{v_{i,b,I_t}^\top Y_t}{q_t(I_t)} + \mathbb{1}\{I_t = b_t\}\frac{v_{i,b_t,I_t}^\top Y_t}{q_t(I_t)}.
$$

Since $\pi_k(x_t)$ is a unit vector, we have $\mathbb{1}\{\pi_k(x_t) = I_t\} = \pi_k(x_t)^\top e_{I_t}$, so we can write

$$
\pi_k(x_t)^\top \hat{\Delta}_t = \pi_k(x_t)^\top e_{I_t}\frac{v_{I_t,b_t,I_t}^\top Y_t}{q_t(I_t)} + \mathbb{1}\{I_t = b_t\}\frac{v_{b_t,\pi_k(x_t),I_t}^\top Y_t}{q_t(I_t)}.
$$

The variance term can therefore be bounded by

$$
\begin{aligned}
\mathbb{E}_t\left[(\|\Pi(x_t)^\top \hat{\Delta}_t\|_{w_t}^*)^2\right] &= \mathbb{E}\left[\sum_{k=1}^K w_t(k)\left(\pi_k(x_t)^\top e_{I_t}\frac{v_{\pi_k(x_t),b_t,I_t}^\top Y_t}{q_t(I_t)} + \mathbb{1}\{I_t = b_t\}\frac{v_{b_t,\pi_k(x_t),I_t}^\top Y_t}{q_t(b_t)}\right)^2\right] \\
&\leq V_\infty^2 \mathbb{E}\left[\sum_{k=1}^K w_t(k)\left(\frac{\pi_k(x_t)^\top e_{I_t}}{q_t(I_t)} + \frac{\mathbb{1}\{I_t = b_t\}}{q_t(b_t)}\right)^2\right] \\
&= V_\infty^2 \mathbb{E}\left[\sum_{k=1}^K w_t(k)\left(\frac{\pi_k(x_t)^\top e_{I_t}}{q_t(I_t)^2} + 2\frac{\pi_k(x_t)^\top e_{I_t}\mathbb{1}\{I_t = b_t\}}{q_t(I_t)q_t(b_t)} + \frac{\mathbb{1}\{I_t = b_t\}}{q_t(b_t)^2}\right)\right] \\
&\leq V_\infty^2 \mathbb{E}\left[\frac{q_t^\top e_{I_t}}{q_t(I_t)^2} + 2\frac{q_t^\top e_{I_t}\mathbb{1}\{I_t = b_t\}}{q_t(I_t)q_t(b_t)} + \frac{\mathbb{1}\{I_t = b_t\}}{q_t(b_t)^2}\right] \\
&\leq V_\infty^2 \mathbb{E}\left[\frac{q_t(I_t)}{q_t(I_t)^2} + 2\frac{q_t(I_t)\mathbb{1}\{I_t = b_t\}}{q_t(I_t)q_t(b_t)} + \frac{\mathbb{1}\{I_t = b_t\}}{q_t(b_t)^2}\right] \\
&= V_\infty^2\left(3 + \frac{1}{q_t(b_t)}\right).
\end{aligned}
$$

Since we choose $b_t = \arg\max_i q_t(i)$, we must have $q_t(b_t) \geq 1/N$ and the previous term is bounded by $V_\infty^2(N+3)$. Combining this inequality with (13) and setting $\gamma = 0$ produces

$$
\mathbb{E}[\mathcal{R}_t] \leq \eta T V_\infty^2 (N+3) + \frac{\log(K)}{\eta}.
$$

The theorem statement follows from setting

$$\eta = \sqrt{\frac{\log(K)}{TV_\infty^2(N+3)}}.$$

$\square$

B Proofs for Relaxations

Lemma 4. *For any random vector Z_t with $\mathbb{E}[(Z_t^\top e_i)^2] \le W$, $\epsilon_{t,i}$ i.i.d. Rademacher random variables, and $\boldsymbol{\epsilon}_t$ denoting the collection across i of $\boldsymbol{\epsilon}_{t,i} = \mathrm{diag}(e_i \epsilon_{t,i})$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{Z}_{1:T},\boldsymbol{\epsilon}_{1:T}}\left[\sum_{i=1}^N \sup_{\pi\in\Pi}\sum_{t=1}^T \pi(x_t)^\top \boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t\right] \le N\sqrt{2TW\log(|\Pi|)}. \tag{14}$$

Proof. This reasoning is a small refinement of the proof of Lemma 2 in [17]. We evaluate

$$\begin{aligned}
\sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T},\boldsymbol{\epsilon}_{1:T}}\left[\sup_{\pi\in\Pi}\sum_{t=1}^T \pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t\right] &= \sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\frac{1}{\lambda}\mathbb{E}_{\boldsymbol{\epsilon}_{1:T}}\left[\log\left(\sup_{\pi\in\Pi} e^{\lambda\sum_{t=1}^T \pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right)\right]\\
&= \sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\left[\frac{1}{\lambda}\mathbb{E}_{\boldsymbol{\epsilon}_{1:T}}\left[\log\left(\sup_{\pi\in\Pi} e^{\lambda\sum_{t=1}^T \pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right)\right]\right]\\
&\le \sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\left[\frac{1}{\lambda}\log\left(\mathbb{E}_{\boldsymbol{\epsilon}_{1:T}}\left[\sum_{\pi\in\Pi} e^{\lambda\sum_{t=1}^T \pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right]\right)\right]\\
&\le \sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\left[\frac{1}{\lambda}\log\left(\sum_{\pi\in\Pi}\prod_{t=1}^T \mathbb{E}_{\boldsymbol{\epsilon}_{1:T}}\left[e^{\lambda\pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right]\right)\right].
\end{aligned}$$

We have the inequality

$$\mathbb{E}_{\boldsymbol{\epsilon}_t}\left[e^{\lambda\pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right] = \mathbb{E}_{\epsilon_t}\left[e^{\lambda(\pi(x_t)^\top e_i)\epsilon_t Z_{t,i}}\right] = \frac{e^{\lambda(\pi(x_t)^\top e_i)Z_{t,i}} + e^{-\lambda(\pi(x_t)^\top e_i)Z_{t,i}}}{2} \le e^{\frac{1}{2}\lambda^2 Z_{t,i}^2}.$$

Combining with the above expression produces

$$\begin{aligned}
\sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\left[\frac{1}{\lambda}\log\left(\sum_{\pi\in\Pi}\prod_{t=1}^T \mathbb{E}_{\boldsymbol{\epsilon}_{1:T}}\left[e^{\lambda\pi(x_t)^\top\boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t}\right]\right)\right] &\le \sum_{i=1}^N \mathbb{E}_{\boldsymbol{Z}_{1:T}}\left[\frac{1}{\lambda}\log\left(\sum_{\pi\in\Pi}\prod_{t=1}^T e^{\frac{\lambda^2 Z_{t,i}^2}{2}}\right)\right]\\
&\le N\frac{1}{\lambda}\log(|\Pi|) + NT\lambda\mathbb{E}_{\boldsymbol{Z}}\left[\frac{\lambda Z_{t_i}^2}{2}\right]\\
&\le \frac{N}{\lambda}\log(|\Pi|) + \frac{N}{2}WT\lambda
\end{aligned}$$

