Journal of Machine Learning Research 25 (2024) 1-36 Submitted 2/24; Published 10/24

Information Capacity Regret Bounds for Bandits with Mediator Feedback

Khaled Eldowa KHALED.ELDOWA@UNIMI.IT
Università degli Studi di Milano
Milano, 20133, Italy

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Università degli Studi di Milano and Politecnico di Milano
Milano, 20133, Italy

Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
Politecnico di Milano
Milano, 20133, Italy

Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Politecnico di Milano
Milano, 20133, Italy

Editor: Tor Lattimore

Abstract

This work addresses the mediator feedback problem, a bandit game where the decision set consists of a number of policies, each associated with a probability distribution over a common space of outcomes. Upon choosing a policy, the learner observes an outcome sampled from its distribution and incurs the loss assigned to this outcome in the present round. We introduce the policy set capacity as an information-theoretic measure for the complexity of the policy set. Adopting the classical EXP4 algorithm, we provide new regret bounds depending on the policy set capacity in both the adversarial and the stochastic settings. For a selection of policy set families, we prove nearly-matching lower bounds, scaling similarly with the capacity. We also consider the case when the policies' distributions can vary between rounds, thus addressing the related bandits with expert advice problem, which we improve upon its prior results. Additionally, we prove a lower bound showing that exploiting the similarity between the policies is not possible in general under linear bandit feedback. Finally, for a full-information variant, we provide a regret bound scaling with the information radius of the policy set.

Keywords: regret minimization, multi-armed bandits, expert advice, information theory, best of both worlds

1. Introduction

The framework of multi-armed bandits (MAB) models sequential decision-making problems with partial feedback. Real-world applications of this framework span a wide array of domains and include problems such as dynamic pricing (Misra et al., 2019) and advert placement (Schwartz et al., 2017). In the classical non-stochastic MAB problem (Auer et al., 1995), a learner, faced with a fixed set of actions (also referred to as "arms"), repeatedly interacts with the environment in a series of rounds by selecting an action and subsequently observing a numerical loss assigned beforehand to this action. The performance of the learner is measured via the notion of regret, which compares the cumulative loss of the learner with that of the best action in hindsight. The minimax regret for this problem, that is, the smallest achievable regret in the worst-case, is known to be of order $\sqrt{KT}$ (Audibert and Bubeck, 2009), with K being the number of actions and T the number of rounds.

This formulation, however, fails to model situations in which, aside from observing the loss of the played action, the learner could—in the same round—obtain information concerning the losses of other actions. This information leakage could be a result of prior knowledge of an underlying structure for the losses, or owing to more explicit side observations. Regardless of form, such extra information can lead to more efficient learning in the face of large (or even infinite) action sets. A prominent example of structured losses is exhibited by the (adversarial) linear bandits problem, in which the action set is a subset of $\mathbb{R}^d$ and the loss assigned to an action in a given round is the inner product between the action and a common latent loss vector associated with that round. For this setting, Bubeck et al. (2012) provide an algorithm achieving nearly optimal regret bounds of order $\sqrt{dT \log K}$ for finite action sets and $d\sqrt{T \log T}$ for compact actions sets. On the other hand, a simple form of side observations is modelled by the framework of online learning with graph feedback, in which, upon choosing an action, the learner additionally observes the losses of the actions adjacent to the chosen one in a given graph. Algorithms exploiting this extra feedback can enjoy improved regret guarantees depending on the structure of the graph (see, e.g., Alon et al., 2015).

In this work, we study a certain bandit model where the information leakage results from a combination of side observations and a structured assignment of losses. In the basic template of this model, the learner is faced with a policy set whose elements are each associated with a probability distribution over a common (finite) space of outcomes. At each round, a (latent) loss map associates each outcome with a numerical loss. Upon choosing a policy, an outcome is sampled from its distribution, and the learner subsequently observes *both* the outcome and its associated loss. Depending on the problem, the loss map could be fixed over the rounds or changing in a stochastic or adversarial manner. The regret in this framework is defined as the difference between the (expected) cumulative loss of the learner and that of the optimal policy in hindsight. Our main goal is to understand how the structure of the policy set, in particular, how the similarity between the policies affects the achievable regret. One aspect of this problem reminiscent of linear bandits is that having chosen a policy, the learner's expected loss is linear in the policy's distribution, seen as a vector in the simplex. The distinction, however, is that the learner does not observe this quantity, as would be the case under linear bandit feedback; instead, one outcome, sampled via the chosen policy, and its assigned loss are observed.

This framework has been studied in the works of Papini et al. (2019) and Metelli et al. (2021), where it was referred to as the *mediator feedback* model. The name here highlights the role of the outcomes as an extra layer of feedback "mediating" between the chosen policy and the observed loss, thereby allowing additional information gain regarding other policies. This feedback model arises in these two works from a certain formulation of the online policy optimization problem in episodic reinforcement learning. An instance of their setting is characterized by a Markov decision process (MDP) and a set of policies that map states to distributions over actions. At every round of interaction, the learner selects a policy through which they interact with the MDP for a fixed horizon. Naturally, the learner observes both the sampled trajectory and the accumulated reward, and aims to compete with the best policy from the given policy set. In this case, the trajectories are the outcomes over which each policy induces a probability distribution.

The feedback structure of the mediator framework is shared with the more classical problem of *bandits with expert advice* (Auer et al., 1995). This problem is a variation of the (non-stochastic) MAB problem described above, where at the beginning of every round, the learner receives "advice" from each of a number of "experts" in the form of a probability distribution over the actions. The goal then becomes competing with the (expected) cumulative loss of the best expert in hindsight. Here, the actions of the MAB instance play the role of the outcomes, which are sampled from the distributions provided by the experts. Exactly fitting this problem into the mediator feedback framework additionally requires restricting the learner to only access the actions by way of sampling from (a mixture of) the experts' distributions, though this requirement is already satisfied by most state-of-the-art approaches. A more important distinction of the expert advice problem is the incorporation of a contextual element in that the distributions recommended by the experts can vary from round to round. Given our stated goal of studying the extent to which the similarity between the available distributions can be exploited, the addition of this contextual element is somewhat orthogonal to the main focus of this work, though it will still be briefly treated.

1.1 Prior Results

The Exp4 algorithm was proposed in the work of Auer et al. (1995) to address the bandits with expert advice problem, and remains an important benchmark in the contextual bandits literature. It was shown to enjoy a regret bound of order $\sqrt{KT \log N}$, where N is the number of experts and K still denotes the number of actions. As the recommendations of an expert can be seen as a strategy against which the learner is competing, this result shows that one can achieve a regret scaling only logarithmically with the number of such strategies. This was later shown to be nearly optimal by Seldin and Lugosi (2016), who proved a lower bound of order $\sqrt{KT \log N / \log K}$. However, this lower bound concerns an instance where the number of experts is exponential in the number of actions, and the experts' distributions are deterministic. The former is not surprising; if the number of experts is small compared to the arms, one can always achieve a regret of order $\sqrt{NT}$ by playing a minimax optimal bandit algorithm directly over the experts, entirely casting aside the structure of the problem. More importantly, the bound of Auer et al. (1995) does not reflect one's expectation that the problem should become easier if the recommended distributions are more similar, and the lower bound does not address this question.

Via an elaborate modification of Exp4, McMahan and Streeter (2009) did address these very two points. For a fixed set of expert recommendations, their algorithm achieves a bound of order $\sqrt{\mathcal{S}T \log N}$, where $\mathcal{S}$, formally defined in Section 3, is a notion of effective size of the set of recommended distributions (i.e., the policy set). It satisfies $\mathcal{S} \leq \min\{K, N\}$, but can be smaller depending on the similarity between the distributions—or the "agreement between the experts". Specifically, it reaches its smallest value, $\mathcal{S} = 1$, when all the distributions are identical. One issue with this bound is that when the number of experts is small (say, only two), the fact that $\mathcal{S} \geq 1$ means that no substantial improvement is achieved over the $\sqrt{NT}$ bound, no matter how similar the distributions are. Indeed, $\mathcal{S} - 1$ is arguably a more apt metric as it can shrink arbitrarily if the distributions are similar enough. In particular, it reduces to the total variation distance when there are only two distributions. Nevertheless, even if the bound were to scale with this quantity, one may ask whether this is the best achievable dependence on the structure of the policy set. We note in passing that the bound of McMahan and Streeter (2009) can also be achieved by plain Exp4 in the general case (see Lattimore and Szepesvári, 2020, Theorem 18.3), where it takes the form $\sqrt{\sum_t \mathcal{S}_t \log N}$ with $\mathcal{S}_t$ measuring the (dis)agreement between the experts at the t-th round.

While these results concern the adversarial regime, where no statistical constraints are placed on the losses, similar mediator feedback problems have been studied in the stochastic regime, where the losses are drawn at every round from a fixed distribution. This includes the aforementioned works of Papini et al. (2019) and Metelli et al. (2021), and that of Sen et al. (2018), who consider a stochastic variant of the expert advice problem. Unlike the worst-case flavour (in terms of the dependence of the loss map) of the bounds in the adversarial regime, the results in these works are generally instance-based; the bounds enjoy a logarithmic dependence on the time horizon, but degrade in harder instances where suboptimal policies are difficult to discern. Still, the dependence of these bounds on the policy set structure is largely independent of the loss map as it is primarily represented via diameter-like quantities measuring the pairwise maximum "distance" between the distributions according to some dissimilarity measure, mainly the chi-squared divergence or related quantities. Comparing $\mathcal{S} - 1$ with this chi-squared "diameter", neither quantity uniformly dominates the other, though the former is always bounded, while the latter need not be.

In the online learning and bandits literature, best-of-both-worlds (BOBW) algorithms (Bubeck and Slivkins, 2012) address the adversarial and stochastic regimes simultaneously without prior knowledge of the nature of the environment. They guarantee regret (poly) logarithmic in the time horizon when the faced environment is stochastic, while retaining sub-linear regret against general environments. Of particular relevance to our setting is the recent work of Dann et al. (2023), where they obtain the first BOBW guarantees for bandits with expert advice (or contextual bandits) using Exp4 as a black-box decision rule within a cascade of two meta-algorithms. For stochastic environments, the bound is of order $K \log T \log N / \Delta$, where Δ denotes the minimum sub-optimality gap for the experts, whereas the traditional bound of $\sqrt{KT \log N}$ is guaranteed for all environments. As apparent, these bounds feature the usual coarse dependence on the number of actions and are therefore unable to reflect the affinity of the recommended distributions.

1.2 Contributions[1]

In this work, we consider a generic mediator feedback setting with finite policy and outcome sets. We propose a new complexity (or effective size) measure for the policy set, which we refer to as the chi-squared capacity of the policy set, or simply the policy set capacity. This quantity (defined in Section 3) can be interpreted as the information capacity of a certain hypothetical communication channel induced by the policy set, albeit we define the mutual information between the input and output of said channel in terms of the chi-squared

[1] A preliminary version of this work appeared as (Eldowa et al., 2023).

divergence in lieu of the standard Kullback–Leibler divergence. The channel referred to here is one whose input alphabet is the policy set and output alphabet is the outcome set, while its transition matrix is defined such that conditioned on an input policy, the output is drawn from the policy's distribution. Besides its more natural information-theoretic interpretation, this notion of capacity is never larger than either $\mathcal{S}-1$ or the maximum pairwise chi-squared divergence. Moreover, when the policy set consists of only two policies, the capacity reduces to a (symmetric) divergence measure of the same order as the squared Hellinger distance and the triangular discrimination.

In Section 3, we consider the adversarial regime and provide an improved regret bound of order $\max\{\sqrt{\mathcal{C}T\log N}, \log N\}$ for Exp4, where $\mathcal{C}$ denotes the capacity. Unlike prior results, the horizon-dependent term in this bound can shrink arbitrarily if the distributions are similar enough. We then extend this result to the case when the policies' distributions can vary between rounds, thus addressing the general bandits with expert advice problem. In particular, we show that the same algorithm run with an adaptive learning rate enjoys a bound essentially of order $\sqrt{\sum_t \mathcal{C}_t \log N}$, where $\mathcal{C}_t$ is the capacity of the policies' distributions at round t. This bound is obtained as an implication of a stronger history-dependent bound, where the complexity of the policy set in a given round is represented through the mutual (chi-squared-)information between the chosen policy and the drawn outcome, conditioned on the events up to the previous round.

Still considering the Exp4 algorithm, we provide best-of-both-worlds bounds in Section 4. In the stochastic regime, we show that the algorithm enjoys a bound of order $\mathcal{C}\log T\log(NT)/\Delta$, where Δ is the minimum sub-optimality gap for the policies. Simultaneously, the algorithm is shown to retain a worst-case bound of order $\sqrt{\mathcal{C}T\log T\log N}$. The former bound follows from a more general guarantee that we provide for the adversarially corrupted stochastic regime, an intermediate regime commonly considered in BOBW works (see, e.g., Ito et al., 2022; Dann et al., 2023). The proof of this result builds upon the techniques developed in (Ito et al., 2022) for proving BOBW bounds for a related algorithm in the setting of online learning with graph feedback.

We complement these results in Section 5 by proving worst-case lower bounds for three families of policy sets. These lower bounds scale with the policy set capacity in the same manner as the regret bound we provided for Exp4 in the adversarial setting, asserting its optimality up to factors logarithmic in the number of policies. Additionally, in Section 6, we prove another lower bound through which we aim to compare the studied feedback model with that of linear bandits. As alluded to before, the policies' distributions can be treated as a set of arms belonging to the simplex under an alternative linear bandit formulation of the problem, where the learner observes the inner product between the chosen policy and the latent loss map. Considering a particular family of policy sets, we prove that under linear bandit feedback, one must incur regret of order at least $\sqrt{NT}$ against any policy set in this family, even as the capacity approaches zero. This shows that the attainability of regret guarantees that improve as policies become more similar is a distinctive feature of mediator feedback, a key aspect of which is that the observed loss is attributed to a single outcome sampled via the chosen policy and revealed to the learner.

Finally, in Section 7, we consider a full-information variant of the problem, where the learner observes the entire loss map at every round. This can be seen as a generalization of the prediction with expert advice problem (Cesa-Bianchi et al., 1997). For this setting, we show that a simple Online Mirror Descent strategy enjoys a regret bound of order $\sqrt{\mathcal{C}_{\mathrm{KL}} T}$, where $\mathcal{C}_{\mathrm{KL}}$ is an altered form of the policy set capacity, based on the more standard KL-divergence. This quantity is no larger than $\log N$, and can be interpreted as an information radius of the policy set.

1.3 Additional Related Works

Krishnamurthy et al. (2020) study a contextual bandit problem with continuous actions, where the learner competes with a set of competitor policies mapping states (contexts) to actions. Instead of placing a smoothness assumption on the loss function, they opt for minimizing a notion of smoothed regret. More precisely, they fix a smoothing kernel that maps each action to a distribution over action. Accordingly, they obtain a new competitor class of smoothed policies that map states to distributions over actions by composing the original policies with the smoothing kernel. This has the effect of forcing a favourable structure on the policy set. Indeed, they obtain (via Exp4) a bound of order $\sqrt{\kappa T \log N}$ in the adversarial regime, where N is the number of policies and κ, called the kernel complexity, is defined as the largest possible density assigned by the smoothing kernel with respect to some base probability measure. This latter quantity upper bounds the continuous analogue of $\mathcal{S}$ for any given context. Similar smoothed benchmarks are studied in (Majzoubi et al., 2020) and (Zhu and Mineiro, 2022) under realizability assumptions.

