Proceedings of Machine Learning Research vol 125:1–52, 2020 33rd Annual Conference on Learning Theory

Taking a hint: How to leverage loss predictors in contextual bandits?

Chen-Yu Wei CHENYU.WEI@USC.EDU
University of Southern California

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research, Redmond

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We initiate the study of learning in contextual bandits with the help of loss predictors. The main question we address is whether one can improve over the minimax regret $\mathcal{O}(\sqrt{T})$ for learning over T rounds, when the total error of the predicted losses relative to the realized losses, denoted as $\mathcal{E} \leq T$, is relatively small. We provide a complete answer to this question, with upper and lower bounds for various settings: adversarial and stochastic environments, known and unknown $\mathcal{E}$, and single and multiple predictors. We show several surprising results, such as 1) the optimal regret is $\mathcal{O}(\min\{\sqrt{T}, \sqrt{\mathcal{E}}T^{\frac{1}{4}}\})$ when $\mathcal{E}$ is known, in contrast to the standard and better bound $\mathcal{O}(\sqrt{\mathcal{E}})$ for non-contextual problems (such as multi-armed bandits); 2) the same bound cannot be achieved if $\mathcal{E}$ is unknown, but as a remedy, $\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ is achievable; 3) with M predictors, a linear dependence on M is necessary, even though logarithmic dependence is possible for non-contextual problems.

We also develop several novel algorithmic techniques to achieve matching upper bounds, including 1) a key *action remapping* technique for optimal regret with known $\mathcal{E}$, 2) computationally efficient implementation of Catoni's robust mean estimator via an ERM oracle in the stochastic setting with optimal regret, 3) an underestimator for $\mathcal{E}$ via estimating the histogram with bins of exponentially increasing size for the stochastic setting with unknown $\mathcal{E}$, and 4) a self-referential scheme for learning with multiple predictors, all of which might be of independent interest.

1. Introduction

Online learning with the help of loss predictors has been widely studied over the past decade. In these problems, before making a decision at each round t, the learner is given some prediction m_t of the true gradient or loss vector ℓ_t. The goal is to ensure regret that is much smaller than the worst-case bound as long as these predictions are indicative of the loss vectors. For example, for most problems with $\Theta(\sqrt{T})$ minimax regret for learning over T rounds, it has been shown that a more adaptive regret bound of order $\mathcal{O}(\sqrt{\mathcal{E}})$ is possible, where $\mathcal{E} = \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$ is the total error of the predictions, which is at most $\mathcal{O}(T)$ but could be much smaller if a good predictor is available. Such a bound is achievable for problems with full information feedback (Rakhlin and Sridharan, 2013a; Steinhardt and Liang, 2014), as well as partial information feedback such as multi-armed bandits (Wei and Luo, 2018) and linear bandits (Rakhlin and Sridharan, 2013a).

In contextual bandits (Auer et al., 2002; Langford and Zhang, 2008), a generalization of multi-armed bandits that has been proven to be useful for applications such as personalized recommendation systems in practice, loss predictors are also commonly used to construct doubly-robust esti-

Table 1: Summary of main results. T is the total number of rounds. For single predictor, $\mathcal{E} \leq T$ is the total error of predictions. For multiple predictors, $\mathcal{E}^*$ is the total error of the best predictor and M is the number of predictors. Dependence on other parameters is omitted. Note that for the case with known $\mathcal{E}$ or $\mathcal{E}^*$, one can achieve the minimum of $\mathcal{O}(\sqrt{T})$ and the stated upper bound by simply comparing the two bounds and choosing between the minimax algorithm and our algorithms.

	Single predictor with known $\mathcal{E}$	Single predictor with unknown $\mathcal{E}$	Multiple predictors with known $\mathcal{E}^*$
Lower bound for $\mathcal{E}, \mathcal{E}^*, M = \mathcal{O}(\sqrt{T})$	$\Omega(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ [Theorem 1]	$\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ is impossible [Theorem 2]	$\Omega(\sqrt{\mathcal{E}^*}T^{\frac{1}{4}} + M)$ [Theorem 3]
Upper bound in the adversarial setting	$\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ [Theorem 4]	$\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ [Theorem 5]	$\mathcal{O}(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}})$ [Theorem 10]
Upper bound in the i.i.d. setting with oracle-efficient algorithms	$\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ [Theorem 8]	$\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ [Theorem 9]	$\mathcal{O}(M^{\frac{2}{3}}(\mathcal{E}^*T)^{\frac{1}{3}})$ [Theorem 11]

mators, both for off-policy evaluation (Dudík et al., 2014) and online exploration (Agarwal et al., 2014). The potentially lower variance of these doubly-robust estimators has been used to motivate this line of work, and resulting improvements are well established for policy evaluation in both finite sample (Dudík et al., 2014) and asymptotic settings (Robins and Rotnitzky, 1995). In the online exploration setting, however, the extent of benefits from a good loss predictor and potential rate improvements beyond the worst case $O(\sqrt{T})$ bound have not been studied at all, despite all the works mentioned above for the simpler non-contextual settings.