Setting $\lambda = \sqrt{2\log(|\Pi|)/WT}$ finishes the proof. $\square$

Proof of Corollary 1. Lemma 4 with $W = V_\infty\gamma^{-1}$ yields a bound of

$$\mathop{\mathbb{E}}_{\boldsymbol{Z}_{1:T},\boldsymbol{\epsilon}_{1:T}}\left[\sum_{i=1}^N \sup_{\pi\in\Pi}\sum_{t=1}^T \pi(x_t)^\top \boldsymbol{\epsilon}_{t,i}\boldsymbol{Z}_t\right] \le N\sqrt{2TV_\infty\gamma^{-1}\log(|\Pi|)},$$

which produces the theorem with the given value of γ. $\square$

Proof of Theorem 2. The base case is easy. Using the convexity of supremum and the unbiasedness of $\hat{\Delta}_t$,

$$\begin{aligned}
\mathbb{E}_{I_{1:T},\hat{Z}_{1:T}}\left[\mathbf{Rel}(\mathcal{H}^T)\right] &= \mathbb{E}_{I_{1:T},\hat{Z}_{1:T}}\left[\sum_{i=1}^{N}\sup_{\pi\in\Pi} -\sum_{s=1}^{T}\pi(x_t)^\top D_i\hat{\Delta}_t\right] \\
&\geq \mathbb{E}_{I_{1:T},\hat{Z}_{1:T}}\left[\sup_{\pi\in\Pi} -\sum_{i=1}^{N}\sum_{s=1}^{T}\pi(x_t)^\top D_i\hat{\Delta}_t\right] \\
&= \mathbb{E}_{I_{1:T},\hat{Z}_{1:T}}\left[\sup_{\pi\in\Pi} -\sum_{s=1}^{T}\pi(x_t)^\top \hat{\Delta}_t\right] \\
&\geq \sup_{\pi\in\Pi}\sum_{s=1}^{T}\mathbb{E}_{I_{1:T},\hat{Z}_{1:T}}\left[-\pi(x_t)^\top\hat{\Delta}_t\right] \\
&= \sup_{\pi\in\Pi} -\sum_{s=1}^{T}\pi(x_t)^\top\Delta_t.
\end{aligned}$$

We now check the inductive step. We define $\rho = (x_t, \boldsymbol{\epsilon}_t, \boldsymbol{Z}_t)_{t+1:T}$ to the collection of random variables in the relaxation. Recall that our aim is to prove admissibility of the strategy $q_t(\rho) = (1-N\gamma)q_t^*(\rho) + \gamma\mathbf{1}$ where $q_t^*(\rho)$ was defined by

$$q_t^*(\rho) = \operatorname*{arg\,min}_{q\in\triangle_N}\sup_{p_t\in\triangle'_{\mathcal{D}}}\mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[q^\top\hat{\Delta}_t + \sum_{i=1}^{N}R_i\left(\mathcal{H}^t,\rho\right)\right].$$

By construction, no entry of $q_t(\rho)$ is below γ. It is then easy to check that, for $q_t^* = \mathbb{E}_\rho[q_t^*(\rho)]$ and $q_t = \mathbb{E}_\rho[q_t(\rho)]$,

$$\mathbb{E}_{I_t\sim q_t}\left[\Delta_t^\top e_{I_t}\right] = \Delta_t^\top q_t \leq \Delta_t^\top q_t^* + N\gamma = \mathbb{E}_{I_t\sim q_t}\left[\hat{\Delta}_t(I_t,j_t,q_t)^\top q_t^*\right] + N\gamma,$$

where we have been explicit that $\hat{\Delta}_t(I_t,j_t,q_t)$ is a random variable that depends on the player actions and the q_t (since $\boldsymbol{Z}_t$'s distribution is a function of q_t).

For every fixed x_t, we can apply the above inequality and unpack the definition of the relaxation to find

$$\begin{aligned}
\sup_{j_t}\mathop{\mathbb{E}}_{I_t\sim q_t}\left[e_{I_t}^\top\hat{\Delta}_t(I_t,j_t,q_t) + \mathbf{Rel}(\mathcal{H}^t)\right] &= \sup_{j_t}\mathop{\mathbb{E}}_{I_t\sim q_t}\left[e_{I_t}^\top\hat{\Delta}_t(I_t,j_t,q_t) + \sum_{i=1}^{N}\mathop{\mathbb{E}}_{\rho}\left[R_i\left(\mathcal{H}^t,\rho\right)\right]\right] \\
&\quad + (T-t)N\gamma \\
&\leq \sup_{j_t}\mathop{\mathbb{E}}_{I_t\sim q_t}\left[(q_t^*)^\top\hat{\Delta}_t(I_t,j_t,q_t) + \sum_{i=1}^{N}\mathop{\mathbb{E}}_{\rho}\left[R_i\left(\mathcal{H}^t,\rho\right)\right]\right] \\
&\quad + (T-t+1)N\gamma.
\end{aligned}$$

Defining $A_i(\pi) = -\sum_{s=1}^{t-1}\pi(x_s)^\top D_i\hat{\Delta}_s - \sum_{s=t+1}^{T}2\pi(x_s)\boldsymbol{\epsilon}_{s,i}\boldsymbol{Z}_s$, the previous line is equal to

$$\begin{aligned}
&\sup_{j_t}\mathop{\mathbb{E}}_{I_t\sim q_t}\left[(q_t^*)^\top\hat{\Delta}_t(I_t,j_t,q_t) + \mathop{\mathbb{E}}_{\rho}\left[\sum_{i=1}^{N}\sup_{\pi} -\pi(x_t)^\top\hat{\Delta}_t(I_t,j_t,q_t) + A_i(\pi)\right]\right] + (T-t+1)N\gamma \\
&\quad = \sup_{j_t}\sum_{i=1}^{N}\mathop{\mathbb{E}}_{I_t\sim q_t}\left[\mathop{\mathbb{E}}_{\rho}\left[q_t^*(\rho)^\top\hat{\Delta}_t(I_t,j_t,q_t) + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i\hat{\Delta}_t(I_t,j_t,q_t) + A_i(\pi)\right]\right] \\
&\quad + (T-t+1)N\gamma.
\end{aligned}$$

This optimization is intractable without strong assumptions on Π; the j_t optimization needs to account for how the supremum over the policy class will be affected. Therefore, we reduce the constraints on the adversary to relax the problem by allowing play of distributions over $\hat{\Delta}_t$ instead of constraining the $\hat{\Delta}_t$ to correspond to a specific choice of I_t, j_t, and q_t. However, we carefully defined $\hat{\Delta}_t(I_t, j_t, q_t)$ so that it would have coordinate-wise sparseness and only expand the plays of the adversary to distributions with the same sparseness.

Specifically, every coordinate of $\hat{\Delta}_t(I_t, j_t, q_t)$ has large probability of being zero: if we fix q_t by conditioning on ρ, then

$$\begin{aligned}
P\left(\hat{\Delta}_t(I_t, j_t, q_t)^\top e_i = 0\right) &= \mathbb{E}_\rho\left[\mathbb{E}_{I_t\sim q_t(\rho)}\left[P\left(\hat{\Delta}_t(I_t, j_t, q_t)^\top e_i = 0 \middle| \rho, I_t\right)\right]\right] \\
&= \mathbb{E}_\rho\left[\mathbb{E}_{I_t\sim q_t(\rho)}\left[1 - \frac{\gamma |e_i^\top V(I_t)^\top Y_t|}{V_\infty q_t(\rho)(I_t)}\right]\right] \\
&\geq \mathbb{E}_\rho\left[\mathbb{E}_{I_t\sim q_t(\rho)}\left[1 - \frac{\gamma}{q_t(\rho)(I_t)}\right]\right] \\
&= \mathbb{E}_\rho\left[\sum_{i=1}^N q_t(\rho)(i)\left(1 - \frac{\gamma}{q_t(\rho)(i)}\right)\right] \\
&= 1 - N\gamma.
\end{aligned}$$

Therefore, the distribution of $\hat{\Delta}_t(I_t, j_t, q_t)$ is always in $\triangle'_{\mathcal{D}}$, and we obtain an upper bound by allowing the adversary to play distributions over a random variable named $\hat{\Delta}_t$ with distribution in $\triangle'_{\mathcal{D}}$. With this substitution, I_t and j_t no longer appear in the expression. Suppressing the $(T - t + 1)N\gamma$ term, we have

$$\begin{aligned}
&\sup_{j_t} \sum_{i=1}^N \mathop{\mathbb{E}}_{I_t\sim q_t}\left[\mathop{\mathbb{E}}_{\rho}\left[q_t^*(\rho)^\top \hat{\Delta}_t(I_t, j_t, q_t) + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t(I_t, j_t, q_t) + A_i(\pi)\right]\right] \\
&\leq \sup_{p_t\in\triangle'_{\mathcal{D}}} \mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[\sum_{i=1}^N \mathop{\mathbb{E}}_{\rho}\left[q_t^*(\rho)^\top \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right]\right] \\
&\leq \mathop{\mathbb{E}}_{\rho}\left[\sup_{p_t\in\triangle'_{\mathcal{D}}} \sum_{i=1}^N \mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[q_t^*(\rho)^\top \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right]\right].
\end{aligned}$$

Our ability to move the expectation over ρ to the outside allows us to obtain the same bound in expectation by sampling a single ρ and playing $q_t(\rho)$ instead of calculating the infimum over q.