In its dependence on the mutual information between policies and outcomes, one of the regret bound we provide for Exp4 (see Theorem 2) bears some superficial resemblance to the PAC-Bayesian results in (Seldin et al., 2011). For a stochastic contextual bandits problem, Seldin et al. (2011) prove bounds on the per-round instantaneous regret that depend on the mutual information between the observed state and the chosen action. This quantifies the complexity of the adopted decision rule, which is traded off against its empirical regret measured according to past observations. Another notable mention is the information-theoretic analysis of Russo and Van Roy (2016) for Thompson sampling in Bayesian bandit problems, which results in bounds scaling with the mutual information between the faced environment and the optimal action, or some satisficing benchmark (Russo and Van Roy, 2022; Arumugam and Van Roy, 2021). This measures the amount of information that needs to be acquired about the environment to identify the target action.

Another related line of work concerns the best arm identification (BAI) problem, a variant of the MAB problem where the learner's aim is to find the optimal arm efficiently. Reddy et al. (2023) and Poiani et al. (2023) study the BAI problem with the added constraint that the learner can only sample arms via a number of given stochastic policies. As the objective remains identifying the optimal arm, the manner in which the structure of the policy set affects the achievable regret is fundamentally different from our setting. Indeed, in their problem, a more diverse policy set can be advantageous to the learner.

2. Preliminaries

In this section, we start by reviewing some concepts from information theory that will be referenced throughout the rest of the paper. Then, we lay down a formal statement of the main problem setting.

2.1 Information Theory Background

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and define the limits $f(0) = \lim_{x \to 0^+} f(x)$ and $f'(\infty) = \lim_{x \to \infty} f(x)/x$ (either of which could be infinite). If P and Q are two distributions (probability mass functions) on a common finite set Ω, the f-divergence (Ali and Silvey, 1966; Csiszár, 1967; Polyanskiy and Wu, 2023, Section 7.1) between them is defined as:

$$D_f(P \,\|\, Q) \coloneqq \sum_{x \in \Omega} Q(x) f\left(\frac{P(x)}{Q(x)}\right),$$

with the understanding that $0f(0/0) = 0$ and $0f(a/0) = \lim_{x \to 0^+} xf(a/x) = af'(\infty)$ for $a > 0$. Notable properties of f-divergences include joint convexity in P and Q, non-negativity, and the fact that $D_f(P \,\|\, P) = 0$. Examples for f-divergences used in this work include

- $f(x) = (1/2)|x - 1| \longrightarrow$ total variation distance:

$$\delta(P, Q) \coloneqq \frac{1}{2} \sum\nolimits_x |P(x) - Q(x)| = 1 - \sum\nolimits_x \min\{P(x), Q(x)\}\,.$$

- $f(x) = (1/2)(\sqrt{x} - 1)^2 \longrightarrow$ squared Hellinger distance:

$$H^2(P, Q) \coloneqq \frac{1}{2} \sum\nolimits_x (\sqrt{P(x)} - \sqrt{Q(x)})^2\,.$$

- $f(x) = (x - 1)^2/(x + 1) \longrightarrow$ triangular discrimination (also known as the Vincze–Le Cam divergence, see Sason and Verdú, 2016):

$$\Delta(P, Q) \coloneqq \sum\nolimits_x \frac{(P(x) - Q(x))^2}{P(x) + Q(x)}\,.$$

- $f(x) = x \ln x \longrightarrow$ KL-divergence:

$$D(P \,\|\, Q) \coloneqq \sum\nolimits_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)\,.$$

- $f(x) = (x - 1)^2 \longrightarrow$ chi-squared divergence:

$$\chi^2(P \,\|\, Q) \coloneqq \sum\nolimits_x Q(x)\left(\frac{P(x)}{Q(x)} - 1\right)^2 = \sum\nolimits_x \frac{P(x)^2}{Q(x)} - 1\,.$$

Let X and Y be two discrete random variables mapping Ω to the finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. The quantity defined as

$$D_f(P_{Y|X} \,\|\, Q_{Y|X} \mid P_X) \coloneqq \sum_{x \in \mathcal{X}} P_X(x) D_f(P_{Y|X=x} \,\|\, Q_{Y|X=x})$$

is known as the conditional f-divergence, where a summand corresponding to some $x \in \mathcal{X}$ is set to zero if $P_X(x) = 0$. An immediate property of f-divergences is that if $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$, then

$$D_f(P_{X,Y} \,\|\, Q_{X,Y}) = D_f(P_{Y|X} \,\|\, Q_{Y|X} \mid P_X)\,. \tag{1}$$

When jointly distributed according to $P_{X,Y}$, the mutual information between X and Y based on a given f-divergence (or their mutual f-information) is defined as (Polyanskiy and Wu, 2023, Section 7.8):

$$I_f(X;Y) \coloneqq D_f(P_{X,Y} \,\|\, P_X P_Y)\,,$$

which is the divergence between their joint distribution and the product of the marginals; hence, it is non-negative and symmetric in X and Y. Moreover, when f is strictly convex at unity (as is the case for the above examples), we have that $I_f(X;Y) = 0$ if and only if X and Y are independent (Makur and Zheng, 2020). Via (1), it holds that[2]

$$I_f(X;Y) = D_f(P_{Y|X} \,\|\, P_Y \mid P_X) = \sum\nolimits_x P_X(x) D_f(P_{Y|X=x} \,\|\, \sum\nolimits_{x'} P_X(x') P_{Y|X=x'})\,.$$

The previous identity justifies the overloaded notion $I_f(P_X, P_{Y|X}) \coloneqq I_f(X;Y)$ formulating I_f as a function of P_X and the kernel $P_{Y|X}$.

When the f-divergence of choice is the KL-divergence, we obtain the standard mutual information, denoted simply as $I(X;Y)$, whereas $I_{\chi^2}(X;Y)$ will denote the mutual information based on the chi-squared divergence. The latter is bounded from above by $\min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$ considering that

$$\begin{aligned} I_{\chi^2}(X;Y) &= \sum_{x\colon P_X(x)>0} P_X(x) \sum_{y\colon P_{Y|X=x}(y)>0} P_{Y|X=x}(y) \frac{P_{Y|X=x}(y)}{\sum_{x'\in\mathcal{X}} P_X(x') P_{Y|X=x'}(y)} - 1 \\ &\le \sum_{x\colon P_X(x)>0} P_X(x) \sum_{y\colon P_{Y|X=x}(y)>0} P_{Y|X=x}(y) \frac{P_{Y|X=x}(y)}{P_X(x) P_{Y|X=x}(y)} - 1 \le |\mathcal{X}| - 1\,, \end{aligned}$$

which holds with equality if the distributions $\{P_{Y|X=x}\}_{x\in\mathcal{X}}$ have disjoint supports and P_X has full support. Symmetrically, we also have that $I_{\chi^2}(X;Y) \le |\mathcal{Y}| - 1$. On the other hand, as $D(P \,\|\, Q) \le \log(\chi^2(P \,\|\, Q) + 1)$ (Polyanskiy and Wu, 2023, Section 7.6), we have that $I(X;Y) \le \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$. In particular, $I(X;Y) = \log|\mathcal{X}|$ holds if the distributions $\{P_{Y|X=x}\}_{x\in\mathcal{X}}$ have disjoint supports and P_X is uniform. Another distinction between the two quantities concerns their behaviour as functions of P_X. While for a fixed kernel $P_{Y|X}$, $I(P_X, P_{Y|X})$ is continuous in P_X (Cover and Thomas, 2006, Section 7.3), the same does not hold in general for I_{χ^2}. To see this, consider a simple instance where $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $P_{Y|X=0}(0) = 0.5$, and $P_{Y|X=1}(0) = 1$. If $P_X(0) = \varepsilon$ for some $\varepsilon \in (0,1]$, then

$$I_{\chi^2}(P_X, P_{Y|X}) = \frac{2 - (3/2)\varepsilon}{2 - \varepsilon} - \frac{1}{2} \eqqcolon g(\varepsilon)\,,$$

which satisfies $\lim_{\varepsilon\to 0^+} g(\varepsilon) = 1/2$ even though $I_{\chi^2}(P_X, P_{Y|X}) = 0$ at $\varepsilon = 0$ by definition since X and Y become independent. Moreover, since g is decreasing in the interval $(0,1]$, I_{χ^2} as a function of P_X attains no maximum.

Maximizing the standard mutual information $I(P_X, P_{Y|X})$ in P_X gives rise to what we will refer to as the (KL-)information capacity of the kernel $P_{Y|X}$, denoted as:

$$\mathcal{C}_{\mathrm{KL}}(P_{Y|X}) \coloneqq \max_{P_X \in \mathcal{P}_{\mathcal{X}}} I(P_X, P_{Y|X})\,,$$

[2] By symmetry, this also holds when the roles of X and Y are exchanged.

where $\mathcal{P}_{\mathcal{X}}$ is the set of possible distributions over the elements of $\mathcal{X}$. More practically, $\mathcal{C}_{\mathrm{KL}}(P_{Y|X})$ is better known as the information capacity of the discrete memoryless stationary channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition matrix $P_{Y|X}$ (Cover and Thomas, 2006, Chapter 7; Polyanskiy and Wu, 2023, Chapter 19). This quantity has an operational significance as it quantifies the highest rate per channel use at which information can be reliably sent (Cover and Thomas, 2006, Theorem 7.7.1; Polyanskiy and Wu, 2023, Theorem 19.9). Analogously, we define the χ^2-capacity of $P_{Y|X}$ as:

$$\mathcal{C}_{\chi^2}(P_{Y|X}) \coloneqq \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} I_{\chi^2}(P_X, P_{Y|X}) \,,$$

where we use the supremum in place of the maximum as the latter might not exist per the counterexample provided earlier.

2.2 Problem Setting

Let $\mathscr{X} \coloneqq \{1, \dots, K\}$ denote a set of $K \geq 2$ outcomes, and let $\Theta \coloneqq \{\theta_1, \dots, \theta_N\} \subset \Delta_K$ denote a policy set consisting of $N \geq 2$ distributions over the outcomes, where Δ_K is the probability simplex in $\mathbb{R}^K$ defined as $\{u \in \mathbb{R}^K : \sum_{j=1}^K u(j) = 1 \text{ and } u(j) \geq 0 \, \forall j \in [K]\}$. Hence, for an outcome $x \in \mathscr{X}$ and policy $\theta \in \Theta$, $\theta(x)$ denotes the probability assigned to x by θ. We consider a mediator feedback problem where a learner, possessing full knowledge of the policy set Θ, plays a sequential game with an unknown environment for T rounds. From the environment's characteristic distribution,[3] a latent sequence of loss vectors $(\ell_t)_{t=1}^T$ is drawn at the beginning of the game, where $\ell_t \in [0,1]^K$ maps each outcome to a loss at the t-th round. Ensuingly, the learner sequentially interacts with the environment by selecting at each round t a policy $\vartheta_t \in \Theta$, possibly at random, and subsequently observing the pair $(X_t, \ell_t(X_t))$, where X_t is a random outcome distributed according to ϑ_t. Slightly overloading the notation, we let $\ell_t(\theta)$ denote the expected value (conditioned on ℓ_t) of $\ell_t(X_t)$ had the learner picked policy θ at round t; that is, $\ell_t(\theta) \coloneqq \sum_{x \in \mathscr{X}} \theta(x)\ell_t(x)$. The learner's objective is to minimize their regret, defined as:

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell_t(\vartheta_t)\right] - \min_{\theta \in \Theta} \mathbb{E}\left[\sum_{t=1}^T \ell_t(\theta)\right] \,,$$

where the expectation is taken over both the learner's and the environment's randomization. We use a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to define all random variables. For round $t \in [T]$, let $\mathcal{H}_t \coloneqq (\vartheta_s, X_s, \ell_s(X_s))_{s=1}^t$ denote the interaction history up to the end of round t, and let $\mathcal{F}_t \coloneqq \sigma(\mathcal{H}_t)$ denote the σ-algebra generated by $\mathcal{H}_t$. Accordingly, we define $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and $\mathbb{P}^t(\cdot) \coloneqq \mathbb{P}(\cdot \mid \mathcal{F}_{t-1})$, with $\mathcal{F}_0$ being the trivial σ-algebra. Analogously to (Russo and Van Roy, 2016), we define $I^t(X;Y)$ and $I^t_{\chi^2}(X;Y)$ as the mutual information and the mutual chi-squared-information between (discrete) random variables X and Y with $\mathbb{P}^t$ as the base measure. Notice that these quantities are random variables owing to their dependence on the history.

[3] This formulation subsumes the standard (non-adaptive) adversarial case, where an environment is characterized by a deterministic sequence of loss vectors.

3. The Policy Set Capacity: An Improved Regret Bound for EXP4

Before defining the policy set capacity, we provide some context by briefly reviewing some quantities used in related works to describe the richness of the policy set. McMahan and Streeter (2009) introduce the quantity

$$\mathcal{S}(\Theta) \coloneqq \sum_{x \in \mathcal{X}} \max_{\theta \in \Theta} \theta(x)\,.$$

It is easily verified that $1 \leq \mathcal{S}(\Theta) \leq \min\{K, N\}$, where the lower bound is attained in the limit case when all the policies are identical, and the upper bound is attained either when the policies have disjoint supports or when each outcome is matched with a policy entirely concentrated on that outcome. To get a finer sense of this quantity, we define $\mathcal{V}(\Theta) \coloneqq \mathcal{S}(\Theta) - 1$. Notice then that when there are only two policies, that is, $\Theta = \{\theta_1, \theta_2\}$, $\mathcal{V}$ reduces to the total variation distance between the two distributions:

$$\mathcal{V}(\{\theta_1, \theta_2\}) = \sum\nolimits_x \max\{\theta_1(x), \theta_2(x)\} - 1 = 1 - \sum\nolimits_x \min\{\theta_1(x), \theta_2(x)\} = \delta(\theta_1, \theta_2)\,.$$

More generally, one can obtain the somewhat coarse bound:

$$\mathcal{V}(\Theta) \leq \min_{\alpha \in \Delta_K} \sum\nolimits_\theta \delta(\theta, \alpha)\,,$$

where the distribution minimizing the right-hand side acts as the geometric median of Θ in terms of the total variation distance. This inequality follows from Theorem II.1 in (Guntuboyina, 2011), which for any f-divergence and any $\alpha \in \Delta_K$, provides the (implicit) bound:

$$f(\mathcal{S}(\Theta)) + (N-1) f\left(\frac{N - \mathcal{S}(\Theta)}{N-1}\right) \leq \sum\nolimits_\theta D_f(\theta \,\|\, \alpha)\,.$$

Another relevant quantity is the chi-squared "diameter" of the policy set:

$$d_{\chi^2}(\Theta) \coloneqq \max_{\theta, \theta' \in \Theta} \chi^2(\theta \,\|\, \theta')\,,$$

which is featured in the regret bounds of Sen et al. (2018) and Papini et al. (2019), see Section 4. Though it has no general upper bound, d_{χ^2} can be smaller than $\mathcal{V}$ as shown in the examples section below. Sen et al. (2018) also obtain bounds in terms of another diameter-like quantity based on the logarithm of (one plus) the f-divergence with $f(x) = x\exp(x-1) - 1$, though d_{χ^2} is never larger.