In this work, we take the first attempt in addressing this question and provide a rather complete answer on upper and lower bounds for various setups: adversarial and stochastic environments, known and unknown $\mathcal{E}$, and single and multiple predictors. The main message is that good predictors indeed help reduce regret for contextual bandits, but *not to the same extent as the non-contextual settings*. Specifically, our contributions are (see also Table 1 for a summary):

- (Section 3) In the adversarial setting where contexts, losses, and predictions are all decided by an adversary, we show that, somewhat surprisingly, the regret is at least $\Omega(\min\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\})$, and we also provide an algorithm with a matching regret upper bound when $\mathcal{E}$ is known. When $\mathcal{E}$ is unknown, we show that it is impossible to achieve the same bound, and as a remedy, we provide an adaptive version of our algorithm with regret $\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$, which is always sublinear and is better than $\mathcal{O}(\sqrt{T})$ as long as $\mathcal{E} = o(T^{\frac{1}{3}})$. Note that these results are in sharp contrast with the typical bound $\mathcal{O}(\sqrt{\mathcal{E}})$, for non-contextual problems. For multi-armed bandits, even with unknown $\mathcal{E}$, $\mathcal{O}(\sqrt{\mathcal{E}})$ is achievable (Wei and Luo, 2018), indicating that the difficulty indeed comes from the contexts, and not just the bandit feedback.
- (Section 4) In the stochastic setting where contexts, losses, and predictions are jointly i.i.d. samples from a fixed and unknown distribution, we show the exact same lower and upper bounds with known or unknown $\mathcal{E}$, but importantly our algorithms are efficient assuming access to some ERM oracle. This a typical computational model for studying efficient contextual bandits algorithms, and avoiding running time that is polynomial in the number of polices (Langford and

Zhang, 2008; Agarwal et al., 2014; Syrgkanis et al., 2016b). Somewhat surprisingly, we find that an adaptation of ϵ-greedy exploration is optimal when $\mathcal{E} = \mathcal{O}(\sqrt{T})$.

- (Section 5) Finally, we extend our results to the setting where M predictors are available and the goal is to improve the $\mathcal{O}(\sqrt{T})$ regret as long as the total error $\mathcal{E}^*$ of the best predictor is relatively small. For simplicity we assume known $\mathcal{E}^*$. We show a lower bound $\Omega(\min\{\sqrt{\mathcal{E}^*}T^{\frac{1}{4}}+M, \sqrt{T}\})$ when $M = \mathcal{O}(\sqrt{T})$, as well as an upper bound of $\mathcal{O}(\min\{(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}} + M, \sqrt{T}\})$ for the adversarial setting and an upper bound of $\mathcal{O}(\min\{M^{\frac{2}{3}}(\mathcal{E}^*T)^{\frac{1}{3}}, \sqrt{T}\})$ for the stochastic setting with an oracle-efficient algorithm. This is also in contrast with the non-contextual settings where the dependence on M is logarithmic, even with bandit feedback (Rakhlin and Sridharan, 2013a).

Throughout, we focus on finite action and policy sets in this work to cleanly illustrate the key ideas. Extensions to infinite actions and policies are interesting avenues for future work.

Techniques. Our algorithms require several novel techniques, briefly summarized below:

- Most importantly, all our algorithms rely on an *action remapping* technique, which restricts the algorithm's attention to only a subset of actions at each round. This subset consists of actions with predicted loss not larger than that of a baseline action (such as the action with the smallest predicted loss) by a certain amount, and the algorithms pretend that all actions outside this set are just the baseline action. For the adversarial setting with multiple predictors, we also need to apply a self-referential scheme to find the baseline and construct this subset, an idea similar to sleeping experts (Freund et al., 1997). We prove that this action remapping technique reduces both the exploration overhead and the variance of estimators.
- Our algorithms for the stochastic setting require using robust mean estimators. In particular, we use the Catoni's estimator (Catoni, 2012) and show that it can be implemented efficiently using the ERM oracle, which might be of independent interest and useful for developing oracle-efficient algorithms for other problems.
- When $\mathcal{E}$ is unknown, we construct a novel underestimator of $\mathcal{E}$ by estimating the histogram of the distribution of $\|\ell_t - m_t\|$ with bins of exponentially increasing size in the stochastic setting.