For the remainder, fix some ρ. We defined

$$q_t^*(\rho) = \mathop{\arg\min}_{q\in\triangle_N} \sup_{p_t\in\triangle'_{\mathcal{D}}} \mathbb{E}_{\hat{\Delta}_t\sim p_t}\left[q^\top \hat{\Delta}_t + \sum_{i=1}^N \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right],$$

and so

$$\begin{aligned}
&\sup_{p_t\in\triangle'_{\mathcal{D}}} \sum_{i=1}^N \mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[q_t^*(\rho)^\top \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right] \\
&= \inf_{q_t} \sup_{p_t\in\triangle'_{\mathcal{D}}} \sum_{i=1}^N \mathop{\mathbb{E}}_{\hat{\Delta}_t\sim p_t}\left[q_t^\top \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right].
\end{aligned}$$

We continue bounding this saddle point problem from above. Noting that the objective is linear in both q_t and p_t, we may perform a min-max swap:

$$\begin{aligned}
&\inf_{q_t} \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\hat{\Delta}_t \sim p_t} \left[q_t^\top \hat{\Delta}_t + \sup_{\pi \in \Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi) \right] \\
&= \sup_{p_t \in \triangle'_{\mathcal{D}}} \inf_{q_t} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\hat{\Delta}_t \sim p_t} \left[q_t^\top \hat{\Delta}_t + \sup_{\pi \in \Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi) \right] \\
&\leq \sup_{p_t \in \triangle'_{\mathcal{D}}} \inf_{j} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\hat{\Delta}_t \sim p_t} \left[e_j^\top \hat{\Delta}_t + \sup_{\pi \in \Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi) \right] \\
&= \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \inf_{j} \mathop{\mathbb{E}}_{\hat{\Delta}_t \sim p_t} \left[e_j^\top \hat{\Delta}_t \right] + \mathop{\mathbb{E}}_{\hat{\Delta}_t \sim p_t} \left[\sup_{\pi \in \Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi) \right].
\end{aligned}$$

Because π is deterministic, we must have

$$\min_i \mathbb{E}_{\hat{\Delta}_t \sim p_t}[e_i^\top \hat{\Delta}_t] \leq \mathbb{E}_{\hat{\Delta}_t \sim p_t}[\pi(x_t)^\top \hat{\Delta}_t] \leq \mathbb{E}_{\hat{\Delta}_t \sim p_t}[\pi(x_t)^\top D_i \hat{\Delta}_t],$$

which allows us to upper bound the previous expression. Performing the usual symmeterization (using ϵ as a single Rademacher random variable) yields

$$\begin{aligned}
&\sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t} \left[\sup_{\pi \in \Pi} A_i(\pi) + \min_j \mathbb{E}_{\hat{\Delta}'_t \sim p_t} \left[e_j^\top \hat{\Delta}'_t \right] - \pi(x_t)^\top D_i \hat{\Delta}_t \right] \\
&\leq \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t} \left[\sup_{\pi \in \Pi} A_i(\pi) + \mathbb{E}_{\hat{\Delta}'_t \sim p_t} \left[\pi(x_t)^\top D_i \hat{\Delta}'_t \right] - \pi(x_t)^\top D_i \hat{\Delta}_t \right] \\
&\leq \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t, \hat{\Delta}'_t \sim p_t} \left[\sup_{\pi \in \Pi} A_i(\pi) + \pi(x_t)^\top D_i \left(\hat{\Delta}'_t - \hat{\Delta}_t \right) \right] \\
&= \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t, \hat{\Delta}'_t \sim p_t, \epsilon_i} \left[\sup_{\pi \in \Pi} A_i(\pi) + \epsilon_i \pi(x_t)^\top D_i \left(\hat{\Delta}'_t - \hat{\Delta}_t \right) \right] \\
&\leq \sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t, \epsilon_i} \left[\sup_{\pi \in \Pi} A_i(\pi) + 2\epsilon_i \pi(x_t)^\top D_i \hat{\Delta}_t \right].
\end{aligned}$$

Since $D_i \hat{\Delta}_t = e_i \hat{\Delta}_t(i)$, each term in the sum only involves one coordinate if $\hat{\Delta}_t$. Therefore, it is without loss of generality to assume that p_t is a product distribution. Using $\triangle_{\{0, V_\infty \gamma^{-1}\}}$ to denote distributions over the singletons 0 and $V_\infty \gamma^{-1}$, define

$$\triangle'_1 := \left\{ p \in \triangle_{\{0, V_\infty \gamma^{-1}\}} : p(0) \geq 1 - N\gamma \right\}$$

and observe that if $\hat{\Delta}_t \sim p_t \in \triangle_{\mathcal{D}'}$, then $\epsilon_i X_i e_t \overset{\mathcal{L}}{=} \hat{\Delta}_t(i)$ for some $X_i \sim p_i \in \triangle'_1$. We then have

$$\begin{aligned}
\sup_{p_t \in \triangle'_{\mathcal{D}}} \sum_{i=1}^{N} \mathbb{E}_{\hat{\Delta}_t \sim p_t, \epsilon} \left[\sup_{\pi \in \Pi} A_i(\pi) + 2\epsilon \pi(x_t)^\top D_i \hat{\Delta}_t \right] &= \sup_{p_1, \ldots, p_N \in \triangle'_1} \sum_{i=1}^{N} \mathbb{E}_{X_i \sim p_i, \epsilon_{1:N}} \left[\sup_{\pi \in \Pi} A_i(\pi) + 2\pi(x_t)^\top e_i \epsilon_i X_i \right] \\
&= \sum_{i=1}^{N} \sup_{p_i \in \triangle'_1} \mathbb{E}_{X_i \sim p_i, \epsilon_i} \left[\sup_{\pi \in \Pi} A_i(\pi) + 2\pi(x_t)^\top e_i \epsilon_i X_i \right].
\end{aligned}$$

We will now argue that a witness to the supremum of

$$\sup_{p_i \in \triangle'_1} \mathbb{E}_{X_i \sim p_i, \epsilon_i} \left[\sup_{\pi \in \Pi} A_i(\pi) + 2\pi(x_t)^\top e_i \epsilon_i X_i \right]$$

is the distribution that puts mass $N\gamma$ on $V_\infty \gamma^{-1}$ and the rest on 0. Define the convex function $g(x) := \sup_{\pi \in \Pi} A_i(\pi) + 2\pi(x_t)^\top e_i x$. The expectation $\mathbb{E}_\epsilon \left[g(\epsilon x) \right]$ is increasing on $x \geq 0$. To see this, consider some $0 \leq a < b$. We can write $a = \theta b + (1-\theta)(-b)$ for some θ and use the definition of convexity to conclude

$$\mathbb{E}_\epsilon\left[g(\epsilon a)\right] = \frac{g(a)+g(-a)}{2} \leq \frac{\theta g(b) + (1-\theta)g(-b) + (1-\theta)g(b) + \theta g(-b)}{2} = \mathbb{E}_\epsilon\left[g(\epsilon b)\right].$$