3.1 The Policy Set Capacity

Let ϑ and X be two random variables taking values respectively over Θ and $\mathcal{X}$ such that $\mathbb{P}_{X|\vartheta=\theta}(x) = \theta(x)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. Then, we define the (chi-squared) capacity of the policy set as:

$$\mathcal{C}(\Theta) \coloneqq \mathcal{C}_{\chi^2}(\mathbb{P}_{X|\vartheta})\,,$$

which does not depend on the distribution of ϑ. More explicitly, if we define

$$Q_\tau(\Theta) \coloneqq I_{\chi^2}(\tau, \mathbb{P}_{X|\vartheta}) = \sum\nolimits_\theta \tau(\theta) \chi^2\big(\theta \,\big\|\, \sum\nolimits_{\theta'} \tau(\theta')\theta'\big)\,,$$

for some distribution $\tau \in \mathcal{P}_\Theta$;[4] then,

$$\mathcal{C}(\Theta) = \sup_{\tau \in \mathcal{P}_\Theta} Q_\tau(\Theta)\,.$$

As alluded to before, this definition inspires an interpretation of the policy set as inducing a stationary, memoryless channel defined via the kernel $\mathbb{P}_{X|\vartheta}$. Intuitively, $\mathcal{C}(\Theta)$ can be seen to measure the dependency between ϑ and X maximised over the prior distribution of ϑ. Hence, the more dissimilar the distributions are, the larger this quantity.

Since $\mathcal{C}$ is based on I_{χ^2}, it satisfies $0 \leq \mathcal{C}(\Theta) \leq \min\{K, N\} - 1$, which is the same range as that of $\mathcal{V}$. In particular, much like $\mathcal{V}$, the upper bound is attained either when the policies have disjoint supports or when each outcome is matched with a policy entirely concentrated on that outcome. Moreover, $\mathcal{C}(\Theta) = 0$ if and only if X and ϑ are independent no matter how ϑ is distributed, which requires the policies to be identical. More distinctively, it holds in general that $\mathcal{C}(\Theta) \leq \min\{\mathcal{V}(\Theta), d_{\chi^2}(\Theta)\}$. On the one hand, the (joint) convexity of the chi-squared divergence implies that

$$\mathcal{C}(\Theta) \leq \sup_{\tau \in \mathcal{P}_\Theta} \sum\nolimits_{\theta,\theta'} \tau(\theta)\tau(\theta')\chi^2(\theta \,\|\, \theta') \leq \max_{\theta,\theta'} \chi^2(\theta \,\|\, \theta') = d_{\chi^2}(\Theta)\,.$$

On the other hand, we have that

$$\begin{aligned}
\mathcal{C}(\Theta) &= \sup_{\tau \in \mathcal{P}_\Theta} \sum\nolimits_\theta \tau(\theta) \Big(\sum\nolimits_x \frac{\theta(x)^2}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \Big) \\
&= \sup_{\tau \in \mathcal{P}_\Theta} \sum\nolimits_x \frac{\sum_\theta \tau(\theta)\theta(x)^2}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \\
&\leq \sup_{\tau \in \mathcal{P}_\Theta} \sum\nolimits_x \max_{\theta''} \theta''(x) \frac{\sum_\theta \tau(\theta)\theta(x)}{\sum_{\theta'} \tau(\theta')\theta'(x)} - 1 \\
&= \sum\nolimits_x \max_\theta \theta(x) - 1 = \mathcal{S}(\Theta) - 1 = \mathcal{V}(\Theta)\,.
\end{aligned}$$

3.2 A Regret Bound for EXP4 in Terms of the Capacity

Exp4, detailed in Algorithm 1, adopts a simple and natural approach for tackling mediator feedback problems. Its choice of policy in a given round is drawn from a running distribution over the policies taking an exponential weights form. There, each policy θ is weighted according to a proxy of the sum of its losses so far, where an importance-weighted estimator $\widehat{\ell}_t(\theta)$ replaces the inaccessible $\ell_t(\theta)$. The following theorem provides a regret bound for Exp4 that scales with the policy set capacity. This result improves upon the $\sqrt{\mathcal{S}(\Theta) T \log N}$ bound, seemingly the best available worst-case bound for the considered setting. Further, we instantiate the capacity in the ensuing discussion for three families of policy sets for which the bound of this theorem will be shown to be near-optimal in Section 5. While the proposed learning rate schedule requires exact knowledge of the capacity, this requirement will be lifted in Theorem 2, which also addresses the case when the policies' distributions can vary between rounds.

[4] Inline with previous notation, $\mathcal{P}_\Theta$ denotes the set of possible distributions over the policies.

Algorithm 1 Exp4 (Fixed Policy Set)

```
Input: sequence of learning rates $(\eta_t)_{t=1}^T$
Initialize: $\forall \theta \in \Theta$, $\widehat{\ell}_0(\theta) = 0$
for $t = 1, \dots, T$ do
    Draw $\vartheta_t \sim p_t$, where $p_t(\theta) = \frac{\exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta))}{\sum_{\theta'} \exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta'))}$
    Draw $X_t \sim \vartheta_t$, and observe loss $\ell_t(X_t)$
    $\forall \theta \in \Theta$, set $\widehat{\ell}_t(\theta) = \frac{\theta(X_t)}{\sum_{\theta'} p_t(\theta')\theta'(X_t)} \ell_t(X_t)$
end for
```

Theorem 1 *Algorithm 1 with* $\eta_t = \min\Big\{1, \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)t}}\Big\}$ *satisfies*

$$R_T \le 2 \max\Big\{ \sqrt{e\mathcal{C}(\Theta)T \log N}, \log N \Big\} .$$

Proof Let $\theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E} \sum_{t=1}^T \ell_t(\theta)$. For a policy θ, we define a shifted version of the loss at time t as $\zeta_t(\theta) \coloneqq \sum_x \big(\theta(x) - \psi_t(x)\big)\ell_t(x)$, where $\psi_t(x) \coloneqq \sum_\theta p_t(\theta)\theta(x)$. Thus, $\zeta_t(\theta) = \ell_t(\theta) - \sum_{\theta'} p_t(\theta')\ell_t(\theta')$. Notice that for any two policies θ and θ', $\zeta_t(\theta) - \zeta_t(\theta') = \ell_t(\theta) - \ell_t(\theta')$. Hence, $R_T = \mathbb{E} \sum_t \big(\ell_t(\vartheta_t) - \ell_t(\theta^*)\big) = \mathbb{E} \sum_t \big(\zeta_t(\vartheta_t) - \zeta_t(\theta^*)\big)$. Next, we define $\widehat{\zeta}_t(\theta)$ as an estimate of the shifted loss of θ at time t:

$$\widehat{\zeta}_t(\theta) \coloneqq \big(\theta(X_t) - \psi_t(X_t)\big)\frac{\ell_t(X_t)}{\psi_t(X_t)} = \widehat{\ell}_t(\theta) - \ell_t(X_t) . \tag{2}$$

For convenience, we will sometimes treat the distribution p_t as a vector belonging to the simplex $\Delta_N \subset \mathbb{R}^N$, where its i-th coordinate $p_t(i)$ denotes $p_t(\theta_i)$ for each $i \in [N]$. Analogously, the functions $\widehat{\ell}_t$ and $\widehat{\zeta}_t$ will sometimes be handled as vectors in $\mathbb{R}^N$. Notice that p_t and ψ_t are measurable with respect to $\mathcal{F}_{t-1}$, and that ℓ_t is independent of ϑ_t and X_t conditioned on $\mathcal{F}_{t-1}$. Hence, it holds that $\mathbb{E}_t\zeta_t(\vartheta_t) = \mathbb{E}_t \sum_\theta p_t(\theta)\zeta_t(\theta)$, and that $\mathbb{E}_t\widehat{\zeta}_t(\theta) = \mathbb{E}_t\zeta_t(\theta)$ for any fixed $\theta \in \Theta$. Consequently, thanks to the tower rule and the linearity of expectation, we have that

$$\mathbb{E} \sum\nolimits_t \big(\zeta_t(\vartheta_t) - \zeta_t(\theta^*)\big) = \mathbb{E} \sum\nolimits_t \langle p_t - \mathbf{e}_{\theta^*}, \zeta_t \rangle = \mathbb{E} \sum\nolimits_t \langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t \rangle ,$$

where $\mathbf{e}_{\theta^*} \in \mathbb{R}^N$ is the indicator vector for θ^*.

It is well known (see Shalev-Shwartz et al., 2012, Section 2.7) that for every round t, the definition of p_t in Algorithm 1 is equivalent to

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_N} \; \eta_t \Big\langle \sum\nolimits_{s=1}^{t-1} \widehat{\ell}_s, p \Big\rangle - H(p) , \tag{3}$$

where $H(p) \coloneqq \sum_{i=1}^N p(i) \log(1/p(i))$ is the Shannon entropy of p. Note that for any $p \in \Delta_N$,

$$\Big\langle \sum\nolimits_{s=1}^{t-1} \widehat{\zeta}_s, p \Big\rangle = \Big\langle \sum\nolimits_{s=1}^{t-1} \widehat{\ell}_s, p \Big\rangle - \sum\nolimits_{s=1}^{t-1} \ell_s(X_s) .$$

Hence, by adding constant terms (i.e., not depending on p) to the objective function in (3) and changing the scaling, we can arrive at the following alternative characterization of p_t

for $t \in [T+1]$:

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_N} \left\langle \sum\nolimits_{s=1}^{t-1} \widehat{\zeta}_s, p \right\rangle + \frac{1}{\eta_t}\big(\log N - H(p)\big)\,,$$

which is equivalent to the update rule of the follow the regularized leader (FTRL) algorithm when executed on the losses $(\widehat{\zeta}_t)_{t\in[T]}$ with a decision set Δ_N and a sequence of regularizers $(\phi_t)_{t\in[T+1]}$ where

$$\phi_t(p) = \frac{1}{\eta_t}\big(\log N - H(p)\big) \qquad \forall p \in \Delta_N\,,$$

which is the negative Shannon entropy normalized to the range $[0, \log N]$ and scaled by the learning rate. Let $D_{\phi_t}(\cdot\,;\cdot)$ be the Bregman divergence based on ϕ_t, and set $\eta_{T+1} = \eta_T$. We can then use Lemma 7.14 in (Orabona, 2023) to obtain the following regret bound for FTRL on the estimated shifted losses:

$$\sum\nolimits_t \langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle \le \frac{\log N}{\eta_T} + \frac{1}{2}\sum\nolimits_t \eta_t \langle z_t, \widehat{\zeta}_t^{\,2}\rangle\,,$$

where z_t lies on the line segment between p_t and $\widetilde{p}_{t+1} = \arg\min_{u\in\mathbb{R}^N_{\ge 0}} \langle \widehat{\zeta}_t, u\rangle + D_{\phi_t}(u\,;p_t)$. By its definition, it is easy to show that $\widetilde{p}_{t+1}(i) = p_t(i)\exp(-\eta_t\widehat{\zeta}_t(i))$ for every $i \in [N]$. Notice that $\eta_t\widehat{\zeta}_t(i) \ge -\eta_t\ell_t(X_t) \ge -\eta_t \ge -1$ since $\widehat{\ell}_t$ is non-negative, $\eta_t \in (0,1]$, and $\ell_t(X_t) \le 1$. Hence, it holds for every $i \in [N]$ that $\widetilde{p}_{t+1}(i) \le e\, p_t(i)$, implying that $\langle z_t, \widehat{\zeta}_t^{\,2}\rangle \le e\langle p_t, \widehat{\zeta}_t^{\,2}\rangle$.

Overall, we have shown that

$$R_T \le \frac{\log N}{\eta_T} + \frac{e}{2}\sum\nolimits_t \eta_t\, \mathbb{E}\sum\nolimits_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\,. \tag{4}$$

Now, for every $\theta \in \Theta$ and $t \in [T]$, we have that

$$\begin{aligned}
\mathbb{E}_t\widehat{\zeta}_t(\theta)^2 &= \mathbb{E}_t\big(\theta(X_t) - \psi_t(X_t)\big)^2\frac{\ell_t(X_t)^2}{\psi_t(X_t)^2}\\
&\le \mathbb{E}_t\frac{\big(\theta(X_t) - \psi_t(X_t)\big)^2}{\psi_t(X_t)^2}\\
&= \mathbb{E}_t\sum\nolimits_x \frac{\big(\theta(x) - \psi_t(x)\big)^2}{\psi_t(x)^2}\mathbb{I}\{x = X_t\}\\
&= \sum\nolimits_x \frac{\big(\theta(x) - \psi_t(x)\big)^2}{\psi_t(x)}\\
&= \sum\nolimits_x \psi_t(x)\left(\frac{\theta(x)}{\psi_t(x)} - 1\right)^2 = \chi^2(\theta\,\|\,\psi_t) = \chi^2\big(\theta\,\big\|\,\textstyle\sum_{\theta'}p_t(\theta')\theta'\big)\,,
\end{aligned}$$

where the third equality holds since $\mathbb{E}_t\mathbb{I}\{x = X_t\} = \mathbb{P}^t(x = X_t) = \psi_t(x)$. This implies that

$$\mathbb{E}_t\sum\nolimits_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2 \le \sum\nolimits_\theta p_t(\theta)\chi^2\big(\theta\,\big\|\,\textstyle\sum_{\theta'}p_t(\theta')\theta'\big) = Q_{p_t}(\Theta)\,. \tag{5}$$

Consequently, we arrive at the following bound:

$$R_T \le \frac{\log N}{\eta_T} + \frac{e}{2}\,\mathbb{E}\sum\nolimits_t \eta_t Q_{p_t}(\Theta) \le \frac{\log N}{\eta_T} + \frac{e}{2}\mathcal{C}(\Theta)\sum\nolimits_t \eta_t\,.$$

If $T \geq \frac{\log N}{e\mathcal{C}(\Theta)}$, then $\eta_T = \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)T}}$ and $R_T \leq 2\sqrt{e\mathcal{C}(\Theta)T\log N}$, where we have used that $\eta_t \leq \sqrt{\frac{\log N}{e\mathcal{C}(\Theta)t}}$ for $t \in [T]$ and that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. Otherwise, if $T < \frac{\log N}{e\mathcal{C}(\Theta)}$, then $\eta_1 = \cdots = \eta_T = 1$ and $R_T \leq \log N + \frac{e\mathcal{C}(\Theta)T}{2} \leq 2\log N$. ■

Compared to the standard analysis of Exp4 (see, e.g., Bubeck and Cesa-Bianchi, 2012, Section 4.2) or the analysis of McMahan and Streeter (2009), the main nuance in the proof is the use of a more refined bound on the second moment of the estimated losses, thanks to which, the dependence on the capacity occurs naturally in the regret bound. Moreover, carrying out the analysis in terms of the shifted losses (introduced in the proof) causes the horizon-dependent term in the regret bound to feature $\mathcal{C}(\Theta)$ directly instead of $1 + \mathcal{C}(\Theta)$, which can lead to a substantial improvement whenever $\mathcal{C}(\Theta) \ll 1$. This distinction is particularly relevant, as discussed in the introduction, when the number of policies is very small (e.g., just two), as in that case, a bound scaling with $1 + \mathcal{C}(\Theta)$ would not improve much on the trivial worst-case bound of $\sqrt{NT}$, no matter how benign the policy set is. Interestingly, all prior works seem to suffer from this shortcoming.

3.3 Examples

We now examine the quantity $\mathcal{C}(\Theta)$ for a selection of policy set structures and compare it with related quantities.