Related work. Similar to prior work such as (Rakhlin and Sridharan, 2013a) (for non-contextual problems), we consider generic loss predictions given by any predictors as inputs of the algorithm. A series of works focus on choosing specific predictions based on observed data and deriving data-dependent bounds in terms of the variation of the environment (Hazan and Kale, 2010, 2011; Chiang et al., 2012, 2013; Steinhardt and Liang, 2014; Wei and Luo, 2018; Bubeck et al., 2019), which are themselves useful for applications such as faster convergence to equilibria for game playing (Rakhlin and Sridharan, 2013b; Syrgkanis et al., 2015; Wei and Luo, 2018). Whether similar applications can be derived based on our results is an interesting future direction.

EXP4 is the classic algorithm with optimal regret $\mathcal{O}(\sqrt{T})$ for the adversarial setting (Auer et al., 2002), albeit with running time linear in the number of policies. For the stochastic setting, the simple ϵ-greedy algorithm is oracle-efficient but with suboptimal regret $\mathcal{O}(T^{\frac{2}{3}})$ (Langford and Zhang, 2008). Later, Agarwal et al. (2014) proposed an oracle-efficient and optimal algorithm ILOVETOCONBANDITS. All these algorithms are building blocks for our methods.

Developing adaptive regret bounds for contextual bandits is relatively under-explored. The only existing work on contextual learning that considers a similar setting with loss predictors is (Syrgkanis et al., 2016a, Section 6), but they only consider the easier full-information feedback. On a different direction, Allen-Zhu et al. (2018) derived the first small-loss bound for contextual bandits.

Our idea of using robust estimators is inspired by (Krishnamurthy et al., 2019), which studies contextual bandits with continuous actions and uses median-of-means, a standard robust estimator, for a different purpose. It is unclear whether median-of-means can be implemented efficiently via an ERM oracle. Instead, we turn to Catoni's estimator (Catoni, 2012), which provides a similar concentration guarantee and can be implemented efficiently as we show.

2. Problem Description and Lower Bounds

Contextual bandits is a generalization of the classic multi-armed bandit problem, where before choosing one of the K actions at each round, the learner observes a context from some arbitrary context space $\mathcal{X}$. In addition to the context, we consider a variant where a *loss predictor* is also available. Specifically, for each round $t = 1, \ldots, T$, the environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and a loss predictor $m_t \in [0,1]^K$; the learner then receives x_t and m_t; finally, the learner chooses an action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$.

We consider both the adversarial setting and the stochastic setting. In the former, the sequence $(x_t, \ell_t, m_t)_{1:T}$ can be arbitrary and even depend on the learner's strategy. For simplicity we assume it is decided ahead of time before the game starts (also known as the oblivious setting). In the latter, each triple (x_t, m_t, ℓ_t) is drawn independently from a fixed and unknown distribution $\mathcal{D}$.

As in the standard contextual bandits setup, the learner has access to some fixed policy class $\Pi \subseteq [K]^{\mathcal{X}}$, assumed to be finite (for simplicity) with cardinality N, and her goal is to minimize the (pseudo) regret against the best fixed policy:

$$\mathrm{Reg}(\mathcal{A}) \triangleq \max_{\pi \in \Pi} \mathbb{E}\Big[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi(x_t))\Big],$$

where the expectation is with respect to the randomness of the learner, denoted as the algorithm $\mathcal{A}$ which chooses $a_1, \ldots, a_T$, and also that of the environment in the stochastic case. When it is clear from the context, we omit the dependence on $\mathcal{A}$ and simply denote the regret by Reg. It is well-known that the optimal worst-case regret is $\mathcal{O}(\sqrt{dT})$ where we define $d \triangleq K \ln N$.[1] The key question we address in this work is whether one could improve upon this worst-case bound when the predictor is accurate. More specifically, we denote the total loss of the predictor by $\mathcal{E} \triangleq \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$ for the adversarial setting and $\mathcal{E} \triangleq T\mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}\left[\|\ell - m\|_\infty^2\right]$ for the stochastic setting, and we ask the following question:

(Q1) *Can we improve the regret over $\mathcal{O}(\sqrt{dT})$ if $\mathcal{E} = o(T)$?*

Note that for the special case of multi-armed bandits where Π consists of K constant mappings that always pick one of the K actions (that is, contexts are ignored), Wei and Luo (2018) show that $\mathcal{O}(\sqrt{d\mathcal{E}})$ is achievable, an improvement over $\mathcal{O}(\sqrt{dT})$ *as long as* $\mathcal{E} = o(T)$. A natural guess would be that the same holds true for contextual bandits. However, somewhat surprisingly, in the following theorem we show that this is not the case (proofs for all lower bounds are deferred to Appendix B).

