The Sample Complexity of Multiclass and Sparse Contextual Bandits

Liad Erez* LIADEREZ@MAIL.TAU.AC.IL
Tel Aviv University, Israel

Fan Chen* FANCHEN@MIT.EDU
Massachusetts Institute of Technology, USA

Alon Cohen ALONCO@TAUEX.TAU.AC.IL
Tel Aviv University and Google Research Tel Aviv, Israel

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Tel Aviv University and Google Research Tel Aviv, Israel

Yishay Mansour MANSOUR@TAUEX.TAU.AC.IL
Tel Aviv University and Google Research Tel Aviv, Israel

Shay Moran SMORAN@TECHNION.AC.IL
Technion—Israel Institute of Technology, Haifa and Google Research, Tel Aviv, Israel

Alexander Rakhlin RAKHLIN@MIT.EDU
Massachusetts Institute of Technology, USA

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study contextual bandits in the stochastic i.i.d. setting, where a learner observes contexts drawn from an unknown distribution, selects actions from a finite set $\mathcal{A}$, and aims to identify an approximately optimal policy from a given class based on bandit feedback. Motivated by the important special case of bandit multiclass classification with zero-one rewards, we focus on the *s-sparse* setting in which, for every context, the underlying reward vector has L_1-norm at most $s \ll |\mathcal{A}|$. Our main result is the design of algorithms that, with probability at least $1-\delta$, output an ε-optimal policy compared to policy class Π using

$$\widetilde{O}\left(\left(\frac{s}{\varepsilon^2}+\frac{|\mathcal{A}|}{\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right)$$

samples. We further extend this bound to general Natarajan classes and complement it with a matching lower bound (up to logarithmic factors), thereby closing a substantial gap left by prior work (Erez et al., 2024a,b; Erez and Koren, 2025), which incurred an additional $\Theta(|\mathcal{A}|^9)$ dependence.

We obtain these results via two complementary approaches. First, we analyze contextual bandits through the lens of contextual decision making with structured observations, designing an exploration-by-optimization algorithm whose sample complexity is governed by the *decision-estimation coefficient* (DEC; Foster et al., 2021, 2022). We show that, with s-sparse rewards, the induced model class admits a sharp DEC bound that scales with s and directly yields the optimal rate. Since this approach is largely information-theoretic and involves solving complex min-max optimization problems, we also develop a second, more specialized algorithmic method based on a low-variance exploration technique. This approach leads to concrete, tractable algorithms and naturally extends to contextual combinatorial semi-bandits, leading to improved sample complexity guarantees for bandit multiclass list classification.

Keywords: Contextual Bandits, Sample Complexity, Reward Sparsity, Bandit Classification, Multiclass Classification, List Classification

*Equal contribution

1. Introduction

Contextual bandits (Auer et al., 2002a; Langford and Zhang, 2007) are a central model in online learning and decision-making under uncertainty, capturing the fundamental tradeoff between exploration and exploitation in the presence of side information. The framework has been extensively studied over the past two decades, motivated by a wide range of applications including online advertising, recommendation systems, medical decision-making, and adaptive experimentation. From a theoretical perspective, contextual bandits provide a rich setting that interpolates between supervised learning and reinforcement learning, and have driven the development of powerful algorithmic techniques and analytical tools. As a result, the problem has attracted sustained attention in the learning theory community, with a large body of work characterizing achievable performance guarantees under various feedback, structural, and complexity assumptions (e.g., Beygelzimer et al., 2011; Dudik et al., 2011; Agarwal et al., 2014; Foster and Rakhlin, 2020).

We focus on the *stochastic contextual bandit* setting, in which learning proceeds over a sequence of rounds $t = 1, 2, \ldots, T$. In each round, a context $x_t \in \mathcal{X}$ is drawn i.i.d. from an unknown population distribution, and the learner selects an action $a_t \in \mathcal{A}$. The learner then observes a bounded reward corresponding only to the chosen action, drawn from an unknown reward distribution that may depend on the context. Given a policy class $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$, the learner's goal is to compete with the best policy in Π. While much of the contextual bandit literature has focused on minimizing *regret*, which measures the cumulative reward accrued during the learning process, an alternative objective in the stochastic setting is *sample complexity*. Here, success is defined as outputting, after the final round T, a policy whose expected reward is within $\varepsilon > 0$ of that of the best policy in Π with probability at least $1 - \delta$; the sample complexity is the minimal $T = T(\varepsilon, \delta)$ for which this guarantee holds.

One fundamental instance of stochastic contextual bandits, capturing some of the primary historical motivations for this setting, is the *bandit multiclass classification* problem (Kakade et al., 2008; Daniely et al., 2011). Here, contexts correspond to training examples that must be mapped to a finite set of K labels, and Π corresponds to a hypothesis class of classification rules. Upon predicting a label, the learner only observes whether the prediction was correct; notably, *the ground-truth label itself is not revealed*. In this setting, the underlying reward function of primary interest is the zero-one reward, in which a single label is associated with unit reward while all others receive zero reward. In particular, the rewards admit a natural *sparsity* structure: the reward vector at each context has only one non-zero entry. Existing regret-minimizing algorithms for contextual bandits (e.g., Auer et al., 2002b; Dudik et al., 2011; Agarwal et al., 2014) imply a sample complexity of $O((|\mathcal{A}|/\varepsilon^2)\log(|\Pi|/\delta))$, regardless of sparsity, and unfortunately are unable to exploit such a structure to obtain any improvement in rates (unless the sample size is prohibitively large and exceeds $|\Pi|$; see Erez et al., 2024a).

Somewhat surprisingly, focusing exclusively on sample complexity (rather than regret), the sparsity structure of multiclass rewards has recently been shown to enable "fast" sample complexity rates that avoid the leading dependence on $|\mathcal{A}|$ prototypical of bandit problems, and instead *match standard full-information rates* in the primary $\varepsilon \to 0$ regime (Erez et al., 2024a,b; Erez and Koren, 2025).[1] In particular, the dominant term in these fast rates is governed by the sparsity of the problem and depends only (poly-)logarithmically on the number of actions. This improvement, however,

1. A similar phenomenon appears in the regret-minimization version of contextual bandits; however, it exhibits an unavoidable linear dependence on the size of the policy class $|\Pi|$ (Erez et al., 2024a).

comes at the cost of a high-degree polynomial dependence on $|\mathcal{A}|$ in other additive terms. The state-of-the-art bounds, due to Erez and Koren (2025), take the form $\widetilde{O}((s/\varepsilon^2 + |\mathcal{A}|^9)\log(|\Pi|/\delta))$,[2] where s is an upper bound on the L_1-norm of the reward vector at each context. It has been conjectured that the correct dependence on $|\mathcal{A}|$ should be significantly smaller, and the optimal bound should be of the form $\widetilde{\Theta}((s/\varepsilon^2 + |\mathcal{A}|/\varepsilon)\log(|\Pi|/\delta))$. Bridging this substantial gap and obtaining minimax-optimal sample complexity bounds has remained an open question in this line of work.

In this work, we resolve this open question by establishing tight (up to logarithmic factors) sample complexity bounds for contextual bandits with sparse rewards, which directly imply tight bounds for bandit multiclass classification. Specifically, we design algorithms achieving the conjectured sample complexity

$$\widetilde{O}\left(\left(\frac{s}{\varepsilon^2} + \frac{|\mathcal{A}|}{\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right),$$

representing a substantial improvement over prior bounds that include an extraneous $|\mathcal{A}|^9$ dependence.[3] Notably, the scaling with $|\mathcal{A}|$ is linear and comes into the bound only in the lower-order term, whereas the dominant term (as $\varepsilon \to 0$) only depends on the sparsity parameter. Moreover, we show that this bound is minimax optimal up to logarithmic factors.

Our results immediately yield corresponding guarantees for PAC learning in bandit multiclass classification with (possibly infinite) hypothesis classes of finite Natarajan dimension. In addition, our techniques extend to *contextual combinatorial semi-bandits (CCSB)*, and consequently to bandit list classification (Erez and Koren, 2025): a setting in which predictions correspond to subsets (or *lists*) of actions of a fixed size $m \geq 1$, and the reward associated with a subset is the sum of the rewards of its constituent actions. In this setting, we obtain sample complexity upper bounds that significantly improve upon the best previously known results of Erez and Koren (2025), that again contained high-polynomial dependencies in $|\mathcal{A}|$.

We establish our main results via two complementary approaches. The first adopts a general learning-theoretic perspective based on the function approximation framework for interactive decision making and the decision-estimation coefficient (DEC), yielding an information-theoretic characterization of the exploration complexity of sparse contextual bandits. We show that sparsity induces a sharp bound on the DEC, leading directly to optimal sample complexity guarantees and placing bandit multiclass classification within a unified theory of sequential decision making. Complementing this, we develop a more specialized and algorithmic approach based on low-variance exploration, following recent work of Erez et al. (2024b); Erez and Koren (2025). This approach results in concrete, tractable algorithms with closed-form updates and naturally extends to contextual combinatorial semi-bandits. Together, the two approaches provide both a general information-theoretic derivation of the optimal rates and an explicit algorithmic framework for achieving them.

1.1. Summary of contributions

In more detail, our main results are the following. Below, $\mathcal{X}$ is the context space, $\mathcal{A}$ is the set of actions, and $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$ is the policy class.

2. For bandit multiclass classification the polynomial term was improved by Erez and Koren (2025) to $|\mathcal{A}|^7$.
3. While the $|\mathcal{A}|^9$ term is independent of ε, note that $|\mathcal{A}|/\varepsilon \leq |\mathcal{A}|^2 + 1/\varepsilon^2$; thus, any bound containing an additive term larger than $|\mathcal{A}|^2$ is strictly suboptimal.

- As our main result, we develop two methodologies (outlined in Sections 3 and 4) both resulting in contextual bandit algorithms that guarantee with probability at least $1-\delta$ to produce a policy $\hat{\pi}$ that is ε-optimal with respect to the (unknown) reward distribution with sample complexity

$$\widetilde{O}\left(\left(\frac{s}{\varepsilon^2}+\frac{|\mathcal{A}|}{\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right).$$

  provided the reward functions $r \in [0,1]^{\mathcal{A}}$ are s-sparse, namely $\|r\|_1 \le s$. We prove that this bound is tight by providing a matching lower bound (up to logarithmic factors) in Appendix D.
- As a direct corollary, the above result implies that for (single-label, $s = 1$) bandit multiclass classification over a hypothesis class $\mathcal{H}$ of finite Natarajan dimension d_N and a finite label space $\mathcal{Y}$, there exists a PAC learning algorithm with sample complexity

$$\widetilde{O}\left(\left(\frac{1}{\varepsilon^2}+\frac{|\mathcal{Y}|}{\varepsilon}\right)\left(d_N+\log\frac{1}{\delta}\right)\right).$$

- Our main result extends to contextual combinatorial semi-bandits (CCSB), and in turn to bandit multiclass list classification, where we establish a sample complexity bound of

$$\widetilde{O}\left(\left(\frac{s\min(s,m)}{\varepsilon^2}+\frac{K\min(s,m)}{m\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right),$$

  where K is the dimensionality of the action space, m is the fixed subset size and the rewards satisfy $\|r\|_1 \le s$ (with probability 1). Due to space constraints, the details of this result are deferred to Appendix A.

1.2. Technical overview

We now outline the key technical challenges and novelties in establishing our main results.

Information-theoretic upper bound via the DEC framework. Our first approach builds on the general *exploration-by-optimization* (ExO) framework of Lattimore and Szepesvári (2020); Lattimore and Gyorgy (2021); Foster et al. (2022), which reduces sample-efficient exploration in any online decision-making problem to analyzing a complexity measure known as the *decision-estimation coefficient* (DEC). We instantiate this framework in the setting of contextual bandits with structured observations, and develop a refined DEC analysis in the sparse reward setting that yields optimal dependence on both the sparsity parameter and the number of actions. At a high level, the algorithm proceeds by repeatedly solving a min–max optimization problem over exploration-exploitation distributions and estimators, as prescribed by the ExO framework. In each round, the learner selects an exploration policy distribution that balances immediate information gain against estimation error, and collects bandit feedback accordingly.