3.3.1 Two Policies

We start with the case when the policy set consists of only two policies, i.e., $\Theta = \{\theta_1, \theta_2\}$. As mentioned before, we have that $\mathcal{V}(\Theta) = \delta(\theta_1, \theta_2)$, while $d_{\chi^2} = \max\{\chi^2(\theta_1 \,\|\, \theta_2), \chi^2(\theta_2 \,\|\, \theta_1)\}$. These two quantities are incomparable in general, and this can be seen by specializing the next example to the two policies case. For a fixed $r \in [0,1]$, define $q_r(\theta_1 \,\|\, \theta_2) = Q_\tau(\Theta)$ with $\tau(\theta_1) = r$; hence, $\mathcal{C}(\Theta) = \sup_{r\in[0,1]} q_r(\theta_1 \,\|\, \theta_2) =: \mathcal{C}(\theta_1, \theta_2)$. An explicit form for $q_r(\theta_1 \,\|\, \theta_2)$ is given by:

$$q_r(\theta_1 \,\|\, \theta_2) = r(1-r)\sum\nolimits_x \frac{(\theta_1(x) - \theta_2(x))^2}{r\theta_1(x) + (1-r)\theta_2(x)} \,.$$

As a function of $r \in [0,1]$, $q_r(\theta_1 \,\|\, \theta_2)$ is concave with $q_0(\theta_1 \,\|\, \theta_2) = q_1(\theta_1 \,\|\, \theta_2) = 0$. This quantity is known in the literature as the Vincze–Le Cam divergence of order r (Raginsky, 2016; Makur and Zheng, 2020), which is an f-divergence with $f(x) = \frac{r(1-r)(x-1)^2}{r(x-1)+1}$. Corresponding to $r = 1/2$ is (half) the triangular discrimination Δ, immediately implying a lower bound for $\mathcal{C}$:

$$\mathcal{C}(\theta_1, \theta_2) \geq q_{1/2}(\theta_1 \,\|\, \theta_2) = \frac{1}{2}\sum\nolimits_x \frac{(\theta_1(x) - \theta_2(x))^2}{\theta_1(x) + \theta_2(x)} = \frac{1}{2}\Delta(\theta_1, \theta_2) \,.$$

On the other hand, since $\frac{r(1-r)(x-1)^2}{r(x-1)+1} \leq (\sqrt{x} - 1)^2$ for any $r \in (0,1)$ and $x \in [0, \infty)$, we can bound $\mathcal{C}$ in terms of the squared Hellinger distance H^2:

$$\mathcal{C}(\theta_1, \theta_2) \leq \sum\nolimits_x \theta_2(x)\left(\sqrt{\theta_1(x)/\theta_2(x)} - 1\right)^2 = 2H^2(\theta_1, \theta_2) \,. \tag{6}$$

Combining these observations with known inequalities (Topsoe, 2000), we obtain that

$$\delta^2 \le \frac{\Delta}{2} \le \mathcal{C} \le 2H^2 \le \Delta \le 2\delta\,,$$

which shows that the capacity of two policies is of the same order as the squared Hellinger distance and the triangular discrimination. For the two policies case, we prove in Theorem 5 a lower bound of $\Omega(\sqrt{\mathcal{C}(\theta_1,\theta_2)T})$, which order-wise matches the bound of Theorem 1.

3.3.2 ε-Greedy Policies

Consider now a case in which $N = K$ and each policy θ is associated (one-to-one) with an outcome x_θ such that for any outcome x, $\theta(x) = (1-\varepsilon)/N + \varepsilon\mathbb{I}\{x = x_\theta\}$, where $\varepsilon \in [0,1]$. At $\varepsilon = 0$, all policies collapse to the uniform distribution, and we get that $\mathcal{C}(\Theta) = 0$. On the other hand, when $\varepsilon = 1$, the problem essentially reduces to a standard (unstructured) bandit problem with $\mathcal{C}(\Theta) = N - 1$. Generally, for $\tau \in \mathcal{P}_\Theta$, $Q_\tau(\Theta)$ takes the following form:

$$Q_\tau(\Theta) = \varepsilon^2 \sum\nolimits_\theta \frac{\tau(\theta)(1-\tau(\theta))}{\frac{1-\varepsilon}{N} + \varepsilon\tau(\theta)}\,.$$

For intermediate values of $\varepsilon \in (0,1)$, $Q_\tau(\Theta)$ is a strictly concave function in τ attaining its maximum value at the uniform distribution, entailing that $\mathcal{C}(\Theta) = \varepsilon^2(N-1)$. In comparison, we have that

$$\mathcal{V}(\Theta) = \varepsilon(N-1) \qquad \text{and} \qquad d_{\chi^2}(\Theta) = \frac{\varepsilon(N-2)+2}{\varepsilon(N-1)+1} \cdot \frac{\varepsilon^2}{1-\varepsilon}N\,.$$

Notice that even though d_{χ^2} grows unbounded as ε approaches 1, it can be smaller than $\mathcal{V}$ for small enough ε. In Theorem 6, we prove a lower bound of order $\varepsilon\sqrt{NT}$ for this policy set structure, which matches the upper bound of Theorem 1 up to a logarithmic factor.

3.3.3 M-Supported Uniform Policies

Consider another policy set structure where all policies are uniform distributions over a support of $M \le K$ outcomes. That is, if we denote by $\mathrm{Supp}(\theta)$ the support for policy θ, then we have that $\mathrm{Supp}(\theta) = M$ and for any outcome x, $\theta(x) = (1/M)\mathbb{I}\{x \in \mathrm{Supp}(\theta)\}$. Assume further that each outcome belongs to the support of at least one policy. For this structure, one can verify that $\mathcal{C}(\Theta) = \mathcal{V}(\Theta) = K/M - 1$. In fact, we have that $Q_\tau(\Theta) = \mathcal{C}(\Theta)$ for any $\tau \in \mathcal{P}_\Theta$ with full support. On the other hand, $d_{\chi^2} = \infty$ outside of the trivial case when $M = K$. In Theorem 7, we show that for a special family of M-supported uniform policies (where $N \ge K/M$), the regret of any algorithm is $\Omega\big(\sqrt{(K/M-1)\,T\log(N)/\log(K/M)}\big)$. This lower bound particularly shows that the logarithmic factor in the regret bound of Theorem 1 is at least partly unavoidable.

3.4 A Generalization for Time-Varying Policy Distributions

The next theorem extends the result of Theorem 1 by allowing the distributions of the policies to vary between rounds, modelling in this manner the problem of bandits with expert advice. We rely on an adaptive learning rate schedule, whose form is common in

Algorithm 2 Exp4 (Time-Varying Policy Distributions)

```
**Input:** sequence of learning rates $(\eta_t)_{t=1}^T$
**Initialize:** $\forall \theta \in \Theta$, $\widehat{\ell}_0(\theta) = 0$
**for** $t = 1, \dots, T$ **do**
    Observe distributions $\theta(\cdot; t) \quad \forall \theta \in \Theta$
    Draw $\vartheta_t \sim p_t$, where $p_t(\theta) = \frac{\exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta))}{\sum_{\theta'} \exp(-\eta_t \sum_{s=0}^{t-1} \widehat{\ell}_s(\theta'))}$
    Draw $X_t \sim \vartheta_t$, and observe loss $\ell_t(X_t)$
    $\forall \theta \in \Theta$, set $\widehat{\ell}_t(\theta) = \frac{\theta(X_t;t)}{\sum_{\theta'} p_t(\theta')\theta'(X_t;t)} \ell_t(X_t)$
**end for**
```

the online learning and bandits literature (Auer et al., 2002b; McMahan and Streeter, 2010; Neu, 2015a). The resulting bound replaces the dependence on the (per-round) capacity with the history-conditioned mutual chi-squared-information between the chosen policy and the drawn outcome, recalling that the former is an upper bound for the latter by definition. Additionally, the adopted learning rate in a given round only requires an upper bound on the capacity of the policies' distributions, thus affording one the flexibility of providing a quantity of simpler form like $\mathcal{V}$, or even just $\min\{N, K\}$. In terms of regret, this flexibility is paid for through a solitary additive term depending primarily on the largest of these provided bounds.

Only in the current scope, a member θ of the policy set Θ is not synonymous with a distribution over the outcomes; it serves solely as an identifier for a policy. Along with the sequence of losses, the environment draws for each $\theta \in \Theta$ a sequence of distributions $(\theta(\cdot; t))_{t=1}^T$ at the beginning of the game, where $\theta(x; t)$ is the probability assigned to outcome x by policy θ at round t. The distributions $(\theta(\cdot; t))_{\theta \in \Theta}$ associated with a given round t are revealed to the learner at the beginning of the round. Accordingly, we redefine the interaction history as $\mathcal{H}_t \coloneqq \big((\vartheta_s, X_s, \ell_s(X_s))_{s \in [t]}, (\theta(\cdot; s))_{\theta \in \Theta, s \in [t+1]}\big)$, which includes all the information available to the learner before choosing a policy at round $t + 1$.[5] The definition of the regret remains the same, only that the loss of policy θ at round t is now defined as $\ell_t(\theta) \coloneqq \sum_x \theta(x; t)\ell_t(x)$. For a distribution $\tau \in \mathcal{P}_\Theta$ and round $t \in [T]$, we define

$$Q_{t,\tau} \coloneqq \sum\nolimits_\theta \tau(\theta)\chi^2\big(\theta(\cdot; t) \,\big\|\, \sum\nolimits_{\theta'} \tau(\theta')\theta'(\cdot; t)\big)$$

and $\mathcal{C}_t \coloneqq \sup_{\tau \in \mathcal{P}_\Theta} Q_{t,\tau}$ as the time-varying analogues of Q_τ and $\mathcal{C}$. The following theorem still concerns the plain EXP4 algorithm, reformulated in Algorithm 2 for the time-varying case. Observe that for any round t, $I^t_{\chi^2}(\vartheta_t; X_t) = Q_{t,p_t}$.

Theorem 2 *Let $Z_t \coloneqq \sum_{s=1}^t I^s_{\chi^2}(\vartheta_s; X_s)$ for all $t \in [T]$, and let $(J_t)_{t=1}^T$ be a non-decreasing sequence of non-negative real numbers such that J_t is $\mathcal{F}_{t-1}$-measurable and $J_t \geq \mathcal{C}_t$. Then,*

[5] Consistently, the usage of the filtration $(\mathcal{F}_t)_t$ (and dependent quantities) in the context of the following result refers to this augmented definition of the history.

Algorithm 2 with $\eta_t = \sqrt{\frac{\log N}{\log N + e(Z_{t-1}+J_t)}}$ *satisfies*

$$\begin{aligned} R_T &\le \mathbb{E}\left[2\sqrt{e\sum\nolimits_t I^t_{\chi^2}(\vartheta_t; X_t)\log N} + \log N + \sqrt{eJ_T \log N}\right] \\ &\le \mathbb{E}\left[2\sqrt{e\sum\nolimits_t \mathcal{C}_t \log N} + \log N + \sqrt{eJ_T \log N}\right]. \end{aligned}$$

Proof For any $\theta \in \Theta$ and $t \in [T]$, let, similarly to (2), $\widehat{\zeta}_t(\theta) := \widehat{\ell}_t(\theta) - \ell_t(X_t)$, with $\widehat{\ell}_t(\theta)$ as defined in Algorithm 2. Notice that the distributions $(\theta(\cdot;t))_{\theta\in\Theta}$ are measurable with respect to $\mathcal{F}_{t-1}$. Moreover, the sequence $(\eta_t)_t$ is non-increasing and $\eta_t \le 1$ holds for all rounds. Hence, with the same arguments laid out in the proof of Theorem 1, one can show that

$$R_T \le \mathbb{E}\left[\frac{\log N}{\eta_T} + \frac{e}{2}\sum\nolimits_t \eta_t \sum\nolimits_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2\right].$$

Furthermore, similar to what was shown in the proof of Theorem 1, it holds for every t that $\mathbb{E}_t \sum_\theta p_t(\theta)\widehat{\zeta}_t(\theta)^2 \le Q_{t,p_t} = I^t_{\chi^2}(\vartheta_t; X_t)$. Hence, since η_t is $\mathcal{F}_{t-1}$-measurable, we get that

$$R_T \le \mathbb{E}\left[\frac{\log N}{\eta_T} + \frac{e}{2}\sum\nolimits_t \eta_t I^t_{\chi^2}(\vartheta_t; X_t)\right].$$

We then conclude the proof by bounding the two terms inside the expectation. Starting with the second term, we have that

$$\begin{aligned} \frac{e}{2}\sum\nolimits_t \eta_t I^t_{\chi^2}(\vartheta_t; X_t) &= \frac{\sqrt{\log N}}{2}\sum\nolimits_t \frac{eI^t_{\chi^2}(\vartheta_t; X_t)}{\sqrt{\log N + e\left(Z_{t-1} + J_t\right)}} \\ &\le \frac{\sqrt{\log N}}{2}\sum\nolimits_t \frac{eI^t_{\chi^2}(\vartheta_t; X_t)}{\sqrt{eZ_t}} \le \sqrt{eZ_T \log N}, \end{aligned}$$

where the last inequality follows via Lemma 3.5 in (Auer et al., 2002b). Whereas

$$\frac{\log N}{\eta_T} = \sqrt{\log^2 N + e\left(Z_{T-1} + J_T\right)\log N} \le \sqrt{eZ_T \log N} + \log N + \sqrt{eJ_T \log N}.$$

■

Let $\mathcal{S}_t := \sum_x \max_\theta \theta(x;t)$, and let $\mathcal{V}_t := \mathcal{S}_t - 1$. A reasonable choice is to set $J_t = \max_{s\le t}\mathcal{V}_s$, which would only cause the bound to concede an added term of $\sqrt{e\max_{t\le T}\mathcal{V}_t \log N}$ while lifting the more burdensome requirement of computing the capacity at each round. Notice that the first bound of Theorem 2 depends (in expectation) on the observed sequence of losses through its dependence on the algorithm's decision at each round (i.e., p_t). However, it is unclear whether this bound can take advantage of any particular benign property of the losses when compared with the second bound, which only depends on the policies' distributions. Nevertheless, for what concerns the bandits with expert advice problem, these bounds improve upon the state of the art bound of $\sqrt{\sum_t \mathcal{S}_t \log N}$ reported in (Lattimore and Szepesvári, 2020, Theorem 18.3).

4. Best-of-Both-Worlds Bounds

Besides the adversarial regime considered thus far, we study in this section a more benign setting where the dependence of the regret on the time horizon can be improved. Specifically, we will consider what we will refer to as the adversarially corrupted stochastic regime, where it is assumed that there exists a policy $\xi \in \Theta$ such that for every round t and policy $\theta \neq \xi$,

$$\mathbb{E}\big[\ell_t(\theta) - \ell_t(\xi)\big] \geq \Delta - B_t\,,$$

for some $\Delta \in (0,1]$ and $B_t \geq 0$. Additionally, we define $B \coloneqq \sum_{t=1}^{T} B_t$. This includes, as a special case, the canonical stochastic regime where the loss functions $(\ell_t)_t$ are independently and identically distributed across rounds. Notice that besides the addition of corruption, the more general stochastic regime we consider does not require the losses to be distributed either stationarily or independently. For similar setups, see, for example, (Wei and Luo, 2018; Zimmert and Seldin, 2021; Ito et al., 2022; Dann et al., 2023).