Since $\mathbb{E}_\epsilon\left[g(\epsilon x)\right]$ is increasing, the supremum of p_i puts maximum mass on $V_\infty \gamma^{-1}$. Hence, defining the random vector Z_t with elements $P(Z_{t,i} = 0) = 1 - N\gamma$ and $P(Z_{t,i} = V_\infty \gamma^{-1}) = N\gamma$, we have

$$\begin{aligned}\sup_{p \in \triangle_1'} \mathbb{E}_{X_i \sim p, \epsilon_i}\left[\sup_{\pi\in\Pi} A_i(\pi) + 2\pi(x_t)^\top e_i \epsilon_i X_i\right] &= \mathbb{E}_{\epsilon_i}\left[\sup_{\pi\in\Pi} A_i(\pi) + 2\epsilon_i \pi(x_t)^\top D_i \boldsymbol{Z}_t\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}_t}\left[\sup_{\pi\in\Pi} A_i(\pi) + 2\pi(x_t)^\top \boldsymbol{\epsilon}_{t,i} \boldsymbol{Z}_t\right].\end{aligned}$$

In total, we have shown that, for every fixed x_t and ρ, playing $q_t(\rho)$ allows the bound

$$\begin{aligned}\sup_{j_t} \mathop{\mathbb{E}}_{I_t \sim q_t}\left[e_{I_t}^\top \Delta_t + \mathbf{Rel}(\mathcal{H}^t)\right] &\leq \mathbb{E}_\rho\left[\mathbb{E}_{\epsilon_t, \boldsymbol{Z}_t}\left[\sum_{i=1}^N \sup_{\pi\in\Pi} A_i(\pi) + 2\pi(x_t)^\top \boldsymbol{\epsilon}_{t,i} \boldsymbol{Z}_t\right]\right] + (T-t+1)N\gamma \\ &= \mathbf{Rel}(\mathcal{H}^{t-1}),\end{aligned}$$

as required. □

C Proof of Lemma 3

Proof. Recall that the relaxation algorithm sampled ρ_t then calculated the $q_t(\rho)$ that minimized

$$\sup_{p_t \in \triangle_{\mathcal{D}}'} \sum_{i=1}^N \mathop{\mathbb{E}}_{\hat{\Delta}_t(i) \sim p_i}\left[q^\top D_i \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right],$$

where $A_i(\pi) = -\sum_{s=1}^{t-1} \pi(x_s)^\top D_i \hat{\Delta}_s - \sum_{s=t+1}^T 2\pi(x_s)\boldsymbol{\epsilon}_{s,i} \boldsymbol{Z}_s$ and the $(T-t)N\gamma$ term is suppressed. This optimization decomposes over coordinates of $\hat{\Delta}_t$ and therefore over marginals of p_t. Since $\triangle_{\mathcal{D}}'$ only puts support on vectors with coordinates in $\{-V_\infty\gamma^{-1}, 0, V_\infty\gamma^{-1}\}$, we can fully parameterize the problem with $p_i^+ := p_i(\cdot = V_\infty\gamma^{-1})$ and $p_i^- = p_i(\cdot = -V_\infty\gamma^{-1})$ for $i = 1, \ldots, N$. The definition of $\triangle_{\mathcal{D}}'$ implies that $p_i^+ \leq \gamma$ and $\pi_i^- \leq \gamma$. The objective becomes

$$\begin{aligned}&\sup_{p_t \in \triangle_{\mathcal{D}}'} \sum_{i=1}^N \mathop{\mathbb{E}}_{\hat{\Delta}_t(i) \sim p_i}\left[q^\top D_i \hat{\Delta}_t + \sup_{\pi\in\Pi} -\pi(x_t)^\top D_i \hat{\Delta}_t + A_i(\pi)\right] \\ &= \sum_{i=1}^N \sup_{p_i^+ \leq \gamma, p_i^- \leq \gamma} \frac{V_\infty q_i}{\gamma}(p_i^+ - p_i^-) + \mathop{\mathbb{E}}_{\hat{\Delta}_t(i) \sim p_i}\left[\sup_{\pi\in\Pi} -\pi(x_t)^\top e_i \hat{\Delta}_t(i) + A_i(\pi_i)\right] \\ &= \sum_{i=1}^N \sup_{p_i^+ \leq \gamma, p_i^- \leq \gamma} \frac{V_\infty q_i}{\gamma}(p_i^+ - p_i^-) + p_i^+\left(\sup_{\pi\in\Pi} -\pi(x_t)^\top e_i \frac{V_\infty}{\gamma} + A_i(\pi_i)\right) \\ &\quad + p_i^-\left(\sup_{\pi\in\Pi} \pi(x_t)^\top e_i \frac{V_\infty}{\gamma} + A_i(\pi_i)\right) + (1 - p_i^+ - p_i^-)\left(\sup_{\pi\in\Pi} A_i(\pi_i)\right),\end{aligned}$$

and we can immediately see that we only require invoking the oracle $3N$ times to evaluate

$$\psi_i^+ := \sup_{\pi\in\Pi} -\pi(x_t)^\top \frac{V_\infty e_i}{\gamma} + A_i(\pi), \psi_i^- := \sup_{\pi\in\Pi} \pi(x_t)^\top \frac{V_\infty e_i}{\gamma} + A_i(\pi), \text{ and } \psi_i^0 := \sup_{\pi\in\Pi} A_i(\pi)$$

for $i = 1, \ldots, N$. In terms of these quantities, the objective becomes

$$\sum_{i=1}^N \sup_{p_i^+ \leq \gamma, p_i^- \leq \gamma}\left(p_i^+\left(\psi_i^+ + \frac{V_\infty q_i}{\gamma} - \psi_i^0\right) + p_i^-\left(\psi_i^- - \frac{V_\infty q_i}{\gamma} - \psi_i^0\right) + \psi_i^0\right). \tag{15}$$

Since p_i^+ and p_i^- are bounded by γ, the supremum will be at $p_i^+ = \gamma \mathbb{1}\left\{\left(\psi_i^+ + \frac{V_\infty q_i}{\gamma}\right) > \psi_i^0\right\}$ and $p_i^- = \gamma \mathbb{1}\left\{\left(\psi_i^- - \frac{V_\infty q_i}{\gamma}\right) > \psi_i^0\right\}$. Hence, (15) evaluates to

$$N\psi_0 + \sum_{i=1}^{N} \max\left\{-\gamma(\psi_i^0 - \psi_i^+) + V_\infty q_i, 0\right\} + \max\left\{\gamma(\psi_i^- - \psi_i^0) - V_\infty q_i, 0\right\}.$$

The positivity of $\pi(x_t)$ ensures that $\psi_i^+ \leq \psi_i^0 \leq \psi_i^-$ which implies that $(\psi^0 - \psi_i^+) \geq 0$ and $(\psi_i^- - \psi^0) \geq 0$. Since the $\max$ switches at $\frac{\gamma}{V_\infty}(\psi^0 - \psi_i^+)$ and the second at $\frac{\gamma}{V_\infty}(\psi_i^- - \psi^0)$, the minimizer of q_i is between these two values where the slope vanishes. If $\psi_i^- - \psi^0 \leq \psi^0 - \psi_i^+$, then the minimum is 0; otherwise, it is $\gamma(\psi_i^+ + \psi_i^- - 2\psi_0)$.

By defining $a_i := \frac{\gamma}{V_\infty} \min\left\{\psi^0 - \psi_i^+, \psi_i^- - \psi^0\right\}$ and $b_i := \frac{\gamma}{V_\infty} \max\left\{\psi^0 - \psi_i^+, \psi_i^- - \psi^0\right\}$, we can compactly write the objective as

$$\sum_{i=1}^{N} V_\infty \max\{q_i - a_i, 0\} + V_\infty \max\{b_i - q_i, 0\} + \max\{\gamma(\psi_i^+ + \psi_i^- - 2\psi_0), 0\}.$$

Let A_t be the rectangle of $\mathbb{R}^N$ with ith coordinate in $[a_i, b_i]$. If A_t has has nonempty intersection with $\triangle_N$, then any point in the intersection is optimal. Otherwise, we can exploit the fact that the objective value at $a_i - \epsilon$ is exactly the same as the objective value at $b_i + \epsilon$. Hence, the value of any point x is the L_1 distance to A_t.