Our analysis then centers on bounding the DEC of the stochastic contextual bandit problem. The primary technical innovation lies in our upper bound on the DEC for the class of models with expected squared reward bounded by s. To this end, instead of relating the difference in mean rewards $|f^M(a) - f^{\overline{M}}(a)|$ between a candidate model M and a reference model $\overline{M}$ solely to the Hellinger distance, we utilize an inequality that scales with the local reward second moment $\lambda_a = \mathbb{E}_{o\sim\overline{M}(a)}[R(o)^2]$ under the reference model $\overline{M}$. This bound takes the form

$$|f^M(a) - f^{\overline{M}}(a)| \lesssim \sqrt{\lambda_a D^2_{\mathrm{H}}(M(a),\overline{M}(a))} + D^2_{\mathrm{H}}(M(a),\overline{M}(a)),$$

where D_{H}^2 denotes the squared Hellinger distance. To capitalize on this variance-sensitive bound, we construct a specific exploration distribution q that mixes a uniform distribution with a component proportional to the sparsity contribution λ_a of the reference model. This choice is critical as it cancels the variance term λ_a in the error bound, allowing us to sum over actions and bound the DEC by the sparsity level s rather than the ambient cardinality, which directly implies the optimal sample complexity $\widetilde{O}(s/\varepsilon^2 + |\mathcal{A}|/\varepsilon)$.

Algorithmic upper bound via low-variance exploration. In our second method, outlined in Section 4, we adopt a high-level approach similar to that of Erez et al. (2024b); Erez and Koren (2025). Concretely, we design an algorithm that operates in two phases: in the first phase, the algorithm computes an *exploration distribution* $\hat{p} \in \Delta(\Pi)$ whose induced importance-weighted reward estimator has low (independent of $|\mathcal{A}|$) variance *simultaneously for all policies* $\pi \in \Pi$; in the second phase, this exploration distribution is used to uniformly estimate the expected rewards of all policies in Π in a variance-sensitive manner, using Bernstein-type concentration bounds.

For a given policy $\pi \in \Pi$, the variance of the resulting estimator for the reward of π can be written explicitly as

$$\mathbb{E}_{(x,r)}\left[\sum_{a\in\mathcal{A}} r(a)^2 \frac{\mathbf{1}\{\pi(x)=a\}}{\widetilde{Q}_{x,a}(\hat{p})}\right],$$

where $\widetilde{Q}_{x,a}(\hat{p})$ denotes the marginal probability of selecting action a when sampling a policy from $\hat{p}$ (mixed with a uniform distribution). Prior work (Erez et al., 2024b; Erez and Koren, 2025) observed that this variance is, up to constant factors, equal to the partial derivative (with respect to π) of a convex log-barrier potential function $\Phi : \Delta(\Pi) \to \mathbb{R}$, defined as $\Phi(p) = \mathbb{E}[-\sum_{a\in\mathcal{A}} \log \widetilde{Q}_{x,a}(p)]$. Accordingly, the goal of the first phase was achieved by using stochastic convex optimization methods to minimize Φ and thereby to approximately minimize (the L_∞-norm of) its gradient. However, this approach incurs a penalty in sample complexity on the order of $|\mathcal{A}|^9$, stemming from the large smoothness parameter of Φ (which scales as $|\mathcal{A}|^2$) and the high accuracy to which Φ needs to be minimized, and ultimately leads to suboptimal bounds. In fact, minimizing the gradient of Φ using only $\approx |\mathcal{A}|/\varepsilon$ samples—sublinear in the smoothness parameter—seems to be a highly technical and nontrivial challenge.

We take a different route that bypasses the direct optimization of a log-barrier potential. Instead, we borrow a technique from a recent work of Cohen et al. (2025) and employ an online approach based on multiplicative weights updates over adaptively chosen "reward vectors" that correspond to the variance of the induced reward estimator. A key observation is that the cumulative reward with respect to these reward vectors of *any* online algorithm is bounded in expectation by roughly the sparsity parameter s. By leveraging the regret guarantees of multiplicative weights, we show that any benchmark policy $\pi \in \Pi$ also incurs small cumulative reward, which in turn corresponds to a small variance of the resulting estimator. After $T \approx |\mathcal{A}|/\varepsilon$ iterations (and samples), this variance is bounded by $\widetilde{O}(s + \varepsilon|\mathcal{A}|)$, which suffices to obtain the tight sample complexity.

This online approach is inspired by the work of Cohen et al. (2025), who employed related ideas in the full-information setting of multiclass classification to reduce the effective size of the label space by learning a list classifier. One key difference between our approach and theirs is that, while their online reward vectors are binary and correspond to prediction accuracy on previously misclassified examples, our reward vectors encode the variance of a bandit reward estimator and are used toward a fundamentally different objective that arises only in the presence of bandit feedback.

1.3. Related work

Contextual bandits. The contextual multi-armed bandit problem was popularized by Langford and Zhang (2007), and has been extensively studied since. Much of the theoretical work has been focused mainly on the regret minimization setting, where regret bounds of the form $\sqrt{|\mathcal{A}|T}$ shown for stochastic environments (Dudik et al., 2011; Agarwal et al., 2014) as well as for adversarial environments (Auer et al., 2002b; McMahan and Streeter, 2009; Beygelzimer et al., 2011). The problem has been studied in numerous previous works in the function approximation framework (Chu et al., 2011; Filippi et al., 2010; Li et al., 2017; Foster et al., 2018; Foster and Rakhlin, 2020; Foster et al., 2020). The role of reward sparsity in contextual bandits has received attention more recently, with a particular focus on bandit multiclass classification. Erez et al. (2024a) investigate the role of sparsity in online regret minimization and show that for moderately large policy classes the leading term of $\sqrt{|\mathcal{A}|T}$ is unavoidable in general. Erez et al. (2024b) and Erez and Koren (2025) later studied the effect of sparsity on sample complexity for PAC learning and showed that the usual dependence of $|\mathcal{A}|/\varepsilon^2$ can in fact be improved to s/ε^2, but their results included suboptimal $\mathrm{poly}(|\mathcal{A}|)$ additive terms.

Interactive Decision Making and the DEC framework. A recent line of work develops a unified, complexity-theoretic view of sequential decision making with function approximation via the *decision-estimation coefficient (DEC)*. Foster et al. (2021) introduces the DEC as a fundamental measure of statistical complexity for interactive decision problems, encompassing structured bandits and reinforcement learning, and provide matching upper and lower bounds (via the Estimation-to-Decisions meta-principle) that elevate the DEC to an analogue of VC/Rademacher complexity for interactive learning. Subsequent works sharpen this theory quantitatively, including tighter guarantees through refined DEC variants (Foster et al., 2023b; Chen et al., 2024) and generalization to various learning goals (Chen et al., 2022; Foster et al., 2023b,a; Glasgow and Rakhlin, 2023; Chen and Rakhlin, 2025; Liu et al., 2025a). The framework has also been instantiated to derive concrete guarantees for reinforcement learning with general function approximation, most notably in model-free settings (Foster et al., 2023c; Liu et al., 2025b). Parallel efforts also study more instance-adaptive notions of complexity (e.g., allocation-based coefficients) that aim at instance-optimality in interactive decision making, complementing the DEC viewpoint (Wagenmaker and Foster, 2023).

Combinatorial semi-bandits. This problem was introduced by György et al. (2007) in the context of online shortest paths, and has since received significant attention in the bandit literature, primarily within the regret minimization framework (Audibert et al., 2014; Wen et al., 2015; Kveton et al., 2015; Neu, 2015; Wei and Luo, 2018; Ito, 2021, etc.). For adversarial losses, Audibert et al. (2014) established a regret bound of $O(\sqrt{mKT})$. We emphasize that in the original formulation the set of available predictions is an arbitrary subset of $\{0,1\}^K$, whereas in our work we restrict attention to m-sets, where the available predictions are all subsets of $\{0,1\}^K$ of size m. In this setting, Lattimore et al. (2018) showed that the $O(\sqrt{mKT})$ regret bound is optimal. The contextual combinatorial semi-bandit (CCSB) problem has been studied in several prior works (Qin et al., 2014; Wen et al., 2015; Takemura et al., 2021; Zierahn et al., 2023), largely focusing on the setting in which the mappings from contexts to rewards are linear functions with additive noise. The works most closely related to ours are Kale et al. (2010); Krishnamurthy et al. (2016), which consider finite, unstructured policy classes and derive regret bounds of order $O(\sqrt{mKT\log|\Pi|})$. Most recently, Erez and Koren (2025)

studied sample complexity in the presence of sparse rewards and bandit multiclass list classification, and is the most directly relevant to our work.

Bandit multiclass classification. This setting was originally introduced by Kakade et al. (2008), with Daniely et al. (2011) characterizing realizable deterministic learnability by the bandit Littlestone dimension. These results were extended by Daniely and Helbertal (2013), who demonstrated that the bandit Littlestone dimension characterizes online learnability whenever the label set $\mathcal{Y}$ is finite, and were further generalized to infinite label sets by Raman et al. (2023). Several prior works (Auer and Long, 1999; Daniely et al., 2011; Long, 2020) studied the price of bandit feedback in the realizable setting, with Filmus et al. (2024) showing that, for randomized learners, this price is bounded by an $O(|\mathcal{Y}|)$ factor relative to the full-information setting. The theoretical framework of multiclass list classification was first introduced by Brukhim et al. (2022). Charikar and Pabbaraju (2023) characterized PAC learnability for multiclass list classification by generalizing the DS-dimension (Daniely and Shalev-Shwartz, 2014), Moran et al. (2023) studied the regret minimization setting and characterized learnability via a generalization of the Littlestone dimension, and Hanneke et al. (2024) investigated uniform convergence and sample compression in this setting.

2. Problem setup

Contextual bandits. We consider a decision making task over a (possibly infinite) *context space* $\mathcal{X}$ and a (finite) *action set* $\mathcal{A}$. A stochastic contextual bandit instance is specified by an unknown joint distribution $\mathcal{D}$ over $\mathcal{X} \times [0,1]^{\mathcal{A}}$. Given $x \in \mathcal{X}$ we denote by $\mathcal{D}_x$ the reward distribution over $[0,1]^{\mathcal{A}}$ conditioned on x. The learner interacts with the environment sequentially according to the following protocol:

- The environment draws $(x_t, r_t) \sim \mathcal{D}$ and x_t is revealed to the learner.
- The learner selects $a_t \in \mathcal{A}$ and observes $r_t(a_t)$.

For every *policy* $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ we define the *population reward* $R_{\mathcal{D}}(\pi) := \mathbb{E}_{(x,r)\sim\mathcal{D}, a\sim\pi(x)}[r(a)]$. Given a *policy class* $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$, the learner's goal is to produce a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ which is ε-optimal with respect to Π. Namely, given $\varepsilon, \delta \in (0,1)$, the output policy must satisfy, with probability at least $1-\delta$, $\max_{\pi^\star \in \Pi} R_{\mathcal{D}}(\pi^\star) - R_{\mathcal{D}}(\pi) \le \varepsilon$. The *sample complexity* of the learner is defined as the smallest number of interaction rounds (as a function of ε, δ) sufficient to achieve this guarantee.

Reward sparsity. We assume that contextual rewards are *s-sparse*, namely

$$\mathbb{P}_{(x,r)\sim\mathcal{D}}[\|r\|_1 \le s] = 1,$$

where $1 \le s \le |\mathcal{A}|$ is the sparsity parameter. We note that some of our results in fact hold under a slightly weaker assumption, where the rewards are bounded by s in squared L_2-norm (this will be noted wherever relevant).

Example: Bandit multiclass classification. This is a special case of combinatorial bandits with sparse rewards where the rewards are constrained to one-hot vectors (i.e. zero-one rewards), and in particular, are s-sparse with $s = 1$. In this setting, the action set $\mathcal{A}$ is referred to as the *label space* and is denoted by $\mathcal{Y}$, and Π is referred to as the *hypothesis class* and is denoted by $\mathcal{H}$.

3. Information-theoretic upper bound via the DEC framework

In this section, we establish a generic upper bound in a more general setting introduced by Foster et al. (2021) and referred to as *contextual decision making*, through the Exploration-by-Optimization technique (Lattimore and Szepesvári, 2020; Lattimore and Gyorgy, 2021; Foster et al., 2022).