The main result of this section concerns, once again, the Exp4 algorithm, which we show to enjoy BOBW bounds when coupled with a certain learning rate schedule. For the stochastic regime, the algorithm achieves a bound linear in the capacity and only polylogarithmic in the time horizon. Simultaneously, it retains roughly the same worst-case guarantee as that of Theorem 1. This result, provided in the next theorem, is obtained by combining elements from the proof of Theorem 1 with the learning rate schedule and analysis technique used by Ito et al. (2022) in the setting of online learning with strongly observable feedback graphs. Before stating the theorem, we define

$$P(\theta) \coloneqq \sum_{t=1}^{T}(1 - p_t(\theta)) \qquad \text{and} \qquad \overline{P}(\theta) \coloneqq \mathbb{E}\sum_{t=1}^{T}(1 - p_t(\theta))$$

for every policy $\theta \in \Theta$, where $p_t(\theta) = \mathbb{P}^t(\vartheta_t = \theta)$ as specified in Algorithm 1. Moreover, for any distribution p over the policies, we let $H(p) \coloneqq \sum_\theta p(\theta)\log(1/p(\theta))$ denote its Shannon entropy as before.

Theorem 3 *Let* $\gamma = \sqrt{\frac{e\mathcal{C}(\Theta)\log(eT)}{2\log(N)}}$, *and suppose Algorithm 1 is run with* $\eta_t = \min\big\{1, \frac{1}{\beta_t}\big\}$ *for* $t \in [T+1]$, *where* $\beta_1 = \gamma$ *and for* $t \in [T]$, $\beta_{t+1} = \beta_t + \frac{\gamma}{\sqrt{1+\frac{1}{\log N}\sum_{s=1}^{t} H(p_t)}}$. *Then, it holds in general that*

$$R_T \leq 3\sqrt{2e\mathcal{C}(\Theta)T\log(eT)\log(eN)} + \log N\,,$$

whereas in the adversarially corrupted stochastic regime, the algorithm additionally satisfies

$$R_T \leq 36e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N + 4\Delta\,.$$

Proof Let $\theta^* \in \arg\min_{\theta\in\Theta}\mathbb{E}\sum_t \ell_t(\theta)$, which need not coincide with ξ. For policy θ and round t, let $\widehat{\zeta}_t(\theta)$ be defined as in (2). As shown in the proof of Theorem 1,[6] we have that $R_T = \mathbb{E}\sum_t\langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t\rangle$, where $\mathbf{e}_{\theta^*} \in \mathbb{R}^N$ is the indicator vector for θ^*. Also, similar to what

[6] We at times treat p_t and $\widehat{\zeta}_t$ as vectors in $\mathbb{R}^N$ in the manner described before in the proof of Theorem 1.

was argued in that proof, Algorithm 1 produces its predictions according to the following FTRL rule:

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_N} \left\langle \sum\nolimits_{s=1}^{t-1} \widehat{\zeta}_s, p \right\rangle + \phi_t(p)\,,$$

where for $p \in \Delta_N$ and $t \in [T+1]$, $\phi_t(p) = -\frac{1}{\eta_t} H(p)$. Note that the sequences β_t and η_t are increasing and non-increasing, respectively. Hence, we have that $\phi_t(p) - \phi_{t+1}(p) \geq 0$. With this in mind, one can extract the following bound from the proof of Theorem 1 and the proof of Lemma 7.14 in (Orabona, 2023):

$$\begin{aligned}\sum\nolimits_t \langle p_t - \mathbf{e}_{\theta^*}, \widehat{\zeta}_t \rangle &\leq \phi_{T+1}(\mathbf{e}_{\theta^*}) - \phi_1(p_1) + \sum\nolimits_t \big(\phi_t(p_{t+1}) - \phi_{t+1}(p_{t+1})\big) \\ &\qquad\qquad\qquad\qquad\qquad\qquad + \frac{e}{2} \sum\nolimits_t \eta_t \sum\nolimits_\theta p_t(\theta) \widehat{\zeta}_t(\theta)^2 \\ &= \frac{1}{\eta_1} \log N + \sum\nolimits_t \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(p_{t+1}) + \frac{e}{2} \sum\nolimits_t \eta_t \sum\nolimits_\theta p_t(\theta) \widehat{\zeta}_t(\theta)^2 \,.\end{aligned}$$

Since, η_t is measurable with respect to $\mathcal{F}_{t-1}$, we have via (5) that $\mathbb{E}_t \eta_t \sum_\theta p_t(\theta) \widehat{\zeta}_t(\theta)^2 \leq \eta_t Q_{p_t}(\Theta) \leq \eta_t \mathcal{C}(\Theta)$. Consequently, it holds that

$$R_T \leq \frac{1}{\eta_1} \log N + \mathbb{E} \sum\nolimits_t \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) H(p_{t+1}) + \frac{e}{2} \mathcal{C}(\Theta)\, \mathbb{E} \sum\nolimits_t \eta_t \,.$$

For every round t, we clearly have that $\eta_t \leq \frac{1}{\beta_t}$, and that $\frac{1}{\eta_t} = \max\{1, \beta_t\}$. Hence, since $\beta_{t+1} \geq \beta_t$, it holds that $\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} = \big(\max\{1, \beta_{t+1}\} - \max\{1, \beta_t\}\big) \leq \beta_{t+1} - \beta_t$. Consequently,

$$R_T \leq \max\{1, \gamma\} \log N + \mathbb{E} \sum\nolimits_t \big(\beta_{t+1} - \beta_t\big) H(p_{t+1}) + \frac{e}{2} \mathcal{C}(\Theta)\, \mathbb{E} \sum\nolimits_t \frac{1}{\beta_t} \,. \tag{7}$$

The following two facts can be extracted from the proof of Proposition 1 in (Ito et al., 2022):

$$\begin{aligned}\sum\nolimits_t \big(\beta_{t+1} - \beta_t\big) H(p_{t+1}) &\leq 2\gamma \sqrt{\log N} \sqrt{\sum\nolimits_t H(p_t)} \\ \sum\nolimits_t \frac{1}{\beta_t} &\leq \frac{\log(eT)}{\gamma \sqrt{\log N}} \sqrt{\log N + \sum\nolimits_t H(p_t)} \,.\end{aligned}$$

Moreover, Lemma 4 in the same paper entails that $\sum_t H(p_t) \leq P(\xi) \log \frac{eNT}{P(\xi)}$. Plugging these inequalities back into (7) yields that

$$\begin{aligned}R_T \leq \max\{1, \gamma\} \log N &+ 2\gamma \sqrt{\log N}\, \mathbb{E} \sqrt{P(\xi) \log \frac{eNT}{P(\xi)}} \\ &+ \frac{e\mathcal{C}(\Theta) \log(eT)}{2\gamma \sqrt{\log N}}\, \mathbb{E} \sqrt{\log N + P(\xi) \log \frac{eNT}{P(\xi)}} \,.\end{aligned} \tag{8}$$

To obtain the worst-case bound, simply observe that $u \log \frac{eNT}{u}$ is an increasing function in u for $0 < u \leq NT$. Hence, $P(\xi) \log \frac{eNT}{P(\xi)} \leq T \log(eN)$, which, together with (8), implies that

$$R_T \leq \max\{1, \gamma\} \log N + \left(2\gamma \sqrt{\log N} + \frac{e\mathcal{C}(\Theta) \log(eT)}{\gamma \sqrt{\log N}} \right) \sqrt{T \log(eN)}\,,$$

from which the desired bound can be seen to hold after plugging in the value of γ.

Towards proving the second bound, we take an alternative route following the proof of Theorem 4 in (Ito et al., 2022). Namely, we argue that if $P(\xi) \le e$, then $P(\xi)\log\frac{eNT}{P(\xi)} \le e\log(NT)$, otherwise, $P(\xi)\log\frac{eNT}{P(\xi)} \le P(\xi)\log(NT)$. Resuming again from (8), we get that

$$\begin{aligned} R_T &\le \max\{1,\gamma\}\log N + \left(2\gamma\sqrt{\log N} + \frac{e\mathcal{C}(\Theta)\log(eT)}{\gamma\sqrt{\log N}}\right)\sqrt{\log(NT)}\,\mathbb{E}\sqrt{\max\{P(\xi),e\}} \\ &\le \max\left\{\log N, \sqrt{\frac{e}{2}\mathcal{C}(\Theta)\log(eT)\log(N)}\right\} + 2\sqrt{2e\mathcal{C}(\Theta)\log(eT)\log(NT)}\left(\sqrt{\overline{P}(\xi)}+\sqrt{e}\right) \\ &\le \log N + 3e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)}\left(1+\sqrt{\overline{P}(\xi)}\right), \end{aligned}$$

where the second inequality follows after plugging in the value of γ and using Jensen's inequality. For what follows, we define $G_1 \coloneqq \log N$ and $G_2 \coloneqq 3e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)}$.

Now, in the adversarially corrupted stochastic regime, we observe that

$$R_T \ge \sum\nolimits_t \mathbb{E}\big[\ell_t(\theta)-\ell_t(\xi)\big] \ge \Delta\,\mathbb{E}\sum\nolimits_t \mathbb{I}\{\vartheta_t \ne \xi\} - B = \Delta\overline{P}(\xi) - B\,.$$

Combining this with the last bound, we obtain that for any $\lambda > 0$,

$$R_T = (1+\lambda)R_T - \lambda R_T \le (1+\lambda)\Big(G_1 + G_2 + G_2\sqrt{\overline{P}(\xi)}\Big) - \lambda\Delta\overline{P}(\xi) + \lambda B\,.$$

Using the fact that $2a\sqrt{u} - bu \le a^2/b$ for all $u,a,b \ge 0$, we get that

$$\begin{aligned} R_T &\le (1+\lambda)(G_1+G_2) + \lambda B + \frac{(1+\lambda)^2G_2^2}{4\lambda\Delta} \\ &= (1+\lambda)(G_1+G_2) + \frac{G_2^2}{2\Delta} + \lambda\left(B + \frac{G_2^2}{4\Delta}\right) + \frac{G_2^2}{4\lambda\Delta}. \end{aligned}$$

If we choose $\lambda \coloneqq \sqrt{\frac{G_2^2}{4\Delta}\big/\big(B+\frac{G_2^2}{4\Delta}\big)}$, we obtain the following bound:

$$R_T \le 2G_1 + 2G_2 + \frac{G_2^2}{2\Delta} + 2\sqrt{\frac{G_2^2B}{4\Delta} + \frac{G_2^4}{16\Delta^2}} \le 2G_1 + 2G_2 + \frac{G_2^2}{\Delta} + \sqrt{\frac{G_2^2B}{\Delta}}\,,$$

where we have also used the fact that $\lambda \le 1$. Using the definitions of G_1 and G_2, we can conclude that

$$\begin{aligned} R_T &\le 18e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N \\ &\qquad\qquad + 6e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)} \\ &\le 36e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta} + 3e\sqrt{\frac{2\mathcal{C}(\Theta)\log(eT)\log(NT)B}{\Delta}} + 2\log N + 4\Delta\,, \end{aligned}$$

where we used that $6e\sqrt{2\mathcal{C}(\Theta)\log(eT)\log(NT)} \le \max\left\{18e^2\frac{\mathcal{C}(\Theta)\log(eT)\log(NT)}{\Delta}, 4\Delta\right\}$. ■

The appeal of this theorem is that it shows that the simple and fundamental Exp4 algorithm can obtain logarithmic regret in stochastic environments, always scaling with the policy set capacity. Notice that we require the uniqueness of the optimal policy to achieve this, though similar assumptions are common in BOBW works adopting the so-called self-bounding technique applied in the last proof (Wei and Luo, 2018; Zimmert and Seldin, 2021; Ito et al., 2022; Dann et al., 2023). There is, however, small room for improvement in terms of the dependence of both bounds on the time horizon. Compared to Theorem 1, the adversarial bound shown above is worse off by an extra $\sqrt{\log T}$ factor. At the same time, the stochastic regime bound scales as $\log^2 T$ instead of the typical $\log T$ rate. Arguably, these shortcomings are mild for BOBW bounds, especially considering the simplicity of the algorithm.

Nevertheless, the recent work of Dann et al. (2023) offers one way for further honing these bounds. There, a general reduction scheme is proposed, allowing the automatic synthesis of BOBW algorithms starting from traditional algorithms satisfying a certain importance-weighting (iw) stability condition. More precisely, this condition requires that if the algorithm receives feedback in round t only with probability q_t (communicated at the start of the round), it achieves a gracefully degrading bound of $\mathbb{E}\left[\sqrt{c_1 \sum_{t \le t'} 1/q_t} + c_2 \max_{t \le t'} 1/q_t\right]$ on the expected regret at any stopping time t' with some constants c_1 and c_2. Let upd_t be an indicator for whether the feedback is received at round t. For bandits with expert advice (or contextual bandits), Lemma 10 in their paper shows that Exp4 is iw-stable with $c_1 = \mathcal{O}(K \log N)$ and $c_2 = 0$ by scaling the loss estimators with upd_t/q_t and using a simple adaptive learning rate. This leads to BOBW bounds depending on the number of actions K. For our setting, one can combine their analysis with that of Theorem 1 (similarly scaling the estimated shifted losses $\widehat{\zeta}_t$ with upd_t/q_t) yielding that Exp4 is iw-stable with $c_1 = \mathcal{O}(\mathcal{C}(\Theta) \log N)$ and $c_2 = \mathcal{O}(\log N)$ using the learning rate

$$\eta_t = \min\left\{ \min_{s \le t} q_s, \sqrt{\frac{\log N}{e\mathcal{C}(\Theta) \sum_{s \le t} 1/q_s}} \right\}.$$

Hence, Theorems 6 and 11 in (Dann et al., 2023) imply the existence of an algorithm enjoying a bound of $\mathcal{O}\left(\sqrt{\mathcal{C}(\Theta) T \log(N)} + \log(N)\log^2(T)\right)$ for the adversarial regime, and

$$\mathcal{O}\left(\frac{\mathcal{C}(\Theta)\log(T)\log(N)}{\Delta} + \sqrt{\frac{\mathcal{C}(\Theta)\log(T)\log(N)B}{\Delta}} + \log(N)\log(T)\log\left(\frac{B}{\Delta}\right) \right)$$

for the adversarially corrupted stochastic regime. These bounds deliver the sought improvements at the minor cost of scaling the trailing terms in both bounds with $\log T$ factors. On the downside, achieving these bounds requires an arguably laborious and contrived combination of Exp4 with two meta-algorithms.

In the stochastic regime, competing results in the literature mainly include a bound of order $(1 + d_{\chi^2}(\Theta))^2 \log T \log N/\Delta$ in (Sen et al., 2018, Corollary 2),[7] and a bound of order $\sqrt{(1 + d_{\chi^2}(\Theta))T \log(NT)}$ in (Papini et al., 2019, Theorem 2). These bounds fall short of

[7] To be precise, Sen et al. (2018) consider a stochastic version of the bandits with expert advice problem where an expert's recommendation is a function of an i.i.d. context. There, the diameter d_{χ^2} is defined with respect to the conditional chi-squared divergence integrated over the context distribution.

this section's results, primarily considering their dependence on the structure of Θ. Another notable result, though incomparable, is the constant (time-independent) regret bound in (Metelli et al., 2021, Theorem 5.2), which nonetheless requires $d_{\chi^2}(\Theta)$ to be finite.