There are three cases to consider. First, if $\sum_i a_i \leq 1 \leq \sum_i b_i$, then any q with $\sum_i q_i = 1$ and $q_i \in [a_i, b_i]$ is optimal. In particular, we can select the q with by a constrained water-filling algorithm as follows. Define $q(x)$ to have coordinates

$$q_i(x) = \begin{cases} a_i & \text{if } x \leq a_i, \\ x & \text{if } a_i \leq x \leq b_i, \text{ and} \\ b_i & \text{if } x \leq b_i. \end{cases}$$

Select $q^* = q(x_{fill})$ where $x_{fill} := \max\{x : \sum_i q_i(x) \leq 1\}$. Because $\sum_i a_i \leq 1 \leq \sum_i b_i$, such an x_{fill} must exist. We can find x_{fill} easily since $\sum_i q_i(x)$ is a piecewise linear increasing function in x with at most $2N$ points where the slope changes.

The second case is $\sum_i b_i < 1$, which implies that A_t does not intersect $\triangle_N$. In this case, the suboptimality is exactly $V_\infty \|q - b\|_1$, which can be minimized a water-filling algorithm as described above with no upper limit to the coordinates of q_i. The final case is $\sum_i a_i > 1$, which results in an inverse water-filling algorithm.

$\square$

D Proof of Lower Bound

For readability, we break up the proof into four sections.

D.1 Defining the alternatives

By assumption, there exist two non-degenerate actions that are not pairwise observable. Without loss of generality, assume that these are actions 1 and 2. Define $S_{1,2} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$. Since actions 1 and 2 are not pairwise observable, we must have $\ell_1 - \ell_2 \notin \text{Im}\,(S_{1,2}^\top)$. Let w be the orthogonal projection of $\ell_1 - \ell_2$ onto $\text{Im}\,(S_{1,2}^\top)$. Since $\mathbb{R}^M = \text{Im}\,(S_{1,2}^\top) \oplus \ker(S_{1,2})$, we must have $w \in \ker(S_{1,2})$. Let v be a scaling of w such that $(\ell_2 - \ell_1)^\top v = 1$, and note that $\mathbf{1} \in \text{Im}\,(S_{1,2})$ implies that $v^\top \mathbf{1} = 0$ (where $\mathbf{1}$ is the all ones vector). To summarize, there exists a vector v with:

1. $v^\top \mathbf{1} = 0$,
2. $(\ell_1 - \ell_2)^\top v = 1$, and

3. $S_1 v = S_2 v = 0$.

We now exploit the existence of such a v to design two adversary strategies, P_1 and P_2, and a policy class Π, such that P_1 and P_2 are difficult to distinguish when playing in Π but the excess loss of playing outside of Π is large.

Recalling that $x_t \sim \mathrm{Uniform}([0,1])$, we will choose an adversary strategy P_1 that plays $q_1 \in C_1$ when $x \leq \beta$ and $q_2 \in C_2$ when $x > \beta$ for some constant $\beta \in [0,1]$ we will optimize later. Strategy P_2 is analogously defined but for distributions $q_3 \in C_1$ and $q_4 \in C_2$. That is,

$$P_1 = q_1 \mathbb{1}\{x \leq \beta\} + q_2 \mathbb{1}\{x > \beta\}$$
$$P_2 = q_3 \mathbb{1}\{x \leq \beta\} + q_4 \mathbb{1}\{x > \beta\}.$$

In particular, for some $\epsilon > 0$, $q_a \in C_1$ and $q_b \in C_2$ (which will be described shortly), the four distribution are

$$q_1 = q_a - \epsilon v, q_2 = q_b - \epsilon v, q_3 = q_a + \epsilon v, \text{ and } q_4 = q_b + \epsilon v. \tag{16}$$

This construction ensures that the feedback distribution of q_1 and q_2 is the same when playing action 1 or 2 (and is the same when following policies π_1 or π_2), but the two policies will get different losses when they disagree because $(\ell_1 - \ell_2)^\top v \neq 0$. We will choose q_a and q_b to maximize this difference.

To be precise, define $Q_1(q) := \max\{\epsilon' \geq 0 : q + \epsilon' v \in C_1, q - \epsilon' v \in C_1\}$ and $Q_2(q) := \max\{\epsilon' \geq 0 : q + \epsilon' v \in C_2, q - \epsilon' v \in C_2\}$ and choose $q_a \in \arg\max_q Q_1(q)$ and $q_b \in \arg\max_q Q_2(q)$. The maximum ϵ translation with $q_1, q_3 \in C_1$ and $q_2, q_4 \in C_2$ is $\epsilon_0 = \max\{Q_1(q_a), Q_2(q_b)\}$. Because C_1 and C_2 are convex, $q_1, q_3 \in C_1$ and $q_2, q_4 \in C_2$ for all $\epsilon \leq \epsilon_0$.

Next, we create a policy class Π that is able to use the context to do better than any fixed action. We will consider the threshold policies

$$\pi_1(x) = e_1 \mathbb{1}\{x \leq \beta + \beta_1\} + e_2 \mathbb{1}\{x > \beta + \beta_1\} \text{ and}$$
$$\pi_2(x) = e_1 \mathbb{1}\{x \leq \beta - \beta_2\} + e_2 \mathbb{1}\{x > \beta - \beta_2\}$$

for constants $0 < \beta_1 \leq 1 - \beta$ and $0 < \beta_2 \leq \beta$, which we will set later. As we can see, both policies use the context to track the optimal action given x_t, but do so with some error. Policy π_1 slightly favors action 1 and strategy P_1 plays distributions which gives lower loss to 1 over action 2. The situation is reversed for π_2 and P_2. Our construction ensures that π_1 holds a slight edge over π_2 under P_1. We will quantify the difference in expected losses in the next section.

D.2 Calculating the loss differences

In the sequel, we will use $\mathbb{E}_1$ to denote the expectation over context X and adversary action $J \sim P_1 | X$. We analogously use $\mathbb{E}_2$ for expectations over X and $J \sim P_2 | X$.

We now show that is possible to set β, β_1, and β_2 so that expected loss difference between π_1 and π_2 is $O(\epsilon)$, but the loss of either policy is a constant better than any fixed action. We denote the expected loss of playing action i and following policy π_i under strategy P_j by $\ell_i^j := \mathbb{E}_{J \sim P_j}[e_i^\top L e_J]$ and $\ell_{\pi_i}^j := \mathbb{E}_{J \sim P_j}[\pi_i(x)^\top L e_J]$, respectively.

Lemma 5. *There exists β, β_1, β_2 and some ϵ_0' such that, for all $\epsilon \leq \epsilon_0'$, the following inequalities are simultaneously satisfied:*

$$\ell_i^1 - \ell_{\pi_1}^1 \geq c_1,$$
$$\ell_i^2 - \ell_{\pi_2}^2 \geq c_1,$$
$$\ell_{\pi_2}^1 - \ell_{\pi_1}^1 \geq c_2 \epsilon,$$
$$\ell_{\pi_1}^2 - \ell_{\pi_2}^2 \geq c_2 \epsilon,$$

for some constants $c_1 > 0$ and $c_2 > 0$ that depend only on the structure of the game.