Theorem 1 *For any algorithm and any value $V \in [0, T]$, there exists a (stochastic or adversarial) environment with $\mathcal{E} \le V$ and $N = \Theta(\sqrt{KT})$ such that* $\mathrm{Reg}(\mathcal{A}) = \widetilde{\Omega}\big(\min\big\{\sqrt{V}(KT)^{\frac{1}{4}}, \sqrt{KT}\big\}\big)$.[2]

1. Throughout the paper, we do not make an effort to optimize the dependence on K and $\ln N$. For example, we often relax $K^2 \ln N$ by d^2 for ease of presentation. For most discussions, we also ignore the dependence on d and only focus on the dependence on T and $\mathcal{E}$.
2. Note that while seemingly a lower bound for the stochastic environments should imply the same for the adversarial environments, there is a subtle technical difference due to the slightly different definitions of $\mathcal{E}$ in these two cases. We provide proofs for both environments.

This theorem gives a negative answer to **(Q1)** when $\mathcal{E} = \Omega(\sqrt{T})$. Even when $\mathcal{E} = \mathcal{O}(1)$, the theorem shows that the best one can achieve is $\mathcal{O}(T^{\frac{1}{4}})$, a sharp contrast with the non-contextual case. Note that we require $N \geq K$ due to Wei and Luo (2018), but perhaps the condition of $N = \Theta(\sqrt{KT})$ can be further weakened. In Sections 3 and 4, we develop algorithms with matching upper bounds for adversarial and stochastic environments respectively, thus completely answering **(Q1)** and confirming that in the worst case, loss predictors are helpful *if and only if* $\mathcal{E} = o(\sqrt{T})$.

Robustness when $\mathcal{E}$ is unknown. One shortcoming of our algorithms with matching upper bounds is that they require knowing the value of $\mathcal{E}$, which is clearly undesirable in practice. Put differently, for each possible value of $\mathcal{E}$, we need a different setting of algorithm parameters to achieve the optimal bound. Therefore, the next general question we ask is:

(Q2) *Is there an algorithm with regret $o(\sqrt{T})$ simultaneously for all environments with $\mathcal{E} = o(\sqrt{T})$?*

One standard method in online learning to deal with unknown parameters is the so-called doubling trick, which is applicable even for some partial-information settings (Hazan and Kale, 2011; Wei and Luo, 2018). However, we show yet another surprising result that the answer to **(Q2)** is *no*.

Theorem 2 *If an algorithm $\mathcal{A}$ achieves $\mathrm{Reg}(\mathcal{A}) = o(\sqrt{T})$ for all environments with $\mathcal{E} = 0$, then there exists another environment with $\mathcal{E} = o(\sqrt{T})$ and $N = \Omega(T)$ for which $\mathrm{Reg}(\mathcal{A}) = \omega(\sqrt{T})$. Thus, no algorithm can achieve $\mathrm{Reg}(\mathcal{A}) = \mathcal{O}\big(\min\big\{\sqrt{\mathcal{E}}(dT)^{\frac{1}{4}}, \sqrt{dT}\big\}\big)$ simultaneously for all $\mathcal{E}$.*

The theorem asserts that no algorithm can improve over $\mathcal{O}(\sqrt{T})$ when good predictors are available while simultaneously maintaining $\mathcal{O}(\sqrt{T})$ worst-case robustness. As a remedy, nevertheless, we develop adaptive versions of our algorithms with regret $\mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ for all environments simultaneously. This bound is $o(\sqrt{T})$ whenever $\mathcal{E} = o(T^{\frac{1}{3}})$ and at the same time provides a robustness guarantee of $\mathcal{O}(T^{\frac{5}{6}})$. As a comparison, a bound of order $\mathcal{O}(\mathcal{E}T^{\frac{1}{4}})$, achievable by naively setting the parameters of our algorithms independent of $\mathcal{E}$, is $o(\sqrt{T})$ only when $\mathcal{E} = o(T^{\frac{1}{4}})$, and more importantly could be linear when $\mathcal{E}$ is large and thus provides no robustness guarantee at all.