5. Lower Bounds

In this section, we complement the regret bounds provided thus far by proving lower bounds for the families of policy sets described in Section 3.3. More precisely, for a given policy set, we aim to prove a lower bound for the minimax regret $\inf_\pi \sup_{(\ell_t)_t} R_T$, where π is the player's strategy. To this end, we will consider a class of environments, each identified by a vector $\mu \in [0,1]^K$ such that, for an outcome x, the loss $\ell_t(x)$ at every round t is drawn from a Bernoulli distribution with mean $\mu(x)$ in an i.i.d. manner. For $t \leq T$, recall that $\mathcal{H}_t \coloneqq (\vartheta_s, X_s, \ell_s(X_s))_{s=1}^t$ denotes the interaction history till the end of round t. The player's strategy π can be represented as a sequence of probability kernels $\{\pi_t\}_{t=1}^T$, each mapping the history so far to a distribution over the policies such that ϑ_t is sampled from $\pi_t(\cdot \mid \mathcal{H}_{t-1})$. Hence, under environment μ, it holds that $\mathbb{P}(\vartheta_t = \cdot \mid \mathcal{H}_{t-1}) = \pi_t(\cdot \mid \mathcal{H}_{t-1})$, $\mathbb{P}(X_t = \cdot \mid \mathcal{H}_{t-1}, \vartheta_t) = \vartheta_t(\cdot)$, and $\mathbb{P}(\ell_t(X_t) = \cdot \mid \mathcal{H}_{t-1}, \vartheta_t, X_t) = p_{\mu,X_t}(\cdot)$, where $p_{\mu,x}$ is the loss distribution of outcome x under μ. Consequently, each environment μ (coupled with the player's strategy) induces a probability distribution P_μ on $\mathcal{H}_T$ such that

$$P_\mu\big((\xi_1, x_1, l_1, \ldots, \xi_T, x_T, l_T)\big) = \prod\nolimits_{t=1}^T \pi_t(\xi_t \mid \xi_1, x_1, l_1, \ldots, \xi_{t-1}, x_{t-1}, l_{t-1}) \xi_t(x_t) p_{\mu,x_t}(l_t)\,,$$

for any $(\xi_1, x_1, l_1, \ldots, \xi_T, x_T, l_T) \in (\Theta \times \mathcal{X} \times \{0,1\})^T$. For an environment μ, we define the stochastic regret as:

$$\overline{R}_T(\mu) \coloneqq \max_{\theta^* \in \Theta} \mathbb{E}_\mu \sum_{t=1}^T \sum_{x \in \mathcal{X}} (\vartheta_t(x) - \theta^*(x))\mu(x)\,, \tag{9}$$

where the subscript in $\mathbb{E}_\mu$ emphasizes the dependence on P_μ. For any μ, $\overline{R}_T(\mu)$ is a lower bound for $\sup_{(\ell_t)_t} R_T$. Thus, to prove a lower bound on the minimax regret, it is sufficient to prove a lower bound on $\sup_\mu \overline{R}_T(\mu)$ that holds for any strategy of the player. In the sequel, we will make use of the following lemma, which provides an expression for the KL-divergence between the probability distributions induced by two environments.

Lemma 4 *For a fixed player's strategy, policy set, and time horizon, any two environments μ and μ' satisfy $D(P_\mu \,\|\, P_{\mu'}) = \sum_\theta N_\mu(\theta; T) \sum_x \theta(x) d\big(\mu(x) \,\big\|\, \mu'(x)\big)$, where $N_\mu(\theta; T) \coloneqq \mathbb{E}_\mu \sum_{t=1}^T \mathbb{I}\{\vartheta_t = \theta\}$, and $d(a \,\|\, b)$ is the KL-divergence between two Bernoulli distributions with means a and b.*

Proof Using the chain rule of the KL-divergence, one can obtain that

$$D(P_\mu \,\|\, P_{\mu'}) = \sum\nolimits_t \mathbb{E}_\mu D(p_{\mu,x_t} \,\|\, p_{\mu',x_t})\,.$$

We then use the tower rule and the linearity of expectation to conclude the proof:

$$\begin{aligned}\sum\nolimits_t \mathbb{E}_\mu D(p_{\mu,x_t} \,\|\, p_{\mu',x_t}) &= \sum\nolimits_t \mathbb{E}_\mu \big[\mathbb{E}_\mu \big[D(p_{\mu,x_t} \,\|\, p_{\mu',x_t}) \mid \vartheta_t\big]\big] \\ &= \sum\nolimits_t \mathbb{E}_\mu \sum\nolimits_x \vartheta_t(x) D(p_{\mu,x} \,\|\, p_{\mu',x}) \\ &= \sum\nolimits_\theta N_\mu(\theta; T) \sum\nolimits_x \theta(x) D(p_{\mu,x} \,\|\, p_{\mu',x}) \,.\end{aligned}$$

■

5.1 The Two Policies Case

The following theorem provides a lower bound for the two policies case examined in Section 3.3.1. The proof mostly follows the needle-in-a-haystack technique of Auer et al. (2002a). The key to obtaining this result is a careful choice of the mean loss for each outcome. This choice leads to a lower bound in terms of the Hellinger squared distance between the two policies, which is then related to the capacity via (6). This shows that the bound of Theorem 1 is order-wise unimprovable for this case.

Theorem 5 *Assume that $\Theta = \{\theta_1, \theta_2\}$. Then, for any algorithm and $T \geq \frac{1}{8\log(4/3)H^2(\theta_1,\theta_2)}$, there exists a sequence of losses such that $R_T \geq \frac{1}{13\sqrt{2}}\sqrt{\mathcal{C}(\theta_1,\theta_2)T}$.*

Proof We will consider two environments μ_1 and μ_2 such that for an outcome x, we choose

$$\mu_1(x) \coloneqq \frac{1}{2} - \Delta\frac{\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}} \qquad \text{and} \qquad \mu_2(x) \coloneqq \frac{1}{2} - \Delta\frac{\sqrt{\theta_2(x)} - \sqrt{\theta_1(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}}\,,$$

where $0 \leq \Delta \leq \frac{1}{4}$ is to be tuned later. We posit that $\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}$ is always positive by assuming, without loss of generality, that each outcome is in the support of at least one policy. Additionally, let μ_0 be an environment such that $\mu_0(x) \coloneqq 1/2$ for any outcome x. Note that θ_1 (θ_2) is the optimal policy in μ_1 (μ_2). Indeed,

$$\sum\nolimits_x (\theta_2(x) - \theta_1(x))\mu_1(x) = \Delta \sum\nolimits_x \big(\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}\big)^2 = 2\Delta H^2(\theta_1,\theta_2) > 0\,.$$

Symmetrically, we have that $\sum_x (\theta_1(x) - \theta_2(x))\mu_2(x) = 2\Delta H^2(\theta_1,\theta_2)$. Hence, it holds that

$$\begin{aligned}\overline{R}_T(\mu_1) &= \mathbb{E}_{\mu_1} \sum\nolimits_t \mathbb{I}\{\vartheta_t = \theta_2\} \sum\nolimits_x (\theta_2(x) - \theta_1(x))\mu_1(x) \\ &= 2\Delta H^2(\theta_1,\theta_2)(T - N_{\mu_1}(\theta_1; T)) \\ &\geq 2\Delta H^2(\theta_1,\theta_2)\left(T - N_{\mu_0}(\theta_1; T) - T\sqrt{\frac{1}{2} D\big(P_{\mu_0} \,\big\|\, P_{\mu_1}\big)}\right), \qquad (10)\end{aligned}$$

where the inequality follows by using that $N_{\mu_1}(\theta_1; T) - N_{\mu_0}(\theta_1; T) \leq T\delta(P_{\mu_0}, P_{\mu_1})$ followed by an application of Pinsker's inequality. Note that for $\Delta \leq 1/4$ and $c \coloneqq 8\log(4/3)$,

$$d(\mu_0(x) \,\|\, \mu_1(x)) = d\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} - \Delta\frac{\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}}\right) \leq c\Delta^2 \left(\frac{\sqrt{\theta_1(x)} - \sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)} + \sqrt{\theta_2(x)}}\right)^2 .$$

We also have that

$$\sum\nolimits_x \theta_1(x)\left(\frac{\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}}{\sqrt{\theta_1(x)}+\sqrt{\theta_2(x)}}\right)^2 \le \sum\nolimits_x \big(\sqrt{\theta_1(x)}-\sqrt{\theta_2(x)}\big)^2 = 2H^2(\theta_1,\theta_2)\,,$$

with an analogous inequality holding for θ_2. Combining these observation with Lemma 4 gets us that

$$\begin{aligned} &D\big(P_{\mu_0}\,\|\,P_{\mu_1}\big)\\ &\quad= N_{\mu_0}(\theta_1;T)\sum\nolimits_x \theta_1(x) d\big(\mu_0(x)\,\|\,\mu_1(x)\big) + N_{\mu_0}(\theta_2;T)\sum\nolimits_x \theta_2(x) d\big(\mu_0(x)\,\|\,\mu_1(x)\big)\\ &\quad\le 2c\Delta^2 H^2(\theta_1,\theta_2)(N_{\mu_0}(\theta_1;T)+N_{\mu_0}(\theta_2;T)) = 2c\Delta^2 H^2(\theta_1,\theta_2)T\,. \end{aligned}$$

Plugging back into (10) yields that

$$\overline{R}_T(\mu_1) \ge 2\Delta H^2(\theta_1,\theta_2)\Big(T - N_{\mu_0}(\theta_1;T) - T\Delta\sqrt{cH^2(\theta_1,\theta_2)T}\Big)\,.$$

An analogous bound can be similarly shown to hold for environment μ_2. Hence, we can proceed by arguing that

$$\sup_\mu \overline{R}_T(\mu) \ge \frac{1}{2}(\overline{R}_T(\mu_1)+\overline{R}_T(\mu_2)) \ge \Delta H^2(\theta_1,\theta_2)T\left(1-2\Delta\sqrt{cH^2(\theta_1,\theta_2)T}\right)\,.$$

The theorem then follows by setting $\Delta \coloneqq \frac{1}{4\sqrt{cH^2(\theta_1,\theta_2)T}}$ and using that (see Section 3.3.1) $2H^2(\theta_1,\theta_2) \ge \mathcal{C}(\theta_1,\theta_2)$. Note that the stated condition on T ensures that $\Delta \le 1/4$. ■

5.2 ε-Greedy Policies

Next, we prove a lower bound for the ε-greedy case discussed in Section 3.3.2. Recall that for this case, $\mathcal{C}(\Theta) = \varepsilon^2(N-1)$; hence, the following lower bound matches the bound of Theorem 1 up to a logarithmic factor. Further, we can conclude from this result that for any $g \in [0, N-1]$, there exists a policy set Θ with $|\Theta| = N$ and $\mathcal{C}(\Theta) = g$ for which one has to incur regret of order at least $\sqrt{\mathcal{C}(\Theta)T}$.

Theorem 6 *Assume that the policy set conforms to the ε-greedy structure. Then, for $T \ge \frac{N}{4\log(4/3)}$ and any algorithm, there exists a sequence of losses such that $R_T \ge \frac{1}{18}\varepsilon\sqrt{NT}$.*

Proof We will consider N environments $\{\mu_\theta\}_{\theta\in\Theta}$ such that for environment μ_θ and outcome x, $\mu_\theta(x) \coloneqq 1/2 - \Delta\mathbb{I}\{x = x_\theta\}$, where $0 \le \Delta \le 1/4$ is to be tuned later. Additionally, let μ_0 be an environment such that $\mu_0(x) \coloneqq 1/2$ for any outcome x. Notice that θ is the optimal policy in environment μ_θ. In particular, for $\theta' \in \Theta \setminus \{\theta\}$, we have that

$$\sum\nolimits_x (\theta'(x)-\theta(x))\mu_\theta(x) = \Delta(\theta(x_\theta)-\theta'(x_\theta)) = \Delta\varepsilon\,.$$

Thus,

$$\overline{R}_T(\mu_\theta) = \Delta\varepsilon(T - N_{\mu_\theta}(\theta;T)) \ge \Delta\varepsilon\bigg(T - N_{\mu_0}(\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_0}\,\|\,P_{\mu_\theta}\big)}\bigg)\,, \qquad (11)$$

where the inequality follows by using that $N_{\mu_\theta}(\theta;T) - N_{\mu_0}(\theta;T) \le T\delta(P_{\mu_0}, P_{\mu_\theta})$ followed by an application of Pinsker's inequality. Starting from Lemma 4, we have that

$$\begin{aligned} D\big(P_{\mu_0} \,\big\|\, P_{\mu_\theta}\big) &= \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T) \sum_x \theta'(x) d\big(\mu_0(x) \,\big\|\, \mu_\theta(x)\big) \\ &= \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T)\theta'(x_\theta) d\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} - \Delta\right) \\ &\le c\Delta^2 \textstyle\sum_{\theta'} N_{\mu_0}(\theta';T)\theta'(x_\theta) = c\Delta^2 \left(\frac{1-\varepsilon}{N} T + \varepsilon N_{\mu_0}(\theta;T)\right) , \end{aligned}$$

where the inequality holds for $\Delta \le 1/4$ with $c \coloneqq 8\log(4/3)$. Plugging this result back into (11) allows us to conclude that

$$\begin{aligned} \sup_\mu \overline{R}_T(\mu) &\ge \frac{1}{N} \textstyle\sum_\theta \overline{R}_T(\mu_\theta) \\ &\ge \Delta\varepsilon \left(T - \frac{T}{N} - T\sqrt{\frac{c}{2}\Delta^2 \left(\frac{1-\varepsilon}{N}T + \varepsilon\frac{T}{N}\right)}\right) \ge \Delta\varepsilon T \left(\frac{1}{2} - \Delta\sqrt{\frac{c}{2}\frac{T}{N}}\right) , \end{aligned}$$

where the second inequality uses the concavity of the square root, and the third holds since $N \ge 2$. The theorem then follows by setting $\Delta \coloneqq \frac{1}{4}\sqrt{\frac{2N}{cT}}$ and verifying that the stated condition on T ensures that $\Delta \le 1/4$. ■

5.3 The Multitask Bandits Structure

In this section, we prove a lower bound for a certain structure belonging to the family of M-supported uniform policies described in Section 3.3.3. Let M be a positive integer such that $q \coloneqq K/M$ is an integer greater than or equal to 2. In this structure, the outcomes are divided into M sections, and each policy is a uniform distribution supported over M outcomes such that its support contains an outcome from each section. Assuming that the policy set contains all such policies, we have that $N = (K/M)^M$. For this particular structure, we will index the outcomes according to the section they belong to and their order therein: $\mathcal{X} = \{x_{i,j} : i \in [M], j \in [q]\}$. With this notation, we can describe the policy set as

$$\Theta = \left\{ \theta \in \mathcal{U}_{K,M} : \forall i \in [M], \sum_{j=1}^{q} \theta(x_{i,j}) = \frac{1}{M} \right\} ,$$

where $\mathcal{U}_{K,M}$ is the set of all M-supported uniform distributions over K outcomes. Seeing the outcomes in one section as arms in a bandit game, this problem is, in a sense, equivalent to playing M bandit games simultaneously, with the choice of policy at each round dictating an arm choice for each game. The distinction is that only the loss incurred in a single randomly sampled game is observed, while the player nonetheless aims at minimizing their regret averaged over the M games. This type of structure (albeit with a different type of feedback) is commonly used to prove lower bounds for combinatorial bandits (see, e.g., Audibert et al., 2014). The following theorem, proved in Appendix A, provides a lower bound for our setting.