Contextual decision making. This framework generalizes the contextual bandit setting as follows. Let $\mathcal{O}$ denote the *observation space* and let $\mathcal{M} \subseteq \mathcal{A} \to \Delta(\mathcal{O})$ be a convex *model class*, known to the learner. We consider an unknown stochastic environment specified by $\mathcal{D} = (\rho, \{M^\star_x\}_{x\in\mathcal{X}})$ where ρ is a distribution over $\mathcal{X}$ and $M^\star_x \in \mathcal{M}$ for all $x \in \mathcal{X}$. The interaction protocol is as follows; for each $t = 1, \ldots, T$:

- The environment draws $x_t \sim \rho$ and x_t is revealed to the learner.
- The learner selects $a_t \in \mathcal{A}$ and observes $o_t \sim M^\star_{x_t}(a_t)$.[4]

The learning objective is specified by a reward function $R : \mathcal{O} \to [0, 1]$. Given a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, the population reward of π is defined as $R_{\mathcal{D}}(\pi) := \mathbb{E}_{x\sim\rho, a\sim\pi(x), o\sim M^\star_x(a)}[R(o)]$. The objective is to output a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ whose population reward is ε-optimal with respect to an underlying policy class Π with high probability.

Algorithm 1 Exploration-by-Optimization (ExO^+)

Require: Model class $\mathcal{M}$, comparator class Π, prior $W_1 = \mathrm{Unif}(\Pi)$, parameter $\gamma > 0$.

1: **for** $t = 1, \ldots, T$ **do**

2: Observe $x_t \in \mathcal{X}$ and compute the weight distribution $w_t = W_t|_{x_t} \in \Delta(\mathcal{A})$ as in (3).

3: Solve the *exploration-by-optimization* objective:

$$(p_t, q_t, \xi_t) \leftarrow \operatorname*{argmin}_{p,q\in\Delta(\mathcal{A}),\, \xi\in\Xi} \Gamma_{w_t,\gamma}(p, q, \xi) \tag{1}$$

4: Sample $a_t \sim q_t$, observe $o_t \sim M^\star_{x_t}(a_t)$ and perform exponential-weight update:

$$W_{t+1}(\pi) \;\propto_\pi\; W_t(\pi) \exp(\xi_t(\pi(x_t); a_t, o_t)) \tag{2}$$

5: **end for**

6: **return** $\widehat{p} : \mathcal{X} \to \Delta(\mathcal{A})$ defined in (6).

The algorithm. The algorithm, ExO^+, is described in Algorithm 1. At each round t, it maintains a reference distribution $W_t \in \Delta(\Pi)$, and uses it to obtain a joint *exploration-exploitation* distribution $p_t, q_t \in \Delta(\mathcal{A})$ and a weight function $\xi_t \in \Xi := (\mathcal{A} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R})$, by solving a joint minimax optimization problem based on the *exploration-by-optimization* objective:

- For any $W \in \Delta(\Pi)$ and $x \in \mathcal{X}$, we define the marginal distribution $W|_x \in \Delta(\mathcal{A})$ as

$$W|_x(a) = \mathbb{P}_{\pi\sim W}(\pi(x) = a), \qquad \forall a \in \mathcal{A}. \tag{3}$$

4. In the function approximation literature, the assumption that the observations are generated by a model from $\mathcal{M}$ is often referred to as *realizability*; in our case this is simply a notation, as $\mathcal{M}$ will be the class of all possible models (environments) with s-sparse rewards.

- Defining

$$\begin{aligned}\Gamma_{w,\gamma}(p,q,\xi;M,a^\star) &:= \mathbb{E}_{a\sim p}\left[f^M(a^\star)-f^M(a)\right] \\ &\quad -\gamma\mathbb{E}_{a\sim q}\mathbb{E}_{o\sim M(a)}\mathbb{E}_{a'\sim w}\left[1-\exp\left(\xi(a';a,o)-\xi(a^\star;a,o)\right)\right]; \\ \Gamma_{w,\gamma}(p,q,\xi) &:= \sup_{M\in\mathcal{M},a^\star\in\mathcal{A}}\Gamma_\gamma(p,q,\xi;M,a^\star),\end{aligned} \tag{4}$$

  the algorithm solves

$$(p_t,q_t,\xi_t)\leftarrow \operatorname*{argmin}_{p,q\in\Delta(\mathcal{A}),\xi\in\Xi}\Gamma_{w_t,\gamma}(p,q,\xi).$$

- The algorithm then selects $a_t\sim q_t$ from the exploration distribution and observes o_t from the environment. Finally, the algorithm updates the reference distribution by performing the exponential weight update (2) with weight function ξ_t.

Output policy. At the end of the sequential interaction, the algorithm outputs a randomized policy $\widehat{\pi}:\mathcal{X}\to\Delta(\mathcal{A})$ described as follows. Note that for each step $t\in[T]$, Algorithm 1 only computes (p_t,q_t,ξ_t) based on the given x_t. By the standard online-to-batch conversion, we consider the following output rule:

- For each $t\in[T]$, we define $P_t:\mathcal{X}\to\Delta(\mathcal{A})$ as follows: For any $x\in\mathcal{X}$, we solve

$$(p,q,\xi)\leftarrow \operatorname*{argmin}_{p,q\in\Delta(\mathcal{A}),\xi\in\Xi}\Gamma_{W_t|_x,\gamma}(p,q,\xi), \tag{5}$$

  and set $P_t(x)=p$. Recall that $W_t|_x\in\Delta(\mathcal{A})$ is defined in Eq. (3).

- The output $\widehat{\pi}\in(\mathcal{X}\to\Delta(\mathcal{A}))$ is then defined, for any $x\in\mathcal{X}$, as

$$\widehat{\pi}(x)=\frac{1}{T}\sum_{t=1}^{T}P_t(x). \tag{6}$$

Guarantee. To state the generic guarantee of Algorithm 1, we first introduce the notion of Decision-Estimation Coefficient (DEC) (Foster et al., 2021, 2022). For any model $M\in\mathcal{M}$, the corresponding value function is given by $f^M(a):=\mathbb{E}_{o\sim M(a)}[R(o)]$, and the optimal action is $a_M=\operatorname{argmax}_{a\in\mathcal{A}}f^M(a)$. For any reference model $\overline{M}\in\operatorname{co}(\mathcal{M})$, we define

$$\textsf{p-dec}^{\mathrm{o}}_\gamma(\mathcal{M},\overline{M}):=\inf_{p,q\in\Delta(\mathcal{A})}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{a\sim p}[f^M(a_M)-f^M(a)]-\gamma\mathbb{E}_{a\sim q}D^2_{\mathrm{H}}\left(M(a),\overline{M}(a)\right)\right\}, \tag{7}$$

where $D^2_{\mathrm{H}}(P,Q):=\frac{1}{2}\int_{\mathcal{Z}}(\sqrt{P(dz)}-\sqrt{Q(dz)})^2$ is the squared Hellinger distance between distributions P,Q over $\mathcal{Z}$. The (offset) DEC (Foster et al., 2021) of $\mathcal{M}$ is then defined as

$$\textsf{p-dec}^{\mathrm{o}}_\gamma(\mathcal{M}):=\sup_{\overline{M}\in\operatorname{co}(\mathcal{M})}\textsf{p-dec}^{\mathrm{o}}_\gamma(\mathcal{M},\overline{M}).$$

The following theorem is an adaptation of the results of Foster et al. (2022) to the contextual decision making setting. We prove this result in Appendix B for completeness.

Theorem 1 *With probability at least* $1-\delta$*, Algorithm 1 outputs* $\widehat{\pi}:\mathcal{X}\to\Delta(\mathcal{A})$ *such that*

$$\max_{\pi^\star\in\Pi} R_D(\pi^\star)-R_D(\widehat{\pi})\le \textsf{p-dec}^{\textsf{o}}_{\gamma/8}(\mathcal{M})+\frac{4\gamma\log(|\Pi|/\delta)}{T}+O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

To apply the above result to the sparse reward setting, we only need to bound the DEC $\textsf{p-dec}^{\textsf{o}}_{\gamma}(\mathcal{M})$.[5] This results in the following theorem which constitutes the main result for this section, and whose proof can be found in Appendix B.

Theorem 2 *Suppose that* $\mathcal{M}=\left\{M:\sum_{a\in\mathcal{A}}\mathbb{E}_{o\sim M(a)}[R(o)^2]\le s\right\}$ *is the class of all models with* s*-sparse rewards. Then it holds that*

$$\textsf{p-dec}^{\textsf{o}}_{\gamma}(\mathcal{M})\lesssim\frac{s}{\gamma},\qquad\forall\gamma\ge 32|\mathcal{A}|.$$

Then, by Theorem 1, we can instantiate Algorithm 1 so that with probability at least $1-\delta$*, with*

$$T=O\left(\left(\frac{s}{\varepsilon^2}+\frac{|\mathcal{A}|}{\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right),$$

it outputs $\widehat{\pi}:\mathcal{X}\to\Delta(\mathcal{A})$ *such that* $\max_{\pi^\star\in\Pi}R_D(\pi^\star)-R_D(\widehat{\pi})\le\varepsilon$.

To see how the fact that $\textsf{p-dec}^{\textsf{o}}_{\gamma}(\mathcal{M})\lesssim s/\gamma$ implies the final sample complexity bound, we set $\gamma\simeq\max\{|\mathcal{A}|,s/\varepsilon\}$. If $|\mathcal{A}|\ge s/\varepsilon$, the leading term in the final sample complexity bound becomes $|\mathcal{A}|/\varepsilon$ giving an error rate of ε, and if $|\mathcal{A}|\le s/\varepsilon$ the leading term is s/ε^2 again resulting in an error rate of ε.

We also show that this rate is optimal up to logarithmic factors by proving a matching lower bound in Appendix D.

Algorithm 1 is adapted from the original $\textsf{ExO}^+$ algorithm (Foster et al., 2022) to the contextual setting. The adaption allows us to only solve an optimization problem (1) over the action space $\mathcal{A}$ (instead of the policy space Π). In addition, while the problem (1) can be non-convex, we can adopt the re-parametrization $\xi(a';a,o)=\widetilde{\xi}(a';a,o)/q(a)$ so that $\widetilde{\Gamma}_{w,\gamma}(p,q,\widetilde{\xi})=\Gamma_{w,\gamma}(p,q,\widetilde{\xi}/q)$ is convex (see Foster et al., 2022). Therefore, assuming that the inner maximization problem can be solved efficiently, the outer minimization problem can be then solved by standard convex optimization methods (e.g., sub-gradient methods). This implies that the optimization problem Eq. (1) can be approximately solved. However, when $|\mathcal{A}|$ is large, we may still want to avoid solving the problem (1), motivating our alternative method described in the next section. Furthermore, to evaluate the output policy $\widehat{\pi}(x)\in\Delta(\mathcal{A})$ at a given context x, we also have to solve the optimization problem T times, making it computationally challenging to employ.

4. Algorithmic upper bound via low-variance exploration

In this section we present an alternative method for obtaining tight rates in contextual bandits with sparse rewards over a finite policy class, detailed in Algorithm 2. Compared to our previous approach, this method has the benefit of admitting an explicit algorithm with closed form updates

5. We note that the result in fact holds under a weaker sparsity assumption by which the expected rewards bounded in squared L_2 norm.

(in particular, it doesn't involve any min-max optimization procedure) as well as being able to output a policy from Π.[6] The high-level approach has a similar structure to that of Erez et al. (2024b); Erez and Koren (2025); namely, the algorithm operates in two phases:

(i) Compute an *exploration distribution* $\hat{p} \in \Delta(\Pi)$ with the property that the importance-weighted reward estimator induced by $\hat{p}$ has low variance.

(ii) Use the reward estimator to uniformly estimate $R_D(\pi)$ for all $\pi \in \Pi$ and output $\hat{\pi}$ with the highest empirical reward.

Computing an exploration distribution. We compute $\hat{p}$ by running Hedge / Multiplicative Weights (MW) over $\Delta(\Pi)$ for $T \approx |\mathcal{A}|/\varepsilon$ iterations with reward vectors defined in Eq. (8) to obtain a sequence of policies $\pi_1, \ldots, \pi_T \in \Pi$, and construct the exploration distribution $\hat{p}$ as the uniform mixture of this sequence, namely $\hat{p}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\pi_t = \pi\}$ for all $\pi \in \Pi$. The policies $\pi_1, \ldots, \pi_T$ can be thought of as a sufficient policy coverage of the true reward function, in the sense that the reward estimator induced by $\hat{p}$ has variance at most $V = \widetilde{O}(s + \varepsilon|\mathcal{A}|)$.