Learning with multiple predictors. Having a complete understanding of the single predictor case, we further consider a more general setup where instead of receiving one predictor m_t, the learner receives M predictors $m_t^1, \ldots, m_t^M \in [0,1]^K$ at the beginning of each round. In the adversarial setting, these are decided ahead of time by an adversary, and we denote by $\mathcal{E}^* \triangleq \min_{i\in[M]} \sum_{t=1}^T \|\ell_t - m_t^i\|_\infty^2$, the total error of the best predictor. On the other hand, for the stochastic setting, each tuple $(x_t, \ell_t, m_t^1, \ldots, m_t^M)$ is an i.i.d. sample of a fixed distribution $\mathcal{D}$, and we denote by $\mathcal{E}^* \triangleq T \min_{i\in[M]} \mathbb{E}_{(x,\ell,m^{1:M})\sim\mathcal{D}}\left[\|\ell - m^i\|_\infty^2\right]$, the expected total error of the best predictor.

The goal of the learner is to improve over the worst-case bound as long as *one of the predictors* is reasonably accurate. Specifically, we ask the following (assuming known $\mathcal{E}^*$ for simplicity).

(Q3) *Can we improve the regret over $\mathcal{O}(\sqrt{dT})$ for $\mathcal{E}^* = o(\sqrt{T})$ and reasonably small M?*

For many online learning problems (even those with partial information), achieving $\mathcal{O}(\sqrt{\mathcal{E} + \ln M})$ is possible (Rakhlin and Sridharan, 2013a). We already know that a worse dependence on T is necessary for contextual bandits, and it turns out that, a worse dependence on M is also unavoidable.

Theorem 3 *For any algorithm $\mathcal{A}$ and any $M \leq \sqrt{T}$ and $V^* \leq \sqrt{T}$, there exists an environment (which can be stochastic or adversarial) with $\mathcal{E}^* \leq V^*$ such that $\mathrm{Reg}(\mathcal{A}) = \widetilde{\Omega}(\sqrt{V^*}(KT)^{\frac{1}{4}} + M)$.*

Compared to the single predictor case, the lower bound has an extra term linear in M. It shows that when $M = \Omega(\sqrt{T})$, there is no hope to improve the worst-case regret even if there is a perfect predictor such that $\mathcal{E}^* = 0$, again a sharp contrast with the non-contextual case. In Section 5, we provide an algorithm with regret $\mathcal{O}(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}})$ for the adversarial setting, and another oracle-efficient algorithm with regret $\mathcal{O}(M^{\frac{2}{3}}(\mathcal{E}^*T)^{\frac{1}{3}})$ for the stochastic setting, answering **(Q3)** positively to some extent. (Note that these bounds are larger than the lower bound when $M \le \sqrt{T}$.)

Other notations. We use $\widetilde{\mathcal{O}}(\cdot)$ to hide the dependence on $\ln T$, and $\widetilde{\Omega}(\cdot)$ to hide the dependence on $1/\ln T$; for an integer n, $[n]$ represents $\{1, \dots, n\}$; for a random variable Z, $\mathbb{V}[Z]$ denotes its variance; Δ_Π and Δ_K are the sets of all distributions over the polices and the actions respectively.

3. Algorithms for Adversarial Environments

In this section, we describe our algorithm for the adversarial setting with one predictor. Similar to existing works on online learning with loss predictors, our algorithm is based on the optimistic Online Mirror Descent (OMD) framework (Rakhlin and Sridharan, 2013a). In particular, with the entropy regularizer, the optimistic OMD update maintains a sequence of distributions

$$Q'_1, \dots, Q'_T \in \Delta_\Pi, \text{ such that } Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp\big(-\eta \widehat{\ell}_t(\pi(x_t))\big),$$

where $\eta > 0$ is the learning rate and $\widehat{\ell}_t$ is some estimator for ℓ_t. Upon seeing a context x_t and a predictor m_t at time t, the algorithm computes $Q_t \in \Delta_\Pi$ such that $Q_t(\pi) \propto Q'_t(\pi) \exp\big(-\eta m_t(\pi(x_t))\big)$, and samples a policy according to Q_t and follows its suggestion to choose an action a_t. Suppose $p_t \in \Delta_K$ is the distribution of a_t, then the standard variance-reduced loss estimator is

$$\widehat{\ell}_t(a) = \frac{(\ell_t(a) - m_t(a))\mathbb{1}[a_t = a]}{p_t(a)} + m_t(a). \tag{1}$$

When $m_t(a) = 0$ for all t and a, this is exactly the EXP4 algorithm (Auer et al., 2002).