Theorem 7 *Suppose the policy set conforms to the multi-task structure. Then, for any algorithm and $T \geq \frac{K}{4\log(4/3)}$, there exists a sequence of losses such that $R_T \geq \frac{1}{18}\sqrt{KT}$.*

Recall from Section 3.3.3 that for M-supported uniform policies, $\mathcal{C}(\Theta) = K/M - 1$. Also, note that $M = \log(N)/\log(K/M)$. Thus, the bound given by the theorem is of order $\sqrt{\mathcal{C}(\Theta)T\log(N)/\log(\mathcal{C}(\Theta)+1)}$. The distinguishing value of this lower bound is that it shows that the logarithmic dependence on the number of policies in the bound of Theorem 1 is not entirely spurious and that it becomes increasingly tight as $C(\Theta)$ decreases. This result can be seen as an analogue for our setting of the $\sqrt{KT\log(N)/\log(K))}$ lower bound proved by Seldin and Lugosi (2016) for the problem of bandits with expert advice, noting that the construction of their bound relies on a (time-varying) sequence of deterministic expert recommendations.

6. An Impossibility Result for Linear Bandits

In this section, we establish a separation between the mediator feedback model and the linear bandit model in terms of achievable regret. With $\Theta \subset \Delta_K$ as the action set, we consider a linear bandit problem where upon choosing a policy ϑ_t in round t, the learner directly observes $\ell_t(\vartheta_t) = \sum_x \vartheta_t(x)\ell_t(x)$. This is in contrast to the setting considered thus far, where the learner observes $(X_t, \ell_t(X_t))$ with X_t sampled from the distribution of ϑ_t. The notion of regret we aim to minimize in the linear bandit variant remains the same as before. The main message conveyed by this section's results is that obtaining information regarding individual outcomes is crucial for achieving regret guarantees that reflect the affinity of the policies' distributions.

Concretely, we will consider once again the ε-greedy structure described in Section 3.3.2. Similar to the previous section, we will rely on a class of environments, each identified by a vector $\mu \in [0,1]^K$. However, we will adopt a different scheme for generating the losses for the outcomes. For every round t, let Z_t be a random variable drawn in an i.i.d. manner from a normal distribution $\mathcal{N}(0,\sigma^2)$, with some $\sigma > 0$. Accordingly, for each outcome x, we set $\ell_t(x) := \mu(x) + Z_t$. Hence, the losses assigned to the outcomes in a given round are correlated, see (Cohen et al., 2017) for a similar approach in the combinatorial bandit problem. It follows then that $\ell_t(\vartheta_t) = \langle \vartheta_t, \mu \rangle + Z_t$. In this section, we let P_μ denote the distribution induced by μ (and the player's strategy) over the interaction history in this variant, namely $(\vartheta_1, \ell_1(\vartheta_1), \dots, \vartheta_T, \ell_T(\vartheta_T))$. We will again study the stochastic regret $\overline{R}_T(\mu)$, still defined as in (9).

For the ε-greedy decision set, Theorem 22.1 in (Lattimore and Szepesvári, 2020) implies the existence of an algorithm enjoying a bound of order $\sqrt{NT\log(NT)}$ on the stochastic regret (when $\sigma = 1$), recalling that the members of Θ in this case are N-dimensional vectors. While for the adversarial regret, an upper bound of order $\sqrt{NT\log(N)}$ is achievable, see (Bubeck et al., 2012). The following two results show that, up to factors logarithmic in N and T, the cited bounds are unimprovable, no matter the value of ε. Hence, the growing similarity between the actions—or the shrinking diameter of Θ—as ε approaches 0 cannot be exploited. This is in sharp contrast to the mediator feedback setting, where Theorems 1 and 6 establish the minimax regret to be essentially of order $\varepsilon\sqrt{NT}$ for this family.

Proposition 8 *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, for the class of linear bandit environments described above (with any given $\sigma > 0$), it holds for any algorithm and $T \geq \frac{\sigma^2 N}{\varepsilon^2}$ that $\sup_\mu \overline{R}_T(\mu) \geq \frac{\sigma}{8}\sqrt{NT}$.*

Proof We will consider N environments $\{\mu_\theta\}_{\theta\in\Theta}$ such that for environment μ_θ and outcome x, we set

$$\mu_\theta(x) \coloneqq \frac{1}{2} + \Delta\Big(\frac{1-\varepsilon}{N} - \mathbb{I}\{x = x_\theta\}\Big)\,,$$

where $0 \leq \Delta \leq 1/2$ is to be tuned later. Thus, under environment μ_θ, we have that

$$\ell_t(\vartheta_t) = \sum\nolimits_x \vartheta_t(x)\ell_t(x) = \frac{1}{2} + \Delta\Big(\frac{1-\varepsilon}{N} - \vartheta_t(x_\theta)\Big) + Z_t = \frac{1}{2} - \Delta\varepsilon\mathbb{I}\{\vartheta_t = \theta\} + Z_t\,. \tag{12}$$

Additionally, let μ_0 be an environment such that $\mu_0(x) = 1/2$ for any outcome x, implying that $\ell_t(\vartheta_t) = 1/2 + Z_t$. Notice that θ is the optimal policy in environment μ_θ. In particular, for $\theta' \in \Theta \setminus \{\theta\}$, we have that

$$\sum\nolimits_x (\theta'(x) - \theta(x))\mu_\theta(x) = \Delta(\theta(x_\theta) - \theta'(x_\theta)) = \Delta\varepsilon\,.$$

Hence, it holds that

$$\overline{R}_T(\mu_\theta) = \Delta\varepsilon(T - N_{\mu_\theta}(\theta;T)) \geq \Delta\varepsilon\bigg(T - N_{\mu_0}(\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_0}\,\big\|\,P_{\mu_\theta}\big)}\bigg)\,,$$

where the inequality follows by using that $N_{\mu_\theta}(\theta;T) - N_{\mu_0}(\theta;T) \leq T\delta(P_{\mu_0}, P_{\mu_\theta})$ followed by an application of Pinsker's inequality. Combining the observation in (12) with standard results, see, for example, Exercise 15.8 (b) and Exercise 14.7 in (Lattimore and Szepesvári, 2020); we can express the KL-divergence term as follows:

$$\begin{aligned}D\big(P_{\mu_0}\,\big\|\,P_{\mu_\theta}\big) &= \sum\nolimits_{\theta'} N_{\mu_0}(\theta';T)D\big(\mathcal{N}(1/2,\sigma^2)\,\big\|\,\mathcal{N}(1/2-\Delta\varepsilon\mathbb{I}\{\theta'=\theta\},\sigma^2)\big)\\ &= N_{\mu_0}(\theta;T)D\big(\mathcal{N}(1/2,\sigma^2)\,\big\|\,\mathcal{N}(1/2-\Delta\varepsilon,\sigma^2)\big) = \frac{\Delta^2\varepsilon^2}{2\sigma^2}N_{\mu_0}(\theta;T)\,.\end{aligned}$$

Consequently, we get that

$$\sup_\mu \overline{R}_T(\mu) \geq \frac{1}{N}\sum\nolimits_\theta \overline{R}_T(\mu_\theta) \geq \Delta\varepsilon\left(T - \frac{T}{N} - \frac{\Delta\varepsilon}{2\sigma}T\sqrt{\frac{T}{N}}\right) \geq \Delta\varepsilon T\left(\frac{1}{2} - \frac{\Delta\varepsilon}{2\sigma}\sqrt{\frac{T}{N}}\right)\,,$$

where the second inequality holds by the concavity of the square root, and the third since $N \geq 2$. The proposition then follows by choosing $\Delta \coloneqq \frac{\sigma}{2\varepsilon}\sqrt{\frac{N}{T}}$. Note that the stated condition on T ensures that $\Delta \leq 1/2$. ■

As the construction of this lower bound relied on normally distributed (hence unbounded) losses, a lower bound in the adversarial setting is not immediately implied. Instead, the following theorem (proved in Appendix B) provides the sought bound at the cost of an extra $1/\sqrt{\log(T)}$ factor by combining Proposition 8 with a simple truncation argument due to Cohen et al. (2017). Notice that the resulting bound can be made arbitrarily larger than the mediator feedback guarantee of Theorem 1 by picking a small enough ε and a suitably long horizon.

Algorithm 3 OMD on the Convex Hull of the Policies Under Full-Information

```
Input: learning rate $\eta$, initial distribution $\alpha^* \in \mathrm{co}(\Theta) : \alpha^*(x) > 0 \, \forall x \in \mathcal{X}$
Initialize: $u_1 = \alpha^*$
for $t = 1, \dots, T$ do
    Pick distribution $p_t \in \mathcal{P}_\Theta$ such that $\sum_\theta p_t(\theta)\theta = u_t$
    Draw $\vartheta_t \sim p_t$
    Observe the loss vector $\ell_t$
    Set $u_{t+1} = \arg\min_{u \in \mathrm{co}(\Theta)} \eta \langle u, \ell_t \rangle + D(u \,\|\, u_t)$
end for
```

Theorem 9 *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, under linear bandit feedback, we have that for any algorithm and $T \geq \frac{N}{8\varepsilon^2}$, there exists a sequence of losses (bounded in $[0, 1]$) such that $R_T \geq \frac{1}{64\sqrt{2\log(16T)}}\sqrt{NT}$.*

7. The Full-Information Case

In this last section, we briefly examine a full-information variant of the problem, where the entire loss map $(\ell_t(x))_{x \in \mathcal{X}}$ is observed at every round. One can see this as a variant of the classical prediction with expert advice problem (Cesa-Bianchi et al., 1997), with the outcomes representing the actions (or experts). The distinction is that the learner can only sample from a mixture of the distributions of the policies in Θ. Moreover, the learner competes with the best policy, aiming to minimize the same notion of regret as before. We show in the following theorem that a simple strategy enjoys a regret guarantee depending on the more standard notion of capacity based on the KL-divergence. Precisely, if ϑ and X are two random variables taking values respectively over Θ and $\mathcal{X}$ such that $\mathbb{P}_{X|\vartheta=\theta}(x) = \theta(x)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. Then, we define the KL-capacity of the policy set as:

$$\mathcal{C}_{\mathrm{KL}}(\Theta) \coloneqq \mathcal{C}_{\mathrm{KL}}(\mathbb{P}_{X|\vartheta}) = \max_{\tau \in \mathcal{P}_\Theta} \sum\nolimits_\theta \tau(\theta) D\big(\theta \,\big\|\, \textstyle\sum_{\theta'} \tau(\theta')\theta'\big) .$$

Via (Polyanskiy and Wu, 2023, Corollary 5.6), $\mathcal{C}_{\mathrm{KL}}(\Theta)$ can alternatively be interpreted as the "radius" of Θ in terms of the KL-divergence:

$$\mathcal{C}_{\mathrm{KL}}(\Theta) = \min_{\alpha \in \Delta_K} \max_{\theta \in \Theta} D(\theta \,\|\, \alpha) . \tag{13}$$

The idea of Algorithm 3 is to run an Online Mirror Descent (OMD) algorithm directly on the outcome space restricted to the convex hull of the policy set, henceforth denoted as $\mathrm{co}(\Theta)$. The key to obtaining the following result is a tailored choice of the initial distribution that utilizes the interpretation in (13).

Theorem 10 *Algorithm 3 run with $\alpha^* \in \arg\min_{\alpha \in \Delta_K} \max_{\theta \in \Theta} D(\theta \,\|\, \alpha)$ and $\eta = \sqrt{\frac{2\mathcal{C}_{\mathrm{KL}}(\Theta)}{T}}$ satisfies $R_T \leq \sqrt{2\mathcal{C}_{\mathrm{KL}}(\Theta)T}$.*

Proof Firstly, we note that setting $u_1 = \alpha^*$ is a valid choice since the minimum value of $\max_{\theta \in \Theta} D(\theta \,\|\, \alpha)$ in $\alpha \in \Delta_K$ can only be attained in $\mathrm{co}(\Theta)$; see Theorem 11.6.1 in Cover and

Thomas (2006). Let $\theta^* \in \arg\min_{\theta\in\Theta} \mathbb{E} \sum_t \ell_t(\theta)$. As ℓ_t and ϑ_t are independent given the events up to the end round $t-1$, we have that $R_T = \mathbb{E}\sum_t \langle \vartheta_t - \theta^*, \ell_t \rangle = \mathbb{E}\sum_t \langle u_t - \theta^*, \ell_t\rangle$. A standard regret bound for OMD with the negative entropy regularizer (see, e.g., Lemma 6.16 and the proof of Theorem 10.2 in Orabona, 2023) allows us to conclude that

$$\sum\nolimits_t \langle u_t - \theta^*, \ell_t\rangle \le \frac{D(\theta^* \,\|\, \alpha^*)}{\eta} + \frac{\eta}{2}\sum\nolimits_t \sum\nolimits_x u_t(x)\ell_t^2(x) \le \frac{D(\theta^* \,\|\, \alpha^*)}{\eta} + \frac{\eta}{2}T \le \frac{\mathcal{C}_{\mathrm{KL}}(\Theta)}{\eta} + \frac{\eta}{2}T\,,$$

where the last step follows from (13) and the definition of α^*. The theorem then follows after plugging in the value of η. ■

We remark that $\mathcal{C}_{\mathrm{KL}}(\Theta) \le \min\{\log N, \log K\}$, see Section 2.1. In particular, the first bound is attained when the policies have non-overlapping supports, while the second is attained when each outcome is matched with a policy entirely concentrated on that outcome. Notice that these two cases essentially reduce the problem to a standard prediction with expert advice problem on the policy and outcome spaces, respectively, where the bound of Theorem 10 matches the minimax regret up to constants (Cesa-Bianchi et al., 1997). Beyond these extreme cases, the bound improves as the capacity of the policy set—or its information radius—shrinks.

8. Conclusions and Future Directions

In this paper, we have focused on the mediator feedback framework and studied to what extent the structure of the problem can be exploited by the learner. In particular, we have introduced the policy set capacity as a measure of the effective size (or complexity) of the policy set. For various setups, we have derived new and improved regret bounds for Exp4 featuring the capacity. Further, the lower bounds we provided establish the capacity as a fundamental indicator of the difficulty of the problem for a rich collection of policy sets. Ultimately, we leave open the study of the optimality of the capacity on a more fine-grained level; specifically, whether a regret lower bound in terms of the capacity can be shown to hold for *any* given policy set.

Another direction for improvement is providing bounds that hold with high probability rather than in expectation, noting that prior works on bandits with expert advice (or contextual bandits) such as (Beygelzimer et al., 2011) and (Neu, 2015b) obtained high probability bounds only of order $\sqrt{KT\log N}$. Yet another direction is proving data-dependent (or small loss) bounds, following again previous works on bandits with expert advice (e.g., Allen-Zhu et al., 2018). Concerning the stochastic regime, improving the dependence of the bounds on the sub-optimality gaps (beyond the crude scaling with the smallest gap) is another interesting problem. Finally, for cases when the policy set is very large, achieving similar regret guarantees via more computationally efficient strategies is a worthy direction.

Acknowledgments

The financial support of the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme, is gratefully acknowledged. KE and NCB also acknowledge the support of the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), funded by the NextGenerationEU program within the PNRR scheme and of the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS.