Uniform reward estimation. Using the reward estimators induced by $\hat{p}$ we uniformly estimate the expected reward of all policies in Π, and invoke the variance sensitive concentration bound of Bernstein's inequality to argue that sufficient number of samples is $\widetilde{O}((V/\varepsilon^2 + |\mathcal{A}|/\varepsilon)\log(|\Pi|/\delta)) = \widetilde{O}((s/\varepsilon^2 + |\mathcal{A}|/\varepsilon)\log(|\Pi|/\delta))$, agreeing with our desired sample complexity bound.

Algorithm and analysis. Our algorithm is detailed in Algorithm 2. It makes use of both a weighted ERM oracle to Π (special case of the argmax oracle of Dudik et al. (2011); Agarwal et al. (2014)) and a sampling procedure from distributions over Π obtained from multiplicative weight updates. When Π is finite, both can be implemented in computational complexity linear in $|\Pi|$ and $|\mathcal{A}|$.

Our main result for Algorithm 2 is given in the following theorem:

Theorem 3 *Let $\Pi : \mathcal{X} \to \mathcal{A}$ be a finite policy class. Assume that $\mathbb{E}_{(x,r)\sim D}\left[\|r\|_2^2\right] \le s$.[7] If we set $T = \widetilde{\Theta}((|\mathcal{A}|/\varepsilon)\log(|\Pi|/\delta))$, $n = \widetilde{\Theta}\big((|\mathcal{A}|/\varepsilon + s/\varepsilon^2)\log(|\Pi|/\delta)\big)$, $\eta = \gamma/|\mathcal{A}|$ and $\gamma = 1/2$, then with probability at least $1-\delta$, Algorithm 2 outputs $\hat{\pi} \in \Pi$ for which $\max_{\pi^\star\in\Pi} R_D(\pi^\star) - R_D(\hat{\pi}) \le \varepsilon$, using a total sample complexity of*

$$\widetilde{O}\left(\left(\frac{s}{\varepsilon^2} + \frac{|\mathcal{A}|}{\varepsilon}\right)\log\frac{|\Pi|}{\delta}\right).$$

As a direct corollary, in the single-label classification setting, using Proposition 1 of Erez et al. (2024b) we can obtain the following sample complexity upper bound for classes with finite Natarajan dimension (and a finite label space):

Corollary 4 *Let $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ be a hypothesis class of finite Natarajan dimension d_N and $|\mathcal{Y}| < \infty$. Then, there exists a bandit multiclass classification algorithm which with probability at least $1-\delta$ outputs $\hat{h}$ with $R_D(h^\star) - R_D(\hat{h}) \le \varepsilon$ using a total sample complexity of*

$$\widetilde{O}\left(\left(\frac{1}{\varepsilon^2} + \frac{|\mathcal{Y}|}{\varepsilon}\right)\left(d_N + \log\frac{1}{\delta}\right)\right).$$

6. This property is referred to as *proper learning* in the learning theory literature.
7. We note that this is a weaker assumption than sparsity with respect to the L_1 norm.

Algorithm 2 Low Variance Exploration with Hedge

parameters: $T \in \mathbb{N}, n \in \mathbb{N}, \eta > 0, \gamma \in (0, 1]$.
initialize: $p_1 = \mathrm{Unif}(\Pi)$.
for $t = 1, 2, \ldots, T$ **do** $\{\leftarrow$ Phase I$\}$
Observe $x_t \in \mathcal{X}$, predict a_t uniformly at random from $\mathcal{A}$ and receive feedback $r_t(a_t)$.
Sample $\pi_t \sim p_t$ and update $p_{t+1}(\pi) \propto p_t(\pi) \cdot e^{\eta u_t(\pi)}$ for all $\pi \in \Pi$, where

$$u_t(\pi) = \frac{r_t(a_t)^2 \mathbf{1}\{\pi(x_t) = a_t\}}{\gamma/|\mathcal{A}| + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1} \mathbf{1}\{\pi_s(x_t) = a_t\}}, \qquad \forall\, \pi \in \Pi. \tag{8}$$

end for
Fix exploration distribution: $\hat{p}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\pi_t = \pi\}$.
for $i = 1, \ldots, n$ **do** $\{\leftarrow$ Phase II$\}$
Observe $x_i \in \mathcal{X}$; with prob. γ sample a_i uniformly at random from $\mathcal{A}$, otherwise sample $\pi_i \sim \hat{p}$ and set $a_i = \pi_i(x_i)$; predict a_i and receive feedback $r_i(a_i)$.
end for
return:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \left\{ \sum_{i=1}^{n} \frac{r_i(a_i)\mathbf{1}\{\pi(x_i) = a_i\}}{\gamma/|\mathcal{A}| + (1-\gamma)\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\pi_t(x_i) = a_i\}} \right\}.$$

Throughout this section, given $p \in \Delta(\Pi)$, $x \in \mathcal{X}$ and $a \in \mathcal{A}$ we denote the probability of predicting the action a for x when sampling from p by $Q_{x,a}(p) \coloneqq \sum_{\pi} p(\pi)\mathbf{1}\{\pi(x) = a\}$. The key property of Algorithm 2 is given in the following theorem which concerns the first phase.

Theorem 5 *For any $0 < \gamma \le \frac{1}{2}$ and $T \ge |\mathcal{A}|/\gamma$, using $\eta = \gamma/|\mathcal{A}|$, with probability at least $1 - \delta$, the first phase results in $\hat{p} \in \Delta(\Pi)$ satisfying*

$$\mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} r(a)^2 \frac{\mathbf{1}\{\pi(x) = a\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})}\right] \le \widetilde{O}\left(s\log T + \frac{|\mathcal{A}|^2}{\gamma T}\log\frac{|\Pi|}{\delta}\right) \quad \forall \pi \in \Pi.$$

We now provide a proof sketch for Theorem 5. For a full proof, see Appendix C.

Proof (sketch). We show that for a fixed $\pi \in \Pi$ the bound holds in expectation (over the randomness in the environment and the algorithm). The high-probability version involves Freedman-style concentration inequalities and is deferred to the full proof. A multiplicative regret bound of Hedge for (nonnegative) rewards bounded by $|\mathcal{A}|/\gamma$ implies that

$$\mathbb{E}\left[\sum_{t=1}^{T} u_t(\pi)\right] \le 2 \cdot \mathbb{E}\left[\sum_{t=1}^{T} u_t^\top p_t\right] + \frac{|\mathcal{A}|\log|\Pi|}{\gamma},$$

so it suffices to upper bound the expected total reward of Hedge by $\widetilde{O}(s\log T)$ and to lower bound the expected total reward of π. Using the fact that the rewards are monotonically decreasing in

expectation, namely $\mathbb{E}_t[u_t(\pi)] \geq \mathbb{E}_t[u_{t+1}(\pi)]$, it can be shown that

$$\mathbb{E}\left[\sum_{t=1}^{T} u_t(\pi)\right] \geq \frac{T}{|\mathcal{A}|} \cdot \mathbb{E}_{(x,r)\sim D,\hat{p}}\left[\sum_{a\in\mathcal{A}} r(a)^2 \frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}| + (1-\gamma)\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{\pi_t(x)=a\}}\right]$$
$$= \frac{T}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim D,\hat{p}}\left[\sum_{a\in\mathcal{A}} r(a)^2 \frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})}\right],$$

Turning to the cumulative reward of Hedge, we make use of reward sparsity and a harmonic sum inequality to obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} u_t^\top p_t\right] \leq \frac{T}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} r(a)^2 \sum_{t=1}^{T}\mathbb{E}_t\left[\frac{\mathbf{1}\{\pi_t(x)=a\}}{1+(1-\gamma)\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x)=a\}}\right]\right]$$
$$\lesssim \frac{T}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim D,\hat{p}}\left[\sum_{a\in\mathcal{A}} r(a)^2 \log\left(1+\sum_{t=1}^{T}\mathbf{1}\{\pi_t(x)=a\}\right)\right] \lesssim \frac{sT}{|\mathcal{A}|}\log T,$$

which implies the result after rearranging. ■

Given the result of Theorem 5, the proof of Theorem 3 follows from a straightforward application of Bernstein's inequality given the small variance of the resulting reward estimator. The proof is deferred to Appendix C.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreements No. 882396 and 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 993/17, 3174/23 and 2250/22), a grant from the Tel Aviv University Center for AI and Data Science (TAD) and from the Len Blavatnik and the Blavatnik Family foundation.

Shay Moran is a Robert J. Shillman Fellow; he acknowledges support by ISF grant 1225/20, by BSF grant 2018385, by Israel PBC-VATAT, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Fan Chen and Alexander Rakhlin acknowledge support from AFOSR through award FA9550-25-1-0375, Simons Foundation and the NSF through awards DMS-2031883 and PHY-2019786, and DARPA AIQ award.

References

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

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer and Philip M Long. Structural results about on-line learning models with and without queries. *Machine Learning*, 36:147–181, 1999.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1713–1726, 2023.

Fan Chen and Alexander Rakhlin. Decision making in changing environments: Robustness, query-based learning, and differential privacy. *arXiv preprint arXiv:2501.14928*, 2025.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for RL with decision-estimation coefficients: PAC, reward-free, preference-based learning, and beyond. *arXiv preprint arXiv:2209.11745*, 2022.

Fan Chen, Dylan J Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, Fano, and Le Cam with interaction: A unifying lower bound framework and characterization for bandit learnability. *arXiv preprint arXiv:2410.05117*, 2024.

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

Alon Cohen, Liad Erez, Steve Hanneke, Tomer Koren, Yishay Mansour, Shay Moran, and Qian Zhang. Sample complexity of agnostic multiclass classification: Natarajan dimension strikes back. *arXiv preprint arXiv:2511.12659*, 2025.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, 2013.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232, 2011.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

Liad Erez and Tomer Koren. Bandit multiclass list classification. *arXiv preprint arXiv:2502.09257*, 2025.

Liad Erez, Alon Cohen, Tomer Koren, Yishay Mansour, and Shay Moran. The real price of bandit information in multiclass classification. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1573–1598. PMLR, 2024a.

Liad Erez, Alon Peled-Cohen, Tomer Koren, Yishay Mansour, and Shay Moran. Fast rates for bandit pac multiclass classification. *Advances in Neural Information Processing Systems*, 37: 75152–75176, 2024b.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Yuval Filmus, Steve Hanneke, Idan Mehalel, and Shay Moran. Bandit-feedback online multiclass classification: Variants and tradeoffs. *arXiv preprint arXiv:2402.07453*, 2024.

Dean Foster, Dylan J Foster, Noah Golowich, and Alexander Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2678–2792. PMLR, 2023a.

Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*, pages 3199–3210, 2020.

Dylan J Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E Schapire. Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088*, 2018.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35:35404–35417, 2022.

Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3969–4043. PMLR, 2023b.

Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. *Advances in Neural Information Processing Systems*, 36:20080–20117, 2023c.

Margalit Glasgow and Alexander Rakhlin. Tight bounds for γ-regret via the decision-estimation coefficient. *arXiv preprint arXiv:2303.03327*, 2023.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10), 2007.

Steve Hanneke, Shay Moran, and Waknine Tom. List sample compression and uniform convergence. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2360–2388. PMLR, 2024.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. *Advances in Neural Information Processing Systems*, 34:2654–2667, 2021.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM, 2008.

Satyen Kale, Lev Reyzin, and Robert E Schapire. Non-stochastic bandit slate problems. *Advances in Neural Information Processing Systems*, 23, 2010.

Akshay Krishnamurthy, Alekh Agarwal, and Miro Dudik. Contextual semibandits via supervised learning oracles. *Advances In Neural Information Processing Systems*, 29, 2016.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems*, 20:817–824, 2007.

Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020.

Tor Lattimore, Branislav Kveton, Shuai Li, and Csaba Szepesvari. Toprank: A practical algorithm for online stochastic ranking. *Advances in Neural Information Processing Systems*, 31, 2018.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Decision making in hybrid environments: A model aggregation approach. *arXiv preprint arXiv:2502.05974*, 2025a.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. An improved model-free decision-estimation coefficient with applications in adversarial mdps. *arXiv preprint arXiv:2510.08882*, 2025b.

Philip M. Long. New bounds on the price of bandit feedback for mistake-bounded online multiclass learning. *Theor. Comput. Sci.*, 808:159–163, 2020. doi: 10.1016/J.TCS.2019.11.017.