While optimistic OMD with entropy regularizer has been used for problems with full-information feedback (Steinhardt and Liang, 2014; Syrgkanis et al., 2015), it in fact cannot be directly applied to the bandit setting since typical analysis requires $\widehat{\ell}_t(a) - m_t(a)$ to be lower bounded by $-1/\eta$, which does not hold if $\ell_t(a_t) \le m_t(a_t)$ and $p_t(a_t)$ is too small. Intuitively this is also the hard case because the predictor over-predicts the loss of a good action and prevents the algorithm from realizing it due to the bandit feedback. A naive approach of enforcing uniform exploration so that $p_t(a_t) \ge \eta$ contributes ηTK regret already, which eventually leads to $\Omega(\sqrt{T})$ regret. Indeed, to get around this issue for multi-armed bandits, Wei and Luo (2018) uses a different regularizer called log-barrier, but this does not work for contextual bandits either since it inevitably introduces polynomial dependence on the number of policies N for the regret.

Our solutions. Our first key observation is that, despite the range of the loss estimators, Optimistic Exp4 in fact *always* guarantees the following (*cf.* Lemma 14): for any $\pi^* \in \Pi$,

$$\sum_{t=1}^{T} \sum_{\pi \in \Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t)) - \sum_{t=1}^{T} \widehat{\ell}_t(\pi^*(x_t)) \le \frac{\ln N}{\eta} + 2\eta \sum_{t=1}^{T} (\widehat{\ell}_t(a_t) - m_t(a_t))^2. \tag{2}$$

Readers familiar with the EXP4 analysis would find that $p_t(a_t)$ is missing in the last term compared to the standard analysis when $\widehat{\ell}_t(a) - m_t(a) \ge -1/\eta$ holds. To see why Eq. (2) is useful, first

Algorithm 1 EXP4.OAR: Optimistic EXP4 with Action Remapping

Parameter: learning rate $\eta > 0$, threshold $\sigma > 0$, exploration probability $\mu \in [0, 1]$.
Initialize: $Q'_1(\pi) = 1/N$ for all $\pi \in \Pi$.
for $t = 1, \dots, T$ **do**

Receive x_t and m_t. Define $a^*_t = \operatorname{argmin}_{a\in[K]} m_t(a)$,

$$\mathcal{A}_t = \{a \in [K] : m_t(a) \le m_t(a^*_t) + \sigma\}, \quad \text{and} \quad \phi_t(a) = \begin{cases} a, & \text{if } a \in \mathcal{A}_t, \\ a^*_t, & \text{otherwise.} \end{cases} \tag{3}$$

Calculate $Q_t \in \Delta_\Pi$: $\quad Q_t(\pi) \propto Q'_t(\pi) \exp\left(-\eta m_t\left(\phi_t(\pi(x_t))\right)\right)$.
Calculate $p_t \in \Delta_K$: $\quad p_t(a) = (1-\mu)\sum_{\pi:\phi_t(\pi(x_t))=a} Q_t(\pi) + \frac{\mu}{|\mathcal{A}_t|}\mathbb{1}[a \in \mathcal{A}_t]$.
Sample $a_t \sim p_t$ and receive $\ell_t(a_t)$.
Construct estimator: $\quad \widehat{\ell}_t(a) = \frac{\ell_t(a)-m_t(a)}{p_t(a)}\mathbb{1}[a_t = a] + m_t(a)$ for all $a \in \mathcal{A}_t$.
Calculate $Q'_{t+1} \in \Delta_\Pi$: $\quad Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp\left(-\eta\widehat{\ell}_t\left(\phi_t(\pi(x_t))\right)\right)$.

take expectation (over a_t) on both sides so the last term is bounded by $2\eta K \sum_t \frac{\|\ell_t - m_t\|^2_\infty}{\min_a p_t(a)}$. Then consider enforcing uniform exploration so that $p_t(a) \ge \mu/K$ holds for some $\mu \in [0, 1]$. Since this contributes μT extra regret, using Eq. (2) we have $\mathrm{Reg} = \mathcal{O}(\frac{\ln N}{\eta} + \frac{\eta K^2 \mathcal{E}}{\mu} + \mu T)$, which, with the optimal tuning of η and μ, already gives a nontrivial bound $\mathrm{Reg} = \mathcal{O}((\mathcal{E}T)^{\frac{1}{3}})$! This bound is also $o(\sqrt{T})$ whenever $\mathcal{E} = o(\sqrt{T})$, but is worse than the bound $\mathrm{Reg} = \mathcal{O}(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ we are aiming for.

To further improve the algorithm, we introduce a novel *action remapping* technique. Specifically, let $a^*_t = \operatorname{argmin}_{a\in[K]} m_t(a)$ be the action with smallest predicted loss and let $\mathcal{A}_t$ (Equation 3) be the set of actions with predicted loss not larger than that of a^*_t by σ, for some threshold $\sigma \ge 0$. Then, we rename the actions according to a mapping $\phi_t : [K] \to \mathcal{A}_t$ such that $\phi_t(a) = a$ for $a \in \mathcal{A}_t$ and $\phi_t(a) = a^*_t$ for $a \notin \mathcal{A}_t$. In other words, we pretend that every action outside $\mathcal{A}_t$ was just a^*_t. We call our algorithm EXP4.OAR and show its pseudocode in Algorithm 1.