Appendix A. Proof of Theorem 7

Theorem 7 *Suppose the policy set conforms to the multi-task structure. Then, for any algorithm and* $T \geq \frac{K}{4\log(4/3)}$*, there exists a sequence of losses such that* $R_T \geq \frac{1}{18}\sqrt{KT}$ *.*

Proof In the following, we will overload the notation and denote by $x_{i,\theta}$—which belongs to $\{x_{i,j}\}_{j=1}^q$—the outcome chosen by policy θ in section i (i.e. we have that $\theta(x_{i,\theta}) = 1/M$). For each policy θ, we construct an environment μ_θ such that for any outcome x, $\mu_\theta(x) \coloneqq 1/2 - \Delta\mathbb{I}\{x \in \operatorname{Supp}(\theta)\}$, where $0 \leq \Delta \leq 1/4$ is to be tuned later. Moreover, we will also use the following variations of each environment. For $i \in [M]$, let μ_θ^{-i} be an environment such that for any outcome x,

$$\mu_\theta^{-i}(x) \coloneqq \begin{cases} \frac{1}{2}\,, & \text{if } x \in \{x_{i,j}\}_{j=1}^q \\ \mu_\theta(x)\,, & \text{otherwise.} \end{cases}$$

In words, μ_θ^{-i} is identical to μ_θ everywhere except in game i, where all outcomes are assigned a mean loss of $1/2$. For any policy θ, we have that

$$\begin{aligned} \overline{R}_T(\mu_\theta) &= \Delta\,\mathbb{E}_{\mu_\theta}\sum_{t=1}^T\sum_{i=1}^M\big(\theta(x_{i,\theta}) - \vartheta_t(x_{i,\theta})\big) \\ &= \frac{\Delta}{M}\,\mathbb{E}_{\mu_\theta}\sum_{t=1}^T\sum_{i=1}^M\big(1 - \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}\big) \\ &= \frac{\Delta}{M}\sum_{i=1}^M\big(T - N_{\mu_\theta}(i,\theta;T)\big)\,, \end{aligned}$$

where for an environment μ, a policy θ, and a section $i \in [M]$, we define $N_\mu(i,\theta;T) \coloneqq \mathbb{E}_\mu\sum_{t=1}^T \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}$. In words, this counts the expected number of times (under μ) that the chosen policy agrees with θ in section i. Next, we use that for any $i \in [M]$, $N_{\mu_\theta}(i,\theta;T) - N_{\mu_\theta^{-i}}(i,\theta;T) \leq T\delta\big(P_{\mu_\theta^{-i}}, P_{\mu_\theta}\big)$ together with Pinsker's inequality to get that

$$\overline{R}_T(\mu_\theta) \geq \frac{\Delta}{M}\sum_{i=1}^M\left(T - N_{\mu_\theta^{-i}}(i,\theta;T) - T\sqrt{\frac{1}{2}D\big(P_{\mu_\theta^{-i}}\,\big\|\,P_{\mu_\theta}\big)}\right). \tag{14}$$

For bounding the KL-divergence term, we start from Lemma 4:

$$\begin{aligned}
D\big(P_{\mu_\theta^{-i}} \,\big\|\, P_{\mu_\theta}\big) &= \sum_{\theta' \in \Theta} N_{\mu_\theta^{-i}}(\theta'; T) \sum_{x \in \mathcal{X}} \theta'(x) d\big(\mu_\theta^{-i}(x) \,\big\|\, \mu_\theta(x)\big) \\
&= \sum_{\theta' \in \Theta} N_{\mu_\theta^{-i}}(\theta'; T) \theta'(x_{i,\theta}) d\big(\mu_\theta^{-i}(x_{i,\theta}) \,\big\|\, \mu_\theta(x_{i,\theta})\big) \\
&= \frac{1}{M} \sum_{\theta' \in \Theta} \mathbb{I}\{x_{i,\theta'} = x_{i,\theta}\} N_{\mu_\theta^{-i}}(\theta'; T) d\bigg(\frac{1}{2} \,\bigg\|\, \frac{1}{2} - \Delta\bigg) \\
&\leq \frac{c\Delta^2}{M} \sum_{\theta' \in \Theta} \mathbb{I}\{x_{i,\theta'} = x_{i,\theta}\} N_{\mu_\theta^{-i}}(\theta'; T) \\
&= \frac{c\Delta^2}{M} \mathbb{E}_{\mu_\theta^{-i}} \sum_{t=1}^{T} \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\} = \frac{c\Delta^2}{M} N_{\mu_\theta^{-i}}(i, \theta; T)\,,
\end{aligned}$$

where the second equality holds since $x_{i,\theta}$ is the only outcome that does not have the same mean loss in the two environments, and the inequality holds for $\Delta \leq 1/4$ with $c := 8 \log(4/3)$. Plugging back into (14), we get that

$$\overline{R}_T(\mu_\theta) \geq \frac{\Delta}{M} \sum_{i=1}^{M} \left(T - N_{\mu_\theta^{-i}}(i, \theta; T) - T\Delta \sqrt{\frac{c}{2M} N_{\mu_\theta^{-i}}(i, \theta; T)} \right). \tag{15}$$

For what follows, we introduce an extra bit of notation. For each $i \in [M]$, we let $\sim_i$ denote an equivalence relation on the policy set such that for θ, $\theta' \in \Theta$,

$$\theta \sim_i \theta' \iff \forall s \in [M] \backslash \{i\}, x_{s,\theta} = x_{s,\theta'}\,.$$

In words, two policies are equivalent according to $\sim_i$ if they agree everywhere outside of section i. Denote the set of all equivalence classes of $\sim_i$ by $\Theta/\sim_i$, which contains q^{M-1} classes, each containing q policies corresponding to the possible outcome choices in section i. For any $Y \in \Theta/\sim_i$, notice that if $\theta, \theta' \in Y$, then μ_θ^{-i} and $\mu_{\theta'}^{-i}$ denote the same environment, which will be referred to in the sequel as μ_Y^{-i}. Now, notice that for any section i,

$$\sum_{\theta \in \Theta} N_{\mu_\theta^{-i}}(i, \theta; T) = \sum_{Y \in \Theta/\sim_i} \sum_{\theta \in Y} N_{\mu_Y^{-i}}(i, \theta; T) = \sum_{Y \in \Theta/\sim_i} \mathbb{E}_{\mu_Y^{-i}} \sum_{t=1}^{T} \underbrace{\sum_{\theta \in Y} \mathbb{I}\{x_{i,\vartheta_t} = x_{i,\theta}\}}_{=1} = q^{M-1} T\,,$$

whereas $\sum_{\theta\in\Theta}\sqrt{N_{\mu_\theta^{-i}}(i,\theta;T)} \le \sqrt{\sum_{\theta\in\Theta}1^2}\sqrt{\sum_{\theta\in\Theta}N_{\mu_\theta^{-i}}(i,\theta;T)} = q^M\sqrt{\frac{T}{q}}$. Hence, we conclude that

$$\begin{aligned}\sup_\mu \overline{R}_T(\mu) &\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\overline{R}_T(\mu_\theta)\\ &\ge \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\frac{\Delta}{M}\sum_{i=1}^{M}\left(T - N_{\mu_\theta^{-i}}(i,\theta;T) - T\Delta\sqrt{\frac{c}{2M}N_{\mu_\theta^{-i}}(i,\theta;T)}\right)\\ &\ge \frac{\Delta}{M}\sum_{i=1}^{M}\left(T - \frac{1}{|\Theta|}q^M T\left(\frac{1}{q} + \Delta\sqrt{\frac{cT}{2qM}}\right)\right)\\ &= \Delta T\left(1 - \frac{1}{q} - \Delta\sqrt{\frac{cT}{2K}}\right) \overset{q\ge 2}{\ge} \Delta T\left(\frac{1}{2} - \Delta\sqrt{\frac{cT}{2K}}\right)\,.\end{aligned}$$

Plugging $\Delta \coloneqq \frac{1}{4}\sqrt{\frac{2K}{cT}}$ into the previous display proves the theorem after observing that $\frac{1}{16}\sqrt{\frac{2}{c}} \ge \frac{1}{18}$. Lastly, notice that the condition imposed on T ensures that indeed $\Delta \le 1/4$. ■

Appendix B. Proof of Theorem 9

Theorem 9 *Assume that the policy set conforms to the ε-greedy structure with $\varepsilon > 0$. Then, under linear bandit feedback, we have that for any algorithm and $T \ge \frac{N}{8\varepsilon^2}$, there exists a sequence of losses (bounded in $[0,1]$) such that $R_T \ge \frac{1}{64\sqrt{2\log(16T)}}\sqrt{NT}$.*

Proof Building on the result of Proposition 8, we follow the technique used in the proof of Theorem 5 in (Cohen et al., 2017). For the class of linear bandit environments specified in Section 6, we define

$$\widehat{R}_T(\mu) \coloneqq \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))(\mu(x) + Z_t) = \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))\mu(x)$$

$$\widetilde{R}_T(\mu) \coloneqq \max_{\theta^*\in\Theta}\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}(\vartheta_t(x) - \theta^*(x))\operatorname{clip}(\mu(x) + Z_t)\,,$$

where $\operatorname{clip}(a) \coloneqq \max\{\min\{a,1\},0\}$. Notice that $\mathbb{E}\widehat{R}_T(\mu) \ge \overline{R}_T(\mu)$, and that $\sup_{(\ell_t)_t} R_T \ge \sup_\mu \mathbb{E}\widetilde{R}_T(\mu)$ considering sequences of losses $(\ell_t)_t$ bounded in $[0,1]$. We also define the event $A_\mu \coloneqq \{\forall t \in [T],\, x \in \mathcal{X} \colon \operatorname{clip}(\mu(x) + Z_t) = \mu(x) + Z_t\}$. We will consider again the environments $\{\mu_\theta\}_{\theta\in\Theta}$ used in the proof of Proposition 8, recalling that $\mu_\theta(x) \coloneqq 1/2 + \Delta\left((1-\varepsilon)/N - \mathbb{I}\{x = x_\theta\}\right)$ for some $0 \le \Delta \le 1/2$. For any θ, we have that

$$\mathbb{E}\widehat{R}_T(\mu_\theta) = \mathbb{E}\big[\widehat{R}_T(\mu_\theta)\mathbb{I}\{A_{\mu_\theta}\}\big] + \mathbb{E}\big[\widehat{R}_T(\mu_\theta)\mathbb{I}\{A^c_{\mu_\theta}\}\big] \le \mathbb{E}\big[\widetilde{R}_T(\mu_\theta)\big] + \Delta\varepsilon T\mathbb{P}\big(A^c_{\mu_\theta}\big)\,, \tag{16}$$

where we have used the fact that $\widetilde{R}_T(\mu_\theta)$ and $\widehat{R}_T(\mu_\theta)$ are identical when A_{μ_θ} occurs, and that $\widehat{R}_T(\mu_\theta)$ is uniformly bounded by $\Delta\varepsilon T$ (see proof of Proposition 8). Assuming we enforce that $\Delta \leq 1/4$, the event $\{\mathrm{clip}(\mu(x)+Z_t) \neq \mu(x)+Z_t\}$ cannot hold for any outcome unless $|Z_t| > 1/4$. Hence, using a union bound and the fact that $Z_t \sim \mathcal{N}(0,\sigma^2)$, we get that

$$\mathbb{P}\big(A^c_{\mu_\theta}\big) \leq \sum\nolimits_t \mathbb{P}\big(|Z_t| > 1/4\big) \leq 2T \exp\left(\frac{-(1/4)^2}{2\sigma^2}\right) .$$

Combining this with (16) and the fact that $\mathbb{E}\widehat{R}_T(\mu_\theta) \geq \overline{R}_T(\mu_\theta)$ allows us to conclude that

$$\sup_{(\ell_t)_t} R_T \geq \sup_{\mu} \mathbb{E}\widetilde{R}_T(\mu) \geq \frac{1}{N}\sum\nolimits_\theta \mathbb{E}\widetilde{R}_T(\mu_\theta) \geq \frac{1}{N}\sum\nolimits_\theta \overline{R}_T(\mu_\theta) - 2\Delta\varepsilon T^2 \exp\left(\frac{-(1/4)^2}{2\sigma^2}\right) .$$

Setting $\Delta := \frac{\sigma}{2\varepsilon}\sqrt{\frac{N}{T}}$, we obtain from the proof of Proposition 8 that $\frac{1}{N}\sum_\theta \overline{R}_T(\mu_\theta) \geq \frac{\sigma}{8}\sqrt{NT}$. Hence, choosing $\sigma := 1/(4\sqrt{2\log(16T)})$ entails that the required bound. Notice that the condition $T \geq \frac{N}{8\varepsilon^2}$ suffices to ensure that $\Delta \leq 1/4$. ■

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1): 131–142, 1966.

Z. Allen-Zhu, S. Bubeck, and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 186–194. PMLR, 2018.

N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35. PMLR, 2015.

D. Arumugam and B. Van Roy. Deciding what to learn: A rate-distortion approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 373–382. PMLR, 2021.

J. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

J. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002a.

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002b.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26. PMLR, 2011.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23. PMLR, 2012.

S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 41.1–41.14. PMLR, 2012.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

A. Cohen, T. Hazan, and T. Koren. Tight bounds for bandit combinatorial optimization. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 629–642. PMLR, 2017.

T. M. Cover and J. A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

I. Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

C. Dann, C. Wei, and J. Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR, 12–15 Jul 2023.

K. Eldowa, N. Cesa-Bianchi, A. M. Metelli, and M. Restelli. Information-theoretic regret bounds for bandits with fixed expert advice. In *2023 IEEE Information Theory Workshop (ITW)*, pages 30–35, 2023.

A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

S. Ito, T. Tsuchiya, and J. Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 28631–28643. Curran Associates, Inc., 2022.

A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *The Journal of Machine Learning Research*, 21(1):5402–5446, 2020.

T. Lattimore and C. Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

M. Majzoubi, C. Zhang, R. Chari, A. Krishnamurthy, J. Langford, and A. Slivkins. Efficient contextual bandits with continuous actions. In *Advances in Neural Information Processing Systems*, volume 33, pages 349–360. Curran Associates, Inc., 2020.

A. Makur and L. Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56:103–156, 2020.

H. B. McMahan and M. J. Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

H. B. McMahan and M. J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory*, pages 244–256. Omnipress, 2010.

A. M. Metelli, M. Papini, P. D'Oro, and M. Restelli. Policy optimization as online learning with mediator feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8958–8966, 2021.

K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1360–1375. PMLR, 2015a.

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2023.

M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

R. Poiani, A. M. Metelli, and M. Restelli. Pure exploration under mediators' feedback. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023.

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning.* Cambridge University Press, 2023. (draft).

M. Raginsky. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

K. S. Reddy, P. N. Karthik, N. Karamchandani, and J. Nair. Best arm identification in bandits with limited precision sampling. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1466–1471, 2023.

D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

D. Russo and B. Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 47(4):2815–2839, 2022.

I. Sason and S. Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Y. Seldin and G. Lugosi. A lower bound for multi-armed bandits with expert advice. In *The 13th European Workshop on Reinforcement Learning (EWRL)*, 2016.

Y. Seldin, P. Auer, J. Shawe-taylor, R. Ortner, and F. Laviolette. Pac-bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 1683–1691. Curran Associates, Inc., 2011.

R. Sen, K. Shanmugam, and S. Shakkottai. Contextual bandits with stochastic experts. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 852–861. PMLR, 2018.

S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

C. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

Y. Zhu and P. Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27574–27590. PMLR, 2022.

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.