Proof. An easy calculation yields

$$\ell_{\pi_1}^1 = \ell_1^\top q_1(\beta - \beta_2) + \ell_1^\top q_1 \beta_2 + \ell_1^\top q_2 \beta_1 + \ell_2^\top q_2(1 - \beta - \beta_1), \text{ and}$$
$$\ell_{\pi_2}^1 = \ell_1^\top q_1(\beta - \beta_2) + \ell_2^\top q_1 \beta_2 + \ell_2^\top q_2 \beta_1 + \ell_2^\top q_2(1 - \beta - \beta_1).$$

Thus,

$$\begin{aligned}
\ell^1_{\pi_2} - \ell^1_{\pi_1} &= \ell_2^\top q_1\beta_2 + \ell_2^\top q_2\beta_1 - \ell_1^\top q_1\beta_2 - \ell_1^\top q_2\beta_1 \\
&= \beta_2(\ell_2 - \ell_1)^\top q_1 - \beta_1(\ell_1 - \ell_2)^\top q_2 \\
&= \beta_2(\ell_2 - \ell_1)^\top q_a - \beta_2\epsilon(\ell_2 - \ell_1)^\top v - \beta_1(\ell_1 - \ell_2)^\top q_b + \beta_1\epsilon(\ell_1 - \ell_2)^\top v \\
&= \beta_2(\ell_2 - \ell_1)^\top q_a - \beta_1(\ell_1 - \ell_2)^\top q_b + (\beta_1 + \beta_2)\epsilon,
\end{aligned}$$

where $(\ell_2 - \ell_1)^\top q_a$ and $(\ell_1 - \ell_2)^\top q_b$ are both positive constants. For readability, we will define the constants $\delta_{1\to i} := (\ell_i - \ell_1)^\top q_a > 0$ and $\delta_{2\to i} := (\ell_i - \ell_2)^\top q_b > 0$, where the first quantity is the excess loss of playing action i instead of action 1 under q_a (where action 1 is optimal) and the second is the analogous quantity for q_b. Hence, under P_1, the excess loss of following policy π_2 instead of the optimal π_1 is

$$\ell^1_{\pi_2} - \ell^1_{\pi_1} = \beta_2\delta_{1\to 2} - \beta_1\delta_{2\to 1} + (\beta_1 + \beta_2)\epsilon.$$

The calculation for $\ell^2_{\pi_2} - \ell^2_{\pi_1}$ is quite similar:

$$\begin{aligned}
\ell^2_{\pi_1} &= \ell_1^\top q_3(\beta - \beta_2) + \ell_1^\top q_3\beta_2 + \ell_1^\top q_4\beta_1 + \ell_2^\top q_4(1 - \beta - \beta_1) \text{ and} \\
\ell^2_{\pi_2} &= \ell_1^\top q_3(\beta - \beta_2) + \ell_2^\top q_3\beta_2 + \ell_2^\top q_4\beta_1 + \ell_2^\top q_4(1 - \beta - \beta_1),
\end{aligned}$$

and so we conclude that

$$\begin{aligned}
\ell^2_{\pi_1} - \ell^2_{\pi_2} &= \beta_1(\ell_1 - \ell_2)^\top q_4 - \beta_2(\ell_2 - \ell_1)^\top q_3 \\
&= \beta_1(\ell_1 - \ell_2)^\top q_b + \epsilon\beta_1(\ell_1 - \ell_2)^\top v - \beta_2(\ell_2 - \ell_1)^\top q_a - \epsilon\beta_2(\ell_2 - \ell_1)^\top v \\
&= \beta_1\delta_{2\to 1} - \beta_2\delta_{1\to 2} + (\beta_1 + \beta_2)\epsilon.
\end{aligned}$$

We also need to evaluate $\ell^j_i - \ell^j_{\pi_1}$ for $j = 1, 2$. It is easy to calculate $\ell^1_i = \beta\ell_i^\top q_1 + (1-\beta)\ell_i^\top q_2$, and therefore,

$$\begin{aligned}
\ell^1_i - \ell^1_{\pi_1} &= \beta\ell_i^\top q_1 + (1-\beta)\ell_i^\top q_2 - \beta\ell_1^\top q_1 - (1 - \beta - \beta_1)\ell_2^\top q_2 - \beta_1\ell_1^\top q_2 \\
&= \beta(\ell_i - \ell_1)^\top q_1 + (1-\beta)(\ell_i - \ell_2)^\top q_2 + \beta_1(\ell_2 - \ell_1)^\top q_2 \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_1\delta_{2\to 1} - \epsilon\left(\beta(\ell_i - \ell_1)^\top v + (1-\beta)(\ell_i - \ell_2)^\top v - \beta_1\right) \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_1\delta_{2\to 1} + \epsilon\left(\beta + \beta_1 - (\ell_i - \ell_2)^\top v\right) \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_1\delta_{2\to 1} + \epsilon\left(\beta + \beta_1 - \frac{1 + (2\ell_i - \ell_2 - \ell_1)^\top v}{2}\right),
\end{aligned}$$

where the last line used $(\ell_i - \ell_2)^\top v = (\ell_i - \ell_2)^\top v - \frac{1}{2}(\ell_1 - \ell_2)^\top v + \frac{1}{2} = (\ell_i - (\ell_1 - \ell_2)/2)^\top v + \frac{1}{2}$. Similarly, the $j = 2$ case is

$$\begin{aligned}
\ell^2_i - \ell^2_{\pi_2} &= \beta\ell_i^\top q_3 + (1-\beta)\ell_i^\top q_4 - \beta_2\ell_2^\top q_3 - (1-\beta)\ell_2^\top q_4 - (1 - \beta - \beta_2)\ell_1^\top q_3 \\
&= \beta(\ell_i - \ell_1)^\top q_3 + (1-\beta)(\ell_i - \ell_2)^\top q_4 + \beta_2(\ell_1 - \ell_2)^\top q_3 \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_2\delta_{1\to 2} + \epsilon\left(\beta(\ell_i - \ell_1)^\top v + (1-\beta)(\ell_i - \ell_2)^\top v + \beta_2\right) \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_2\delta_{1\to 2} + \epsilon\left((\ell_i - \ell_2)^\top v - (\beta - \beta_2)\right) \\
&= \beta\delta_{1\to i} + (1-\beta)\delta_{2\to i} - \beta_2\delta_{1\to 2} + \epsilon\left(\frac{1 + (2\ell_i - \ell_2 - \ell_1)^\top v}{2} - (\beta - \beta_2)\right).
\end{aligned}$$

The last two cases can be deduced by combining the equalities above:

$$\begin{aligned}
\ell^1_1 - \ell^1_{\pi_2} &= \ell^1_1 - \ell^1_{\pi_1} + (\ell^1_{\pi_1} - \ell^1_{\pi_2}) \\
&= (1-\beta)\delta_{2\to 1} - \beta_1\delta_{1\to 2} - \epsilon(1 - (\beta - \beta_2)), \text{ and} \\
\ell^2_2 - \ell^2_{\pi_1} &= \ell^2_2 - \ell^2_{\pi_2} + (\ell^2_{\pi_2} - \ell^2_{\pi_1}) \\
&= \beta\delta_{1\to 2} - \beta_2\delta_{2\to 1} - \epsilon(\beta + \beta_1).
\end{aligned}$$

With these equalities in hand, we now optimize for β, β_1, and β_2. First, it suffices to take $\beta = \frac{1}{2}$, which lets us bound

$$\begin{aligned}
\ell_i^1 - \ell_{\pi_1}^1 &= \frac{\delta_{1\to i} + \delta_{2\to i}}{2} - \beta_1 \delta_{2\to 1} + \epsilon\left(\beta_1 - \frac{(2\ell_i - \ell_2 - \ell_1)^\top v}{2}\right) \\
&\geq \frac{\delta_{1\to i} + \delta_{2\to i}}{2} - \beta_1 \delta_{2\to 1} - \frac{\epsilon}{2}\left|(2\ell_i - \ell_2 - \ell_1)^\top v\right| \text{ and} \\
\ell_i^2 - \ell_{\pi_2}^2 &= \frac{\delta_{1\to i} + \delta_{2\to i}}{2} - \beta_2 \delta_{1\to 2} + \epsilon\left(\beta_2 + \frac{(2\ell_i - \ell_2 - \ell_1)^\top v}{2}\right) \\
&\geq \frac{\delta_{1\to i} + \delta_{2\to i}}{2} - \beta_2 \delta_{1\to 2} - \frac{\epsilon}{2}\left|(2\ell_i - \ell_2 - \ell_1)^\top v\right|.
\end{aligned}$$