H. B McMahan and M. J Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

Shay Moran, Ohad Sharon, Iska Tsubari, and Sivan Yosebashvili. List online classification. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1885–1913. PMLR, 2023.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375. PMLR, 2015.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

Ananth Raman, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learnability under bandit feedback. *arXiv preprint arXiv:2308.04620*, 2023.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9791–9798, 2021.

Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR, 2015.

Lukas Zierahn, Dirk van der Hoeven, Nicolo Cesa-Bianchi, and Gergely Neu. Nonstochastic contextual combinatorial bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 8771–8813. PMLR, 2023.

Appendix A. Extension to combinatorial semi-bandits

Our approach outlined in Section 4 has the additional benefit of extending to the more general setting of *contextual combinatorial semi-bandits* (CCSB, Erez and Koren (2025)), in which the predictions are subsets (or lists) of actions of a fixed size, and the reward of a given subset is the sum of rewards of individual actions in this subset.

Problem setup. Let $\mathcal{A} = \{a \in \{0,1\}^K \mid \|a\|_1 = m\}$ be the prediction space where $1 \le m \le K$ is an integer. We use the notation $j \in a$ for $j \in [K]$ to mean $a_j = 1$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times [0,1]^K$. In this variant, the learner interacts with the environment according to the following protocol. For $t = 1, 2, \ldots, T$:

- The environment draws $(x_t, r_t) \sim \mathcal{D}$ and x_t is revealed to the learner.
- The learner selects $a_t \in \mathcal{A}$ and observes $(r_t(j))_{j \in a_t}$, namely, *semi-bandit feedback*.

Learning objective. Given a policy class $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$, the objective is to produce a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ which satisfies $L_{\mathcal{D}}(\pi) := \max_{\pi^\star \in \Pi} R_{\mathcal{D}}(\pi^\star) - R_{\mathcal{D}}(\pi) < \varepsilon$, where the population reward of a policy π is defined as $R_{\mathcal{D}}(\pi) := \mathbb{E}_{(x,r)\sim\mathcal{D}, a\sim\pi(x)}[a^\top r]$, namely, the reward associated with a subset $a \in \mathcal{A}$ is the sum of individual rewards of actions in a[8]. The special case of binary rewards corresponds to *bandit multiclass list classification*.

Sparse rewards. As in the vanilla contextual bandit setting, we assume the rewards $r \in [0,1]^K$ satisfy $\mathbb{P}_{(x,r)\sim\mathcal{D}}[\|r\|_1 \le s] = 1$.

Main result. In an analogy with our approach described in Section 4, we design an algorithm for contextual combinatorial semi-bandits which admits closed-form updates, outlined in Algorithm 3. The sample complexity guarantee for Algorithm 3 is given in the following theorem.

Theorem 6 *Let $\Pi : \mathcal{X} \to \mathcal{A}$ be a finite policy class, and assume the rewards $r \in [0,1]^K$ satisfy $\|r\|_1 \le s$. If we set*

$$T = \widetilde{\Theta}\left(\frac{K \min(s,m)}{m\varepsilon} \log \frac{|\Pi|}{\delta}\right), \quad n = \widetilde{\Theta}\left(\left(\frac{K \min(s,m)}{m\varepsilon} + \frac{s \min(s,m)}{\varepsilon^2}\right) \log \frac{|\Pi|}{\delta}\right),$$

$\eta = \gamma m/(K \min(s,m))$ and $\gamma = 1/2$, then with probability at least $1-\delta$ Algorithm 3 outputs $\hat{\pi} \in \Pi$ with $\max_{\pi^\star \in \Pi} R_{\mathcal{D}}(\pi^\star) - R_{\mathcal{D}}(\hat{\pi}) \le \varepsilon$ using a total sample complexity of

$$\widetilde{O}\left(\left(\frac{K \min(s,m)}{m\varepsilon} + \frac{s \min(s,m)}{\varepsilon^2}\right) \log \frac{|\Pi|}{\delta}\right).$$[9]

Comparison with Erez and Koren (2025). Erez and Koren (2025) obtained the following sample complexity bound for the same problem:

$$O\left(\frac{K^9}{m^8} + \frac{sm}{\varepsilon^2} \log \frac{|\Pi|}{\delta}\right),$$

over which the bound given in Theorem 6 can be shown to be a strict improvement up to logarithmic factors. We remark that the reduction of the m factor in the $1/\varepsilon^2$ term to $\min(s,m)$ can be shown to hold for of the algorithm by Erez and Koren (2025) with a slightly tighter analysis.

8. Note that contextual bandits is the special case where $m = 1$.
9. The sparsity assumption can be weakened slightly to rewards bounded in squared L_2 norm obtain a rate of $\widetilde{O}((K/\varepsilon + sm/\varepsilon^2)\log(|\Pi|/\delta))$.

Additional notation. Given $p \in \Delta(\Pi)$, $x \in \mathcal{X}$ and $j \in [K]$ we denote

$$Q_{x,j}(p) \coloneqq \sum_{\pi \in \Pi} p(\pi) \mathbf{1}\{j \in \pi(x)\}.$$

Algorithm 3 Low Variance Exploration for CCSB

Parameters: $T \in \mathbb{N}, n \in \mathbb{N}, \eta > 0, \gamma > 0$.
Phase 1:
Initialize $p_1 \in \Delta(\Pi)$ as the uniform distribution.
for $t = 1, 2, \ldots, T$ **do**
Environment generates $(x_t, r_t) \sim \mathcal{D}$, algorithm receives x_t.
Predict $a_t \sim \mathcal{A}$ uniformly at random and receive feedback $(r_t(j))_{j \in a_t}$.
Sample $\pi_t \sim p_t$.
Define the reward $u_t(\cdot) : \Pi \to \mathbb{R}$ by

$$u_t(\pi) = \sum_{j=1}^{K} \frac{r_t(j)\mathbf{1}\{j \in a_t \cap \pi(x_t)\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1} \mathbf{1}\{j \in \pi_s(x_t)\}}, \quad \forall \pi \in \Pi.$$

Update

$$p_{t+1}(\pi) \propto p_t(\pi) \cdot e^{\eta u_t(\pi)}, \quad \forall \pi \in \Pi.$$

end for
Define $\widehat{p}(\pi) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{\pi_t = \pi\}$.
Phase 2:
for $i = 1, \ldots, n$ **do**
Environment generates $(x_i, r_i) \sim \mathcal{D}$, algorithm receives x_i.
With prob. γ sample $a_i \sim \mathcal{A}$ uniformly at random, otherwise sample $\pi_i \sim \hat{p}$ and set $a_i = \pi_i(x_i)$.

Predict a_i and receive feedback $(r_i(j))_{j \in a_i}$.
end for
Return:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} \alpha_{i,j} r_i(j) \mathbf{1}\{j \in \pi(x_i)\} \right\},$$

where $\alpha_{i,j} = \frac{1}{\gamma m/K + (1-\gamma) Q_{x_i,j}(\hat{p})}$ for all $i \in [n], j \in [K]$.

The analysis is analogous to the vanilla combinatorial bandit setting, with a more subtle argument allowing us to provide a tighter upper bound on the variance of the importance weighted estimators given the L_1 sparsity assumption. First, we establish to following guarantee of the first phase of Algorithm 3, which control the variance of the importance weighted reward estimators induced by $\hat{p}$.

Theorem 7 *For any $0 < \gamma \leq \frac{1}{2}$ and $T \geq K/m\gamma$, using $\eta = \gamma m/(K \min(s, m))$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^T$ and the internal randomness of Algorithm 3, the first phase results in a*

distribution $\hat{p} \in \Delta(\Pi)$ *which satisfies*

$$\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{j=1}^{K} r(j)\frac{\mathbf{1}\{j \in \pi(x)\}}{\gamma m/K + (1-\gamma)Q_{x,j}(\hat{p})}\right] \leq \widetilde{O}\left(s\log T + \frac{K^2\min(s,m)}{\gamma m^2 T}\log\frac{|\Pi|}{\delta}\right) \quad \forall \pi \in \Pi.$$

Proof Denote the bound on the Hedge reward functions as $B = (K\min(s,m))/(\gamma m)$. Define the filtration $\mathcal{F}_0 = \emptyset$ and $\mathcal{F}_t$ the σ-algebra generated by $\{(\pi_s, x_s, r_s, a_s)\}_{s\leq t}$ for all $t \in [T]$. By Theorem 14, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} u_t(\pi) \geq \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[u_t(\pi) \mid \mathcal{F}_{t-1}] - B\log\frac{1}{\delta}.$$

Now,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[u_t(\pi) \mid \mathcal{F}_{t-1}] &= \sum_{t=1}^{T}\mathbb{E}_{(x_t,r_t)\sim\mathcal{D},a_t\sim\mathcal{A}}\left[\sum_{j=1}^{K}\frac{r_t(j)\mathbf{1}\{j\in\pi(x_t)\cap a_t\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{j\in\pi_s(x_t)\}} \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{(x,r)\sim\mathcal{D},a\sim\mathcal{A}}\left[\sum_{j=1}^{K}\frac{r(j)\mathbf{1}\{j\in\pi(x)\cap a\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{j\in\pi_s(x)\}} \mid \mathcal{F}_{T}\right] \\
&= \frac{m}{K}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{j=1}^{K} r(j)\sum_{t=1}^{T}\frac{\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{j\in\pi_s(x)\}} \mid \mathcal{F}_{T}\right] \\
&\geq \frac{mT}{K}\cdot\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{j=1}^{K} r(j)\frac{\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{j\in\pi_t(x)\}} \mid \mathcal{F}_{T}\right] \\
&= \frac{mT}{K}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{j=1}^{K} r(j)\frac{\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K + (1-\gamma)Q_{x,j}(\hat{p})}\right],
\end{aligned}$$

so that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} u_t(\pi) \geq \frac{mT}{2K}\cdot\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{j=1}^{K} r(j)\frac{\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K + (1-\gamma)Q_{x,j}(\hat{p})}\right] - 2B\log\frac{1}{\delta}.$$

Now, By Theorem 14 with $\mathcal{F}_t$ defined as the σ-algebra generated by $\{(\pi_s, x_s, r_s, a_s)\}_{s\leq t}\cup\{(x_{t+1}, r_{t+1}, a_{t+1})\}$, with probability at least $1-\delta$ it holds that

$$\sum_{t=1}^{T} u_t(\pi_t) \geq \frac{1}{2}\sum_{t=1}^{T} u_t^\top p_t - 2B\log\frac{1}{\delta}.$$

Again by Theorem 14 with $\mathcal{F}_t$ defined as the σ-algebra generated by $\{(\pi_s, x_s, r_s, a_s)\}_{s\leq t}\cup\{\pi_{t+1}\}$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} u_t(\pi_t) \leq 2\sum_{t=1}^{T}\mathbb{E}[u_t(\pi_t) \mid \mathcal{F}_{t-1}] + B\log\frac{1}{\delta}.$$

We bound the right-hand side as

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[u_t(\pi_t) \mid \mathcal{F}_{t-1}] &= \sum_{t=1}^{T} \mathbb{E}_{(x_t,r_t)\sim D, a_t\sim\mathcal{A}}\left[\sum_{j=1}^{K} \frac{r_t(j)\mathbf{1}\{j \in \pi_t(x_t) \cap a_t\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{j \in \pi_s(x_t)\}} \mid \mathcal{F}_{t-1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{(x,r)\sim D, a\sim\mathcal{A}}\left[\sum_{j=1}^{K} \frac{r(j)\mathbf{1}\{j \in \pi_t(x) \cap a\}}{\gamma m/K + (1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{j \in \pi_s(x)\}} \mid \mathcal{F}_{T}\right] \\
&\le \frac{mT}{K}\mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K} r(j) \sum_{t=1}^{T} \frac{\mathbf{1}\{j \in \pi_t(x)\}}{1 + (1-\gamma)\sum_{s=1}^{t-1}\mathbf{1}\{j \in \pi_s(x)\}} \mid \mathcal{F}_{T}\right] \\
&\le \frac{2mT}{(1-\gamma)K}\mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K} r(j)\log\left(1 + \sum_{t=1}^{T}\mathbf{1}\{j \in \pi_t(x)\}\right) \mid \mathcal{F}_{T}\right] \\
&\le \frac{8smT}{K}\log T,
\end{aligned}$$

where in the second-to-last inequality we used Theorem 15. Combining this with the above implies that with probability $1-2\delta$,

$$\sum_{t=1}^{T} u_t^\top p_t \le \frac{32smT}{K}\log T + 6B\log\frac{1}{\delta}$$