To see why this action remapping is useful, first consider the regret compared to $\sum_{t=1}^T \ell_t(\phi_t(\pi^*(x_t)))$ due to exploration. Note that we only explore actions in $\mathcal{A}_t$ and all actions in this set have predicted loss σ-close to each other. Therefore, exploration leads to regret $\mu T\sigma + 2\mu\sum_t \|\ell_t - m_t\|_\infty \le \mu T\sigma + 2\mu\sqrt{\mathcal{E}T}$, instead of μT compared to the naive approach. On the other hand, the bias due to remapping $\ell_t(\phi_t(\pi^*(x_t))) - \ell_t(\pi^*(x_t))$ is either zero if $\pi^*(x_t) \in \mathcal{A}_t$ or at most $2\|\ell_t - m_t\|_\infty - \sigma$ otherwise (by adding and subtracting $m_t(a^*_t)$ and $m_t(\pi^*(x_t))$). Using the AM-GM inequality and summing over t gives $\mathcal{E}/\sigma$. Combining everything we prove the following theorem.

Theorem 4 EXP4.OAR *(Algorithm 1) ensures* $\mathrm{Reg} \le \frac{\ln N}{\eta} + \frac{2\eta K^2\mathcal{E}}{\mu} + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T} + \frac{\mathcal{E}}{\sigma}$. *Picking* $\mu = \min\left\{\frac{d}{\sqrt{T}}, 1\right\}$, $\eta = \sqrt{\frac{\mu \ln N}{K^2\mathcal{E}}}$, *and* $\sigma = \sqrt{\frac{\mathcal{E}}{\mu T}}$ *gives* $\mathrm{Reg} = \mathcal{O}\big(\sqrt{d\mathcal{E}}(T)^{\frac{1}{4}} + d\sqrt{\mathcal{E}}\big)$.

See Appendix C.1 for the complete proof. This theorem indicates that whenever the predictor is good enough with $\mathcal{E} = o(\sqrt{T})$, our algorithm improves over EXP4 and achieves $o(\sqrt{T})$ regret. Note that this bound requires setting the parameters in terms of the quantity $\mathcal{E}$, and in the case when $\mathcal{E} = \Omega(\sqrt{T})$, one can simply switch to EXP4 and achieve regret $\mathcal{O}(\sqrt{dT})$. Therefore, our result indeed matches the lower bound stated in Theorem 1 (except for a slightly worse dependence on d).

Adaptive version with unknown $\mathcal{E}$. Next, we discuss the case when $\mathcal{E}$ is unknown. Recall that there is no hope to maintain the same bound of Theorem 4 in this case, as indicated by Theorem 2. Standard doubling trick does not work due to the large magnitude of loss estimators (more specifically, the last round before each restart causes some technical problems), even though it works for non-contextual problems with bandit feedback (Hazan and Kale, 2011; Wei and Luo, 2018).

In light of Eq. (2), our solution is to use a time-varying learning rate η_t that is roughly of order $\left(\sum_{s\le t}(\widehat{\ell}_s(a_s) - m_s(a_s))^2\right)^{-1/2}$ to minimize the right hand side of Eq. (2) for each time. While standard analysis requires using the same learning rate in Line 2 and Line 6, due to technical issues we are unable to do so while achieving the desired regret bound. Instead, we use η_{t-1} in Line 2 and η_t in Line 6, and carefully bound the bias introduced by this learning rate mismatch. More details are provided in Appendix C.2. Our algorithm (Algorithm 4 in Appendix C.2) is completely adaptive, requiring no prior information about $\mathcal{E}$. The following theorem gives its regret guarantee.

Theorem 5 EXP4.OVAR *(Algorithm 4 in Appendix C.2) ensures* $\mathrm{Reg} = \widetilde{\mathcal{O}}\left(d\sqrt{\frac{\mathcal{E}}{\mu}} + \mu T\right)$. *Specifically, setting* $\mu = \min\left\{1, (d/T)^{\frac{2}{3}}\right\}$ *gives* $\mathrm{Reg} = \widetilde{\mathcal{O}}\left(d\sqrt{\mathcal{E}} + \sqrt{\mathcal{E}}(d^2T)^{\frac{1}{3}}\right)$.