Now, define the two constants

$$c_3 = \max_i \left|(2\ell_i - \ell_2 - \ell_1)^\top v\right| \text{ and } c_4 = \min_i \frac{\min\{\delta_{1\to i}, \delta_{2\to i}\}}{2},$$

where the latter is strictly positive; the non-degeneracy of action 1 implies that $\delta_{1\to i} > 0$ for all $i \neq 1$, and the non-degeneracy of action 2 implies that $\delta_{2\to i} > 0$ for all $i \neq 2$. Combining $\epsilon \leq \epsilon_0$ with these constants give the bounds

$$\begin{aligned}
\ell_i^1 - \ell_{\pi_1}^1 &\geq c_4 - \beta_1 \delta_{2\to 1} - \epsilon_0 c_3 \text{ and} \\
\ell_i^2 - \ell_{\pi_2}^2 &\geq c_4 - \beta_2 \delta_{1\to 2} - \epsilon_0 c_3.
\end{aligned}$$

So long as we take $\epsilon_0 \leq c_4/(2c_3)$, choosing $\beta_1 \leq \frac{c_4/2 - \epsilon_0 c_3}{\delta_{2\to 1}}$ and $\beta_2 \leq \frac{c_4/2 - \epsilon_0 c_3}{\delta_{1\to 2}}$ yields

$$\ell_i^1 - \ell_{\pi_1}^1 \geq \frac{c_4}{2} \text{ and } \ell_i^2 - \ell_{\pi_2}^2 \geq \frac{c_4}{2}.$$

We can lower bound the cross terms by

$$\begin{aligned}
\ell_1^1 - \ell_{\pi_2}^1 &\geq \frac{\delta_{2\to 1}}{2} - \beta_1 \delta_{1\to 2} - \frac{\epsilon}{2} \geq c_4 - \beta_1 \delta_{1\to 2} - \frac{\epsilon_0}{2} \text{ and} \\
\ell_2^2 - \ell_{\pi_1}^2 &\geq \frac{\delta_{1\to 2}}{2} - \beta_2 \delta_{2\to 1} - \frac{\epsilon}{2} \geq c_4 - \beta_2 \delta_{2\to 1} - \frac{\epsilon_0}{2},
\end{aligned}$$

and so it suffices to take $\beta_1 \leq \frac{c_4 - \epsilon_0}{2\delta_{1\to 2}}$ and $\beta_2 \leq \frac{c_4 - \epsilon_0}{2\delta_{2\to 1}}$ to guarantee that $\ell_1^1 - \ell_{\pi_2}^1 \geq c_4/4$ and $\ell_2^2 - \ell_{\pi_1}^2 \geq c_4/4$.

To summarize, setting $\beta = \frac{1}{2}$,

$$\begin{aligned}
\beta_1 &\leq \min\{\frac{c_4/2 - \epsilon_0 c_3}{\delta_{2\to 1}}, \frac{c_4 - \epsilon_0}{2\delta_{1\to 2}}\}, \quad \text{and} \\
\beta_2 &\leq \min\{\frac{c_4 - \epsilon_0}{2\delta_{2\to 1}}, \frac{c_4/2 - \epsilon_0 c_3}{\delta_{1\to 2}}\},
\end{aligned}$$

yields the first two inequalities in the lemma for $c_1 = c_4/2$.

We now turn to the second two inequalities of the lemma. Choosing $\beta_2 = \beta_1 \frac{\delta_{2\to 1}}{\delta_{1\to 2}}$ implies that

$$\ell_{\pi_2}^1 - \ell_{\pi_1}^1 = \ell_{\pi_1}^2 - \ell_{\pi_2}^2 = \epsilon \beta_1 \left(1 + \frac{\delta_{2\to 1}}{\delta_{1\to 2}}\right).$$

Therefore, we set $\beta_1 = \frac{c_4/2 - \epsilon_0 c_3}{2\delta_{2\to 1}}$ and $\beta_2 = \frac{c_4/2 - \epsilon_0 c_3}{2\delta_{1\to 2}}$, which satisfies the first two inequalities, has $\beta_2 = \beta_1 \frac{\delta_{2\to 1}}{\delta_{1\to 2}}$, and therefore implies that

$$\ell_{\pi_2}^1 - \ell_{\pi_1}^1 = \ell_{\pi_1}^2 - \ell_{\pi_2}^2 \geq \frac{\epsilon}{2}(c_4/2 - \epsilon_0 c_3)\left(\frac{1}{\delta_{2\to 1}} + \frac{1}{\delta_{1\to 2}}\right);$$

that is, we may take $c_2 = \frac{1}{2}(c_4/2 - \epsilon_0 c_3)\left(\frac{1}{\delta_{2\to 1}} + \frac{1}{\delta_{1\to 2}}\right)$ and long as $\epsilon_0 \leq c_4/(2c_3)$, which we may assume (since ϵ_0 is a parameter we control as well). $\square$

D.3 Bounding the KL-Divergence

At a high level, any randomized algorithm must make similar decisions when given similar data. In our setting, the algorithm observes the contexts, which do not depend on the stategy of the adversary, and the feedback symbols. An important step in the argument is to lower bound the difference in feedback distributions as a function of the expected number of plays of different actions. As is standard, we will use the KL-divergence between distribution p and q, denoted by $\mathrm{KL}(p\|q)$, as the measure of distance.

Denote the symbol received at round t by $f_t \in \Sigma$. Let $P_j^*(\cdot|f_{1:t-1}, x_{1:t})$ be the mass function over Σ at round t generated by the algorithm's choices if the adversary uses strategy P_j, and let N_i be the number of times the algorithm plays action i.

Lemma 6. *The relative entropy between the feedback symbols of strategy P_1 and P_2 has the upper bound*

$$\mathrm{KL}(P_1^*||P_2^*) \leq \sum_{i>2} \mathbb{E}_1[N_i] c_5 \epsilon^2, \tag{17}$$

where c_5 is some game-dependent constant.

Proof. Fix some algorithm $\mathcal{A}$ which maps the information available for selecting I_t, denoted $H^t := f_{1:t-1}, x_{1:t}$. We can apply the KL-divergence chain rule T times to conclude

$$\begin{aligned}
\mathrm{KL}(P_2^*||P_1^*) &= \sum_{t=1}^{T-1} \sum_{f_{1:t-1}} P_2^*(f_{1:t-1}) \sum_{f_t} P_2^*(f_t|H^t) \log \frac{P_2^*(f_t|H^t)}{P_1^*(f_t|H^t)} \\
&= \sum_{t=1}^{T-1} \sum_{f_{1:t-1}} P_2^*(f_{1:t-1}) \sum_{i=1}^{N} \mathbb{1}\left\{\mathcal{A}(H^t) = i\right\} \sum_{f_t} P_2^*(f_t|H^t) \log \frac{P_2^*(f_t|H^t)}{P_1^*(f_t|H^t)},
\end{aligned}$$

By a slight abuse of notation, define $S_i P_j$ to be the distribution of feedback symbols from playing action i under strategy P_j with x_t marginalized out, i.e. $S_i P_1 = \beta S_i q_1 + (1-\beta) S_i q_2$ and $S_i P_2 = \beta S_i q_3 + (1-\beta) S_i q_4$. However, because $v \in \ker(S_{1,2})$, we have that $S_i P_1 = S_i P_2$ for $i = 1, 2$. Hence, the $\mathcal{A}(H^t) \in \{1, 2\}$ terms in the summation are zero. We can then evaluate

$$\begin{aligned}
\mathrm{KL}(P_1^*||P_2^*) &= \sum_{t=1}^{T-1} \sum_{f_{1:t-1}} P_1^*(f_{1:t-1}) \sum_{i>2}^{N} \mathbb{1}\left\{\mathcal{A}(H^t) = i\right\} \sum_{f_t} P_1^*(f_t|H^t) \log \frac{P_1^*(f_t|H^t)}{P_2^*(f_t|H^t)} \\
&\leq \sum_{i>2} \mathbb{E}_1[N_i] \mathrm{KL}(S_i P_1 \| S_i P_2) \\
&\leq \sum_{i>2} \mathbb{E}_1[N_i] \mathrm{KL}(P_1 \| P_2).
\end{aligned}$$