Dividing through by mT/K and using Theorem 16 we obtain with probability $1-3\delta$,

$$\begin{aligned}
\mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K} r(j)\frac{\mathbf{1}\{j \in \pi(x)\}}{\gamma m/K + (1-\gamma)Q_{x,j}(\hat{p})}\right] &\le \left(\frac{2K}{mT} + 2\eta B\right)\sum_{t=1}^{T} u_t^\top p_t + \frac{2K}{\eta mT}\log|\Pi| + 4B\log\frac{1}{\delta} \\
&\le 3\left(32s\log T + \frac{6BK}{mT}\log\frac{1}{\delta}\right) + \frac{4BK}{mT}\log\frac{|\Pi|}{\delta} \\
&\le 96s\log T + \frac{18BK}{mT}\log\frac{1}{\delta} + \frac{4BK}{mT}\log\frac{|\Pi|}{\delta}.
\end{aligned}$$

Plugging in the value of B and using a union bound over $\pi \in \Pi$ we conclude the proof. ■

Proof [Proof of Theorem 6] For all $i \in [n]$ define the following importance-weighted reward estimator of a policy $\pi \in \Pi$:

$$R_i(\pi) = \sum_{j=1}^{K} \frac{r_i(j)\mathbf{1}\{j \in \pi(x_i) \cap a_i\}}{\gamma m/K + (1-\gamma)Q_{x_i,j}(\hat{p})}.$$

This is an unbiased estimator for $R_D(\pi)$:

$$\begin{aligned}
\mathbb{E}_{(x_i,r_i,a_i)}[R_i(\pi)] &= \mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K} \Pr[j \in a_i]\frac{r(j)\mathbf{1}\{j \in \pi(x)\}}{\gamma m/K + (1-\gamma)Q_{x,j}(\hat{p})}\right] \\
&= \mathbb{E}_{(x,r)\sim D}\left[r^\top \pi(x)\right] \\
&= R_D(\pi).
\end{aligned}$$

Now, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}_{(x_i,r_i,a_i)}\left[R_i(\pi)^2\right] &= \mathbb{E}_{(x_i,r_i,a_i)}\left[\left(\sum_{j=1}^{K}\frac{r_i(j)\mathbf{1}\{j\in\pi(x_i)\cap a_i\}}{\gamma m/K+(1-\gamma)Q_{x_i,j}(\hat{p})}\right)^2\right] \\
&\le \mathbb{E}_{(x_i,r_i,a_i)}\left[\left(\sum_{j=1}^{K} r_i(j)\mathbf{1}\{j\in\pi(x_i)\}\right)\cdot\left(\sum_{j=1}^{K}\frac{r_i(j)\mathbf{1}\{j\in\pi(x_i)\cap a_i\}}{\left(\gamma m/K+(1-\gamma)Q_{x_i,j}(\hat{p})\right)^2}\right)\right] \\
&\le \min(s,m)\cdot\mathbb{E}_{(x_i,r_i,a_i)}\left[\sum_{j=1}^{K}\frac{r_i(j)\mathbf{1}\{j\in\pi(x_i)\cap a_i\}}{\left(\gamma m/K+(1-\gamma)Q_{x_i,j}(\hat{p})\right)^2}\right] \\
&= \min(s,m)\cdot\mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K}\frac{r(j)\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K+(1-\gamma)Q_{x,j}(\hat{p})}\right],
\end{aligned}$$

where the second inequality follows from the fact that the first sum is bounded both by m (by bounding each $r_i(j)$ by 1 and noting that $\pi(x_i)$ has at most m nonzero entries) and by s (by bounding the indicator terms by 1 and using the sparsity assumption), so it is bounded by their minimum. Next, by Theorem 7 and our choices for γ and T,

$$\begin{aligned}
\mathbb{E}_{(x,r)\sim D}\left[\sum_{j=1}^{K}\frac{r(j)\mathbf{1}\{j\in\pi(x)\}}{\gamma m/K+(1-\gamma)Q_{x,j}(\hat{p})}\right] &\le \widetilde{O}\left(s\cdot\log T+\frac{K^2\min(s,m)}{\gamma m^2 T}\log\frac{|\Pi|}{\delta}\right) \\
&\le \widetilde{O}\left(s+\frac{\varepsilon K}{m}\right),
\end{aligned}$$

Since the reward estimators are also uniformly bounded by $K\min(s,m)/m$, using Bernstein's inequality and a union bound over Π, if $n=\widetilde{\Theta}\big((K\cdot\min(s,m)/(m\varepsilon)+s\cdot\min(s,m)/\varepsilon^2)\log(|\Pi|/\delta)\big)$ then with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}R_i(\pi)-R_D(\pi)\right|\le\varepsilon\quad\forall\pi\in\Pi,$$

which in turn implies that for all $\pi^\star\in\Pi$, with probability at least $1-\delta$,

$$\begin{aligned}
L(\hat{\pi}) &= R_D(\pi^\star)-\frac{1}{n}\sum_{i=1}^{n}R_i(\pi^\star)+\frac{1}{n}\sum_{i=1}^{n}R_i(\pi^\star)-R_D(\hat{\pi}) \\
&\le \varepsilon+\frac{1}{n}\sum_{i=1}^{n}R_i(\hat{\pi})-R_D(\hat{\pi}) \\
&\le 2\varepsilon.
\end{aligned}$$

■

Appendix B. Proofs for Section 3

B.1. Proof of Theorem 1

The following lemma is established by Foster et al. (2022) with minimax analysis.

Lemma 8 *Suppose that $\mathcal{M} \subseteq (\mathcal{A} \to \Delta(\mathcal{O}))$ is convex, and $\mathcal{O}$ is finite. For any $w \in \Delta(\mathcal{A})$, it holds that*

$$\min_{p,q\in\Delta(\mathcal{A}),\xi\in\Xi} \Gamma_{w,\gamma}(p,q,\xi) \le \mathsf{p\text{-}dec}^{\mathrm{o}}_{\gamma/8}(\mathcal{M}). \tag{9}$$

The following lemma guarantees the performance of exponential weight updates.

Lemma 9 *Denote*

$$\mathrm{Err}(q,\xi;w,M,a^\star) := \mathbb{E}_{a\sim q,r\sim M(a)}\mathbb{E}_{a'\sim w}\left[1-\exp\left(\xi(a';a,o)-\xi(a^\star;a,o)\right)\right].$$

Then with probability at least $1-\delta$, it holds that for any $\pi^\star \in \Pi$,

$$\sum_{t=1}^{T}\mathrm{Err}(q_t,\xi_t;w_t,M^\star_{x_t},\pi^\star(x_t)) \le 2\log(|\Pi|/\delta).$$

Proof For simplicity, we denote $z_t = (x_t,a_t,o_t)$ and write $\xi_t(\pi;z_t) := \xi_t(\pi(x_t);a_t,o_t)$. By definition,

$$W_t(\pi) = \frac{W_1(\pi)\exp\left(\sum_{s=1}^{t}\xi_s(\pi;z_s)\right)}{\sum_{\pi'\in\Pi}W_1(\pi')\exp\left(\sum_{s=1}^{t-1}\xi_s(\pi';z_s)\right)},$$

and hence

$$\begin{aligned}\log\mathbb{E}_{\pi\sim W_t}\left[\exp\left(\xi_t(\pi;z_t)\right)\right] = {}& \log\mathbb{E}_{\pi\sim W_1}\exp\left(\sum_{s=1}^{t}\xi_s(\pi;z_s)\right)\\ & -\log\mathbb{E}_{\pi\sim W_1}\exp\left(\sum_{s=1}^{t-1}\xi_s(\pi;z_s)\right).\end{aligned}$$

Therefore, taking summation over $t=1,\cdots,T$, we have

$$-\sum_{t=1}^{T}\log\mathbb{E}_{\pi\sim W_t}\left[\exp\left(\xi_t(\pi;z_t)\right)\right] = -\log\mathbb{E}_{\pi\sim W_1}\left[\exp\left(\sum_{t=1}^{T}\xi_t(\pi;z_t)\right)\right]. \tag{10}$$

Thus, we define

$$A_t(\pi;z_t) := -\xi_t(\pi;z_t)+\log\mathbb{E}_{\pi'\sim W_t}\left[\exp\left(\xi_t(\pi';z_t)\right)\right],$$

and (10) implies that deterministically,

$$\mathbb{E}_{\pi\sim W_1}\exp\left(-\sum_{t=1}^{T}A_t(\pi;z_t)\right)=1,$$

i.e.,

$$-\sum_{t=1}^{T}A_t(\pi;z_t)\le\log|\Pi|,\qquad\forall\pi\in\Pi.$$

Notice that for any $\pi \in \Pi$, we also have

$$\mathbb{E}^{\mathsf{ExO}^+} \exp\left(\sum_{t=1}^{T} A_t(\pi; z_t) - \log \mathbb{E}_{t-1,x_t}\left[\exp\left(A_t(\pi; z_t)\right)\right]\right) = 1,$$

where the expectation $\mathbb{E}^{\mathsf{ExO}^+}$ is taken over the randomness of the interaction between ExO^+ algorithm and the environment, and $\mathbb{E}_{t-1,x_t}[\cdot]$ is the conditional expectation with respect to the history $\mathcal{H}_{t-1} = (x_s, a_s, o_s)_{s<t}$ and x_t.

Further, by the definition of A_t and Err, it holds that for any fixed $\pi \in \Pi$,

$$\begin{aligned}
\mathbb{E}_{t-1,x_t}\left[\exp\left(A_t(\pi; z_t)\right)\right] &= \mathbb{E}_{a_t \sim q_t, o_t \sim M^\star_{x_t}(a_t)} \mathbb{E}_{\pi' \sim W_t} \exp(\xi_t(\pi'(x_t); a_t, o_t) - \xi_t(\pi(x_t); a_t, o_t)) \\
&= \mathbb{E}_{a \sim q_t, r \sim M^\star_{x_t}(a)} \mathbb{E}_{a' \sim w_t} \exp(\xi_t(a'; a, o) - \xi_t(\pi(x_t); a, o)) \\
&= 1 - \mathrm{Err}(q_t, \xi_t; w_t, M^\star_{x_t}, \pi(x_t)),
\end{aligned}$$

where we use $w_t = W_t|_{x_t}$ is the marginal defined in (3).

Hence, by Markov's inequality and union bound, we can conclude that with probability at least $1 - \delta$, for any $\pi \in \Pi$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{Err}(q_t, \xi_t; w_t, M^\star_{x_t}, \pi(x_t)) &\le \sum_{t=1}^{T} -\log \mathbb{E}_{t-1}\left[\exp\left(A_t(\pi; z_t)\right)\right] \\
&\le -\sum_{t=1}^{T} A_t(\pi; z_t) + \log(|\Pi|/\delta) \le 2\log(|\Pi|/\delta).
\end{aligned}$$

■

Theorem 10 *For any $p \in \Delta(\mathcal{A})$, we denote*

$$f^\star(x, p) := \mathbb{E}_{a \sim p, o \sim M^\star_x(a)}[R(o)]. \tag{11}$$

Then with probability at least $1 - \delta$, it holds that

$$\sum_{t=1}^{T} \left(f^\star(x_t, \pi^\star(x_t)) - f^\star(x_t, p_t)\right) \le T \cdot \mathsf{p\text{-}dec}^{\mathrm{o}}_{\gamma/8}(\mathcal{M}) + 2\gamma \log(|\Pi|/\delta). \tag{12}$$

Proof Fix $\pi^\star \in \Pi$, we can organize

$$
\begin{aligned}
&\sum_{t=1}^{T}\big(f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,p_t)\big)\\
&= \sum_{t=1}^{T}\underbrace{\big(f^\star(x_t,\pi^\star(x_t)) - f^\star(x_t,p_t) - \gamma\mathrm{Err}(q_t,\xi_t;w_t,M^\star_{x_t},\pi^\star(x_t))\big)}_{=\Gamma_{w_t,\gamma}(p_t,q_t,\xi_t;M^\star_{x_t},\pi^\star(x_t))} + \gamma\underbrace{\sum_{t=1}^{T}\mathrm{Err}(q_t,\xi_t;w_t,M^\star_{x_t},\pi^\star(x_t))}_{\le 2\log(|\Pi|/\delta)}\\
&\le \sum_{t=1}^{T}\Gamma_{w_t,\gamma}(p_t,q_t,\xi_t) + 2\gamma\log(|\Pi|/\delta)\\
&= \sum_{t=1}^{T}\min_{p,q,\xi}\Gamma_{w_t,\gamma}(p,q,\xi) + 2\gamma\log(|\Pi|/\delta)\\
&\le T\cdot \textsf{p-dec}^{\mathrm{o}}_{\gamma/8}(\mathcal{M}) + 2\gamma\log(|\Pi|/\delta),
\end{aligned}
$$

where the second line uses the definition of $\Gamma_{w,\gamma}(p,q,\xi;M,a^\star)$ in (4) and Lemma 9, the third line uses $\Gamma_{w,\gamma}(p,q,\xi;M,a^\star) \le \Gamma_{w,\gamma}(p,q,\xi)$, the forth line uses the optimality of (p_t,q_t,ξ_t) for each $t\in[T]$, and the last line uses Lemma 8.

Taking maximum over $\pi^\star\in\Pi$ completes the proof. ■

Note that $p_t = P_t(x_t)$ deterministically. Further, we can calculate the conditional distribution:

$$\mathbb{E}_{t-1}[f^\star(x_t,p_t)] = \mathbb{E}_{x\sim\rho}[f^\star(x,P_t(x))],$$

where $\mathbb{E}_{t-1}[\cdot]$ is the conditional expectation with respect to the history before the tth step: $\mathcal{H}_{t-1} = (x_s,a_s,o_s)_{s<t}$. Then, appling the martingale concentration inequality gives with probability at least $1-\delta$,

$$\sum_{t=1}^{T} f^\star(x_t,p_t) \le \sum_{t=1}^{T}\mathbb{E}_{x\sim\rho}[f^\star(x,P_t(x))] + O(\sqrt{T\log(1/\delta)}).$$

This immediately results in the following corollary, which concludes the proof of Theorem 1.

Corollary 11 *with probability at least* $1-\delta$*, it holds that*

$$
\begin{aligned}
\max_{\pi^\star\in\Pi} R_D(\pi^\star) - R_D(\widehat{\pi}) &= \frac{1}{T}\max_{\pi^\star\in\Pi}\sum_{t=1}^{T}\mathbb{E}_{x\sim\rho}\big(f^\star(x,\pi^\star(x)) - f^\star(x,P_t(x))\big)\\
&\le \textsf{p-dec}^{\mathrm{o}}_{\gamma/8}(\mathcal{M}) + \frac{2\gamma\log(|\Pi|/\delta)}{T} + O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right).
\end{aligned}
$$

B.2. Proof of Theorem 2

In the following, we fixed any $\overline{M}\in\mathrm{co}(\mathcal{M}) = \mathcal{M}$ and bound $\textsf{p-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M},\overline{M})$.

For any $M \in \mathcal{M}$ and $a \in \mathcal{A}$, we recall $f^M(a) := \mathbb{E}_{o\sim M(a)}[R(o)]$. By Lemma 17, we can bound

$$\begin{aligned}|f^M(a) - f^{\overline{M}}(a)| &= |\mathbb{E}_{o\sim M(a)}[R(o)] - \mathbb{E}_{o\sim \overline{M}(a)}[R(o)]| \\ &\le 4\sqrt{\mathbb{E}_{o\sim \overline{M}(a)}[R(o)^2] \cdot D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right)} + 4D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right).\end{aligned}$$

We denote $\lambda_a := \mathbb{E}_{o\sim \overline{M}(a)}[R(o)^2]$ and $C := \sum_{a\in\mathcal{A}} \lambda_a \le s$. We consider $q \in \Delta(\mathcal{A})$ defined as

$$q(a) = \frac{1}{|\mathcal{A}|}\left(1 - \frac{C}{2s}\right) + \frac{\lambda_a}{2s}.$$

Then, we can bound

$$\begin{aligned}|f^M(a) - f^{\overline{M}}(a)| &\le 4\sqrt{\lambda_a \cdot D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right)} + 4D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right) \\ &\le 4\sqrt{2s\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right)} + 8|\mathcal{A}|\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right).\end{aligned}$$

In particular,

$$\begin{aligned}f^M(a_M) - f^M(a_{\overline{M}}) &\le 2\max_{a\in\mathcal{A}} |f^M(a) - f^{\overline{M}}(a)| \\ &\le 8\sqrt{2s\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right)} + 16|\mathcal{A}|\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right).\end{aligned}$$

Letting p to be supported on $a_{\overline{M}}$, we can now conclude that

$$\begin{aligned}\text{p-dec}^{\mathrm{o}}_\gamma(\mathcal{M}, \overline{M}) &\le \sup_{M\in\mathcal{M}} f^M(a_M) - f^M(a_{\overline{M}}) - \gamma\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right) \\ &\le \sup_{M\in\mathcal{M}} 8\sqrt{2s\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right)} - (\gamma - 16|\mathcal{A}|)\mathbb{E}_{a\sim q} D^2_{\mathrm{H}}\left(M(a), \overline{M}(a)\right) \\ &\le \frac{64s}{\gamma}, \qquad \forall \gamma \ge 32|\mathcal{A}|.\end{aligned}$$

Appendix C. Proofs for Section 4

We begin by proving the main guarantee required from the first phase of Algorithm 2 given in Theorem 5; namely, the exploration distribution $\hat{p}$ computed in the first phase induces a reward estimator with sufficiently low variance.

Proof [Proof of Theorem 5] Define the filtration $\mathcal{F}_0 = \emptyset$ and $\mathcal{F}_t$ the σ-algebra generated by $\{(\pi_s, x_s, r_s, a_s)\}_{s\le t}$ for all $t \in [T]$, and note that $u_t(\cdot) \in [0, |\mathcal{A}|/\gamma]^\Pi$ is $\mathcal{F}_t$-adapted. Fix $\pi \in \Pi$. By Theorem 14 with $b = 2$, $B = |\mathcal{A}|/\gamma$ and $X_t = u_t(\pi)$, with probability at least $1 - \delta$,

$$\sum_{t=1}^T u_t(\pi) \ge \frac{1}{2}\sum_{t=1}^T \mathbb{E}[u_t(\pi) \mid \mathcal{F}_{t-1}] - \frac{2|\mathcal{A}|}{\gamma}\log\frac{1}{\delta}.$$

Now,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[u_t(\pi)\mid\mathcal{F}_{t-1}] &= \sum_{t=1}^{T}\mathbb{E}_{(x_t,r_t)\sim\mathcal{D},a_t\sim\mathcal{A}}\left[\frac{r_t(a_t)^2\mathbf{1}\{\pi(x_t)=a_t\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x_t)=a_t\}}\mid\mathcal{F}_{t-1}\right]\\
&= \sum_{t=1}^{T}\mathbb{E}_{(x,r)\sim\mathcal{D},a\sim\mathcal{A}}\left[\frac{r(a)^2\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x)=a\}}\mid\mathcal{F}_{T}\right]\\
&= \frac{1}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\sum_{t=1}^{T}\frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x)=a\}}\mid\mathcal{F}_{T}\right]\\
&\geq \frac{T}{|\mathcal{A}|}\cdot\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{\pi_t(x)=a\}}\mid\mathcal{F}_{T}\right]\\
&= \frac{T}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)Q_{x,a}(\hat{p})}\right],
\end{aligned}$$

which together with the above implies that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}u_t(\pi)\geq\frac{T}{2|\mathcal{A}|}\cdot\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)Q_{x,a}(\hat{p})}\right]-\frac{2|\mathcal{A}|}{\gamma}\log\frac{1}{\delta}.$$

Now, By Theorem 14 with $b=2$, $B=|\mathcal{A}|/\gamma$ and $\mathcal{F}_t$ defined as the σ-algebra generated by $\{(\pi_s,x_s,r_s,a_s)\}_{s\leq t}\cup\{(x_{t+1},r_{t+1},a_{t+1})\}$, with probability at least $1-\delta$ it holds that

$$\sum_{t=1}^{T}u_t(\pi_t)\geq\frac{1}{2}\sum_{t=1}^{T}u_t^\top p_t-\frac{2|\mathcal{A}|}{\gamma}\log\frac{1}{\delta}.$$

Another use of Theorem 14 with $a=1$, this time with $\mathcal{F}_t$ defined as the σ-algebra generated by $\{(\pi_s,x_s,r_s,a_s)\}_{s\leq t}\cup\{\pi_{t+1}\}$, implies that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}u_t(\pi_t)\leq 2\sum_{t=1}^{T}\mathbb{E}[u_t(\pi_t)\mid\mathcal{F}_{t-1}]+\frac{|\mathcal{A}|}{\gamma}\log\frac{1}{\delta}.$$

Proceeding to upper bound the right-hand side,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[u_t(\pi_t)\mid\mathcal{F}_{t-1}] &= \sum_{t=1}^{T}\mathbb{E}_{(x_t,r_t)\sim\mathcal{D},a_t\sim\mathcal{A}}\left[\frac{r_t(a_t)^2\mathbf{1}\{\pi_t(x_t)=a_t\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x_t)=a_t\}}\mid\mathcal{F}_{t-1}\right]\\
&= \sum_{t=1}^{T}\mathbb{E}_{(x,r)\sim\mathcal{D},a\sim\mathcal{A}}\left[\frac{r(a)^2\mathbf{1}\{\pi_t(x)=a\}}{\gamma/|\mathcal{A}|+(1-\gamma)\frac{1}{T}\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x)=a\}}\mid\mathcal{F}_{T}\right]\\
&\leq \frac{T}{|\mathcal{A}|}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\sum_{t=1}^{T}\frac{\mathbf{1}\{\pi_t(x)=a\}}{1+(1-\gamma)\sum_{s=1}^{t-1}\mathbf{1}\{\pi_s(x)=a\}}\mid\mathcal{F}_{T}\right]\\
&\leq \frac{2T}{|\mathcal{A}|(1-\gamma)}\mathbb{E}_{(x,r)\sim\mathcal{D}}\left[\sum_{a\in\mathcal{A}}r(a)^2\log\left(1+\sum_{t=1}^{T}\mathbf{1}\{\pi_t(x)=a\}\right)\mid\mathcal{F}_{T}\right]\\
&\leq \frac{8sT}{|\mathcal{A}|}\log T,
\end{aligned}$$

where in the second-to-last inequality we used Theorem 15. Combining this with the above implies that with probability $1-2\delta$,

$$\sum_{t=1}^{T} u_t^\top p_t \le \frac{32sT}{|\mathcal{A}|}\log T + \frac{6|\mathcal{A}|}{\gamma}\log\frac{1}{\delta}$$

Dividing through by $T/|\mathcal{A}|$ and using Theorem 16 we obtain with probability $1-3\delta$,

$$\begin{aligned}
\mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} r(a)^2 \frac{\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})}\right] &\le \left(\frac{2|\mathcal{A}|}{T} + \frac{2\eta|\mathcal{A}|^2}{\gamma T}\right)\sum_{t=1}^{T} u_t^\top p_t + \frac{2|\mathcal{A}|}{\eta T}\log|\Pi| + \frac{4|\mathcal{A}|^2}{\gamma T}\log\frac{1}{\delta} \\
&\le 3\left(32s\log T + \frac{6|\mathcal{A}|^2}{\gamma T}\log\frac{1}{\delta}\right) + \frac{4|\mathcal{A}|^2}{\gamma T}\log\frac{|\Pi|}{\delta} \\
&\le 96s\log T + \frac{18|\mathcal{A}|^2}{\gamma T}\log\frac{1}{\delta} + \frac{4|\mathcal{A}|^2}{\gamma T}\log\frac{|\Pi|}{\delta},
\end{aligned}$$

and using a union bound over $\pi \in \Pi$ we conclude the proof. ■

We can now prove our main result of Theorem 3 by applying Bernstein's inequality on the reward estimators constructed in the second phase of the algorithm, with the variance bound obtained in Theorem 5.