Thus, we may turn a bound on $\mathrm{KL}(P_1\|P_2)$ into a bound on $\mathrm{KL}(P_1^*\|P_2^*)$, so let us consider the first quantity. We will briefly be explicit about the X dependence. Let $P_j(J, X)$, for $j = 1, 2$, denote the joint distribution of the context $X \sim \mathrm{Uniform}([0, 1])$ and adversary choice $J \sim P_j|X$. The chain rule yields

$$\begin{aligned}
\mathrm{KL}(P_1(J, X)\|P_2(J, X)) &= \mathrm{KL}(P_1(X)\|P_2(X)) + \mathrm{KL}(P_1(J|X)\|P_2(J|X)) \\
&= 0 + \beta \mathrm{KL}(q_1\|q_3) + (1-\beta)\mathrm{KL}(q_2\|q_4).
\end{aligned}$$

To bound each term, recall that $q_1 = q_a - \epsilon v$ and $q_3 = q_a + \epsilon v$ and apply Lemma 12 from [5], which implies that, for all ϵ small enough and some positive constants c_1' and c_2',

$$\mathrm{KL}(q_a - \epsilon v||q_a + \epsilon v) \leq c_1' \epsilon^2 \|v\|_\infty^2 \text{ and } \mathrm{KL}(q_b - \epsilon v||q_b + \epsilon v) \leq c_2' \epsilon^2 \|v\|_\infty^2.$$

Thus, our total relative entropy bound is

$$\mathrm{KL}(P_1^*||P_2^*) \leq \sum_{i>2} \mathbb{E}_1[N_i] c_5 \epsilon^2 \tag{18}$$

with $c_5 = \max\{c_1', c_2'\} \|v\|_\infty^2$. $\square$

The next tool we need is a way to translate relative entropy bounds into high probability bounds: a high-probability version of Pinsker's inequality.

Lemma 7. *[18, Lemma 2.6] For probability distribution P and Q with $P \ll Q$,*

$$\int \min\{dP, dQ\} \geq \frac{1}{2} \exp(-\mathrm{KL}(P\|Q)).$$

In particular, for some event A with $P(A) \leq Q(A)$, integrating the indicator function of A and A^c yields $P(A) \geq \frac{1}{2}\exp(-\mathrm{KL}(P\|Q))$ and $Q(A^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(P\|Q))$. In the case of $P(A) > Q(A)$, the same argument applied to A^c yields the same conclusion, which implies that

$$P(A) + Q(A^c) \geq \exp(-\mathrm{KL}(P\|Q)). \tag{19}$$

D.4 Finalizing the Bound

We are now ready to prove a lower bound on regret, and we begin by exploiting the construction of the game which allows the expected regret to be easily calculated from the action counts. Let $N_{\pi_i}(T)$ be the number of times $\mathcal{A}$ chooses policy π_i over a game of length T, which is a random variable dependent on the context and the random choices of the algorithm and adversary. We will also define $N_0(T)$ to be the number of times an $\mathcal{A}$ chooses action $i > 2$.

We assume that the strategy uses some β, β_1, and β_2 that satisfy Lemma 5. Thus, under P_1, the optimal policy is π_1, playing policy π_2 incurs at least $c_2\epsilon$ more expected loss, and playing any other action incurs at least c_1 more expected loss. The following construction is modeled after the slick proof of Lattimore and Szepesvari [11]. We can lower bound the regret under P_1 by

$$\begin{aligned}
\mathbb{E}_1[\mathcal{R}_T] &\geq \mathbb{E}_1\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{1}\left\{\mathcal{A}(H^t) = i\right\}(\ell_i - \ell_{\pi_1(x_t)})^\top J_t\right] \\
&\geq \mathbb{E}_1\left[\sum_{t=1}^{T}\sum_{i=3}^{N} \mathbb{1}\left\{\mathcal{A}(H^t) = i\right\}(\ell_i - \ell_{\pi_1(x_t)})^\top J_t\right] \\
&\quad + \mathbb{E}_1\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(H^t) = \pi_2(x_t)\right\}(\ell_{\pi_2(x_t)} - \ell_{\pi_1(x_t)})^\top J_t\right] \\
&\geq c_1\mathbb{E}_1\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(H^t) > 2\right\}\right] + \epsilon c_2 \mathbb{E}_1\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(H^t) = \pi_2(x_t)\right\}\right] \\
&= c_1\mathbb{E}_1\left[N_0(T)\right] + \epsilon c_2 \mathbb{E}_1\left[N_2(T)\right] \\
&\geq c_1\mathbb{E}_1\left[N_0(T)\right] + \frac{\epsilon c_2 T}{2} P_1\left(N_2(T) \geq \frac{T}{2}\right).
\end{aligned}$$

Similarly, we can bound the regret under P_2 by

$$\begin{aligned}
\mathbb{E}_2[\mathcal{R}_T] &\geq \mathbb{E}_2\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{1}\left\{\mathcal{A}(H^t) = i\right\}(\ell_i - \ell_{\pi_2(x_t)})^\top J_t\right] \\
&\geq c_1\mathbb{E}_2\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(H^t) > 2\right\}\right] + \epsilon c_2 \mathbb{E}_2\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(H^t) = \pi_1(x_t)\right\}\right] \\
&\geq \min\{c_1, \epsilon c_2\}\mathbb{E}_2\left[N_0(T) + N_1(T)\right] \\
&\geq \min\{c_1, \epsilon c_2\}\frac{T}{2} P_2\left(N_0(T) + N_1(T) > \frac{T}{2}\right) \\
&\geq \min\{c_1, \epsilon c_2\}\frac{T}{2} P_2\left(N_2(T) \leq \frac{T}{2}\right),
\end{aligned}$$

with the last line following from $N_2(T) + N_1(T) + N_0(T) \leq T$. Note that we simply dropped the terms where the algorithms plays action 1 or 2 but in disagreement with both policies.

The final adversary strategy will be a uniform mixture between P_1 and P_2, and, under the assumption that $\epsilon \leq c_1/c_2$,

$$\mathbb{E}[\mathcal{R}_T] = \frac{\mathbb{E}_1[\mathcal{R}_T] + \mathbb{E}_2[\mathcal{R}_T]}{2} \geq \mathbb{E}_1\left[N_0(T)\right] c_1 + \frac{\epsilon c_2 T}{2}\left(P_1\left(N_2(T) \geq \frac{T}{2}\right) + P_2\left(N_2(T) \leq \frac{T}{2}\right)\right). \tag{20}$$

We can finally assemble all the ingredients into a lower bound proof.

Proof of Theorem 3. First, by definition, $N_0(T) = \sum_{i>2}^{N} N_i(T)$. Now, combining the regret lower bound from (20) with (19) and Lemma 6 yields

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &\geq \mathbb{E}_1\left[N_0(T)\right] c_1 + \frac{\epsilon c_2 T}{2}\exp(-\mathrm{KL}(P_1 \| P_2)) \\
&\geq \mathbb{E}_1\left[N_0(T)\right] c_1 + \frac{\epsilon c_2 T}{2}\exp(-\mathbb{E}_1[N_0(T)] c_5 \epsilon^2).
\end{aligned}$$

If $\mathbb{E}_1\left[N_0(T)\right] \geq T^{2/3}$, then the desired bound follows immediately. Otherwise, setting $\epsilon = T^{-\frac{1}{3}}$ yields

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &\geq \mathbb{E}_1\left[N_0(T)\right] c_1 + \frac{\epsilon c_2 T}{2}\exp(-\mathrm{KL}(P_1 \| P_2)) \\
&\geq \mathbb{E}_1\left[N_0(T)\right] c_1 + T^{\frac{2}{3}}\frac{c_2}{2}\exp(-\mathbb{E}_1[N_0(T)] c_5 T^{-\frac{2}{3}})
\end{aligned}$$

with the term in the exponential approaching 0. Since ϵ is a decreasing sequence, all the assumptions that ϵ is smaller than certain quantities will eventually be satisfied. $\square$