Proof [Proof of Theorem 3] For all $i \in [n]$ define the following importance-weighted reward estimator of a policy $\pi \in \Pi$:

$$R_i(\pi) = \frac{r_i(a_i)\mathbf{1}\{\pi(x_i)=a_i\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x_i,a_i}(\hat{p})}.$$

This is an unbiased estimator for $R_D(\pi)$:

$$\begin{aligned}
\mathbb{E}_{(x_i,r_i,a_i)}[R_i(\pi)] &= \mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} \Pr[a_i=a]\frac{r(a)\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})}\right] \\
&= \mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} r(a)\mathbf{1}\{\pi(x)=a\}\right] \\
&= R_D(\pi).
\end{aligned}$$

By Theorem 5, we can bound the variance of this estimator by

$$\begin{aligned}
\mathrm{Var}[R_i(\pi)] &\le \mathbb{E}_{(x_i,r_i,a_i)}\left[R_i(\pi)^2\right] \\
&= \mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} \Pr[a_i=a]\frac{r(a)^2\mathbf{1}\{\pi(x)=a\}}{\left(\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})\right)^2}\right] \\
&= \mathbb{E}_{(x,r)\sim D}\left[\sum_{a\in\mathcal{A}} \frac{r(a)^2\mathbf{1}\{\pi(x)=a\}}{\gamma/|\mathcal{A}| + (1-\gamma)Q_{x,a}(\hat{p})}\right] \\
&\le \widetilde{O}\left(s\log T + \frac{|\mathcal{A}|^2}{\gamma T}\log\frac{|\Pi|}{\delta}\right) \\
&\le \widetilde{O}\left(s + \varepsilon|\mathcal{A}|\log\frac{|\Pi|}{\delta}\right),
\end{aligned}$$

where in the last inequality we plug in our choices for γ and T. Using Bernstein's inequality and a union bound over Π, if $n = \widetilde{\Theta}\big((|\mathcal{A}|/\varepsilon + s/\varepsilon^2)\log(|\Pi|/\delta)\big)$ then that with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} R_i(\pi) - R_D(\pi)\right| \le \varepsilon \quad \forall \pi \in \Pi,$$

which in turn implies that for all $\pi^\star \in \Pi$, with probability at least $1-\delta$,

$$\begin{aligned} L(\hat{\pi}) &= R_D(\pi^\star) - \frac{1}{n}\sum_{i=1}^{n} R_i(\pi^\star) + \frac{1}{n}\sum_{i=1}^{n} R_i(\pi^\star) - R_D(\hat{\pi}) \\ &\le \varepsilon + \frac{1}{n}\sum_{i=1}^{n} R_i(\hat{\pi}) - R_D(\hat{\pi}) \\ &\le 2\varepsilon. \end{aligned}$$

■

Appendix D. Lower bound

In this section we present a sample complexity lower bound for combinatorial bandits with s-sparse rewards, of the form

$$\widetilde{\Omega}\left(\frac{|\mathcal{A}|}{\varepsilon} + \frac{s}{\varepsilon^2}\right),$$

which implies the tightness of our upper bounds up to logarithmic factors. To our knowledge, the $|\mathcal{A}|/\varepsilon$ dependence has not been covered in previous works, so we include it for completeness. Our lower bound is formalized in the following theorem.

Theorem 12 *Let* Alg *be any contextual bandit algorithm over a finite action set $\mathcal{A}$ and let $1 \le s \le K$. Then there exists a context space $\mathcal{X}$ of size $|\mathcal{X}| = 2$, a policy class $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$ of size at most* $\mathrm{poly}(K)$ *and a distribution $\mathcal{D}$ over $\mathcal{X} \times [0,1]^{\mathcal{A}}$ with rewards satisfying $\|r\|_1 \le s$, such that in order to output a policy $\hat{\pi} : \mathcal{X} \to \Delta(\mathcal{A})$ satisfying $\max_{\pi^\star \in \Pi} R_D(\pi) - R_D(\hat{\pi}) \le \varepsilon$ with probability at least* $3/4$, Alg *requires a sample complexity of at least*

$$\widetilde{\Omega}\left(\frac{|\mathcal{A}|}{\varepsilon} + \frac{s}{\varepsilon^2}\right).$$

Proof We assume without loss the Alg is deterministic. By Yao's principle, the result will extend to general algorithms, as the hard instances do not depend on the choices made by Alg. The lower bound of $\Omega(s/\varepsilon^2)$ is an immediate consequence of Theorem 5 in Erez and Koren (2025) and holds even for a single context. We thus focus on constructing an instance on which Alg must use a sample complexity of at least $\Omega(|\mathcal{A}|/\varepsilon)$. Indeed, consider two contexts $\mathcal{X} = \{x_1, x_2\}$ where the probability of observing x_1 is ε and the probability of observing x_2 is $1-\varepsilon$. The policy class is defined as $\Pi = (\mathcal{X} \to \mathcal{A})$; note that $|\Pi| = K^2$. The reward function given the context x_2 is identically zero, while $r_{x_1} \in \{0,1\}^{\mathcal{A}}$ assigns a reward of 1 to a unique action $a^\star \in \mathcal{A}$ and 0 for $a \ne a^\star$, and is sampled uniformly at random prior to the interaction. Note that in order for Alg to output an ε-optimal

policy with probability at least $3/4$, it must produce $\hat{\pi} \in \Pi$ for which $\hat{\pi}(x_1) = a^\star$ with probability at least $3/4$. Now, denote by $a_1, a_2, \ldots$ the sequence of actions selected by Alg when the context was x_1. Since $a^\star$ is initially sampled uniformly at random, with probability at least $1/2$ it holds that $a_1, a_2, \ldots, a_{|\mathcal{A}|/2} \neq a^\star$, i.e. Alg has not received any signal regarding the identity of $a^\star$ in the first $|\mathcal{A}|/2$ rounds where the context was x_1. It thus suffices to prove that with probability at least $1/2$ the number of rounds it takes for x_1 to be sampled $|\mathcal{A}|/2$ times is $\Omega(|\mathcal{A}|/\varepsilon)$, which follows immediately from a standard Chernoff bound. ■

Appendix E. Technical lemmas

The following is a version of Freedman's inequality for martingales.

Lemma 13 (Theorem 1 in Beygelzimer et al. (2011)) *Let $X_1, \ldots, X_T \in [-B, B]$ be a martingale difference sequence with respect to the filtration $(\mathcal{F}_t)_{t=0}^T$ and let $S = \sum_{t=1}^T X_t$. $V = \sum_{t=1}^T \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right]$. Then for any $\delta > 0$ and $a \geq 1$,*

$$\Pr\left[S \leq \frac{V}{aB} + aB \log \frac{1}{\delta}\right] \geq 1 - \delta.$$

We use this result to prove the following concentration inequality:

Lemma 14 *Let $X_1, \ldots, X_T \in [0, B]$ be a martingale with respect to the filtration $(\mathcal{F}_t)_{t=0}^T$. Then for any $\delta > 0$ and $a, b \geq 1$,*

$$\Pr\left[\sum_{t=1}^T X_t \leq \left(1 + \frac{1}{a}\right) \sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] + aB \log \frac{1}{\delta}\right] \geq 1 - \delta,$$

and

$$\Pr\left[\sum_{t=1}^T X_t \geq \left(1 - \frac{1}{b}\right) \sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] - bB \log \frac{1}{\delta}\right] \geq 1 - \delta.$$

Proof Let $Y_t = X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$, $S = \sum_{t=1}^T Y_t$ and $V = \sum_{t=1}^T \mathbb{E}\left[Y_t^2 \mid \mathcal{F}_{t-1}\right]$. Using the fact that $X_t^2 \leq BX_t$, we obtain

$$\begin{aligned}
\Pr\left[\sum_{t=1}^T X_t \leq \left(1 + \frac{1}{a}\right) \sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] + aB \log \frac{1}{\delta}\right] &= \Pr\left[S \leq \frac{1}{aB} \sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] + aB \log \frac{1}{\delta}\right] \\
&\geq \Pr\left[S \leq \frac{1}{aB} \sum_{t=1}^T \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] + aB \log \frac{1}{\delta}\right] \\
&\geq \Pr\left[S \leq \frac{V}{aB} + aB \log \frac{1}{\delta}\right] \\
&\geq 1 - \delta,
\end{aligned}$$

where the last inequality follows from Theorem 13. The second inequality follows similarly by considering $Z_t = -Y_t$. ■

Lemma 15 *Let $a_1, \ldots, a_n \in [0,1]$. Then*

$$\sum_{i=1}^{n} \frac{a_i}{1 + \sum_{j=1}^{i-1} a_j} \leq 2\log\left(1 + \sum_{i=1}^{n} a_i\right).$$

Proof Denote $s_i = \sum_{j=1}^{i} a_j$ for $i \in [n]$ and $s_0 = 0$. Using the fact that $z \leq 2\log(1+z)$ for all $z \in [0,1]$, we have

$$\frac{a_i}{1 + s_{i-1}} \leq 2\log\left(1 + \frac{a_i}{1 + s_{i-1}}\right) = 2[\log(1 + s_i) - \log(1 + s_{i-1})].$$

Summing over $i \in [n]$, we note that the sum telescopes and we obtain

$$\sum_{i=1}^{n} \frac{a_i}{1 + \sum_{j=1}^{i-1} a_j} \leq 2\log(1 + s_n),$$

concluding the proof. ■

Lemma 16 *Assume Hedge is run on $\Delta(\Pi)$ with rewards $u_1, \ldots, u_T \in [0,R]^{\Pi}$ and step size $0 < \eta \leq 1/R$. Then for any $p^\star \in \Delta(\Pi)$ it holds that*

$$\sum_{t=1}^{T} u_t^\top p^* \leq (1 + \eta R) \sum_{t=1}^{T} u_t^\top p_t + \frac{\log|\Pi|}{\eta}.$$

Proof This follows directly from the standard second-order bound for Hedge with signed losses $\ell_t = -u_t$ (see e.g. Alon et al. (2015), Lemma 10): for all $p^* \in \Delta(\Pi)$ (and crucially since $\eta\ell_t(i) \geq -1$ for all t, i):

$$\sum_{t=1}^{T} \ell_t^\top p_t - \sum_{t=1}^{T} \ell_t^\top p^\star \leq \frac{\log|\Pi|}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2,$$

which translates to the stated bound by plugging $\ell_t = -u_t$ and using $p_t \cdot \ell_t^2 \leq R p_t^\top u_t$. ■

Lemma 17 *Suppose that $P, Q \in \Delta(\mathcal{Z})$. Then for any $f : \mathcal{Z} \to [-1,1]$, it holds that*

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq 4\sqrt{\mathbb{E}_Q[f^2] \cdot D_{\mathrm{H}}^2(P,Q)} + 4D_{\mathrm{H}}^2(P,Q).$$

Proof We denote $P(z)$ (resp. $Q(z)$) to be the density function of P (resp. Q). Then for any function $f : \mathcal{Z} \to \mathbb{R}$,

$$\begin{aligned}
|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|^2 &= \left(\int_{\mathcal{Z}} f(z)(P(z) - Q(z))dz\right)^2 \\
&\leq \int_{\mathcal{Z}} f(z)^2(\sqrt{P(z)} + \sqrt{Q(z)})^2 dz \cdot \int_{\mathcal{Z}} (\sqrt{P(z)} - \sqrt{Q(z)})^2 dz \\
&\leq 4D_{\mathrm{H}}^2(P,Q) \cdot \Big(\mathbb{E}_Q[f^2] + \mathbb{E}_P[f^2]\Big).
\end{aligned}$$

In particular, when $h : \mathcal{Z} \to [0,1]$, the inequality above implies that

$$|\mathbb{E}_P[h] - \mathbb{E}_Q[h]| \le 2D_{\mathrm{H}}(P,Q)\sqrt{(\mathbb{E}_P[h] + \mathbb{E}_Q[h])} \le \frac{1}{2}(\mathbb{E}_P[h] + \mathbb{E}_Q[h]) + 2D^2_{\mathrm{H}}(P,Q),$$

and hence it holds that $\mathbb{E}_P[h] \le 3\mathbb{E}_Q[h] + 4D^2_{\mathrm{H}}(P,Q)$.

Now, we can bound

$$\begin{aligned}|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|^2 &\le 4D^2_{\mathrm{H}}(P,Q) \cdot \Big(\mathbb{E}_Q[f^2] + \mathbb{E}_P[f^2]\Big) \\ &\le 16D^2_{\mathrm{H}}(P,Q) \cdot \Big(\mathbb{E}_Q[f^2] + D^2_{\mathrm{H}}(P,Q)\Big).\end{aligned}$$

This gives the desired upper bound. ■