This shows that our algorithm is robust and always ensures sublinear regret, since in the worst case $\mathrm{Reg} = \widetilde{\mathcal{O}}(T^{5/6})$ (when $\mathcal{E} = T$). Also, our algorithm improves over EXP4 whenever $\mathcal{E} = o(T^{\frac{1}{3}})$.

4. Algorithms for Stochastic Environments

In this section, we consider learning in a stochastic environment with one predictor. Recall that a stochastic environment is parameterized by an unknown distribution $\mathcal{D}$ such that each triple (x_t, ℓ_t, m_t) is an i.i.d. sample from $\mathcal{D}$ and the total prediction error is $\mathcal{E} = T\mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}\left[\|\ell - m\|_\infty^2\right]$. Clearly, this is a special case of the adversarial environment, and our goal is to derive the same results but with *oracle-efficient* algorithms.

Specifically, an ERM oracle is a procedure that takes any set $\mathcal{S}$ of context-loss pairs $(x, c) \in \mathcal{X} \times \mathbb{R}^K$ as inputs and outputs a policy $\mathrm{ERM}(\mathcal{S}) \in \operatorname{argmin}_{\pi\in\Pi}\sum_{(x,c)\in\mathcal{S}} c(\pi(x))$. An algorithm is oracle-efficient if its total running time and the number of oracle calls are both polynomial in T and d, excluding the running time of the oracle itself. Oracle-efficiency has been proven to be impossible for adversarial environments (Hazan and Koren, 2016), but achievable for stochastic environments. The simplest oracle-efficient algorithm is ϵ-greedy (Langford and Zhang, 2008), with suboptimal regret $\mathcal{O}(T^{\frac{2}{3}})$. However, somewhat surprisingly, we are able to build our algorithm on top of ϵ-greedy and achieve optimal results when $\mathcal{E} = o(\sqrt{T})$.

We first review the ϵ-greedy algorithm and point out the difficulties of improving its regret with loss predictors. In each round t, the algorithm with probability μ samples an action a_t uniformly at random, and with probability $1-\mu$ follows the empirically best policy $\pi_t = \mathrm{ERM}\left(\{x_s, \widehat{\ell}_s\}_{s<t}\right)$ by choosing $a_t = \pi_t(x_t)$, where $\widehat{\ell}_s$ is the standard importance-weighted estimator for round s.

By standard concentration arguments (Freedman inequality), it holds with high probability that the difference between the average estimated loss and the expected loss of any policy π is bounded as

$$\left|\tfrac{1}{t}\sum_{s=1}^t \widehat{\ell}_s(\pi(x_s)) - \mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}[\ell(\pi(x))]\right| \le \widetilde{\mathcal{O}}\left(\tfrac{1}{t}\sqrt{(\ln N)\sum_{s=1}^t \mathbb{V}_s\left[\widehat{\ell}_s(\pi(x_s))\right]} + \tfrac{d}{\mu t}\right),$$

where $\mathbb{V}_s[\widehat{\ell}_s(\pi(x_s))]$ is the conditional variance (given everything before round s) and is at most K/μ. By the optimality of π_t, it is then clear that the total regret of following the empirically best

policy is $\widetilde{\mathcal{O}}\big(\sum_t \big(\sqrt{d/\mu t} + d/\mu t\big)\big) = \widetilde{\mathcal{O}}\big(\sqrt{dT/\mu} + d/\mu\big)$. Further taking the uniform exploration into account shows that the regret of ϵ-greedy has three components: the *variance term* $\widetilde{\mathcal{O}}(\sqrt{dT/\mu})$, the *lower-order term* $\widetilde{\mathcal{O}}(d/\mu)$, and the *exploration term* $\mathcal{O}(\mu T)$. Picking the optimal μ gives $\mathcal{O}\big(T^{\frac{2}{3}}\big)$ regret. To improve the bound, we improve each of these three terms as described below.

Improving variance/exploration terms via action remapping. One natural idea to improve the variance term is to deploy the same variance-reduced (also known as doubly-robust) estimator $\widehat{\ell}_t$ (Eq. (1)) as in the adversarial case. However, the law of total variance implies:

$$\mathbb{V}_t[\widehat{\ell}_t(\pi(x_t))] = \mathbb{E}_{x_t,m_t,\ell_t}[\mathbb{V}_{a_t}[\widehat{\ell}_t(\pi(x_t))|x_t,m_t,\ell_t]] + \mathbb{V}_{x_t,m_t,\ell_t}[\mathbb{E}_{a_t}[\widehat{\ell}_t(\pi(x_t))|x_t,m_t,\ell_t]],$$

where one can verify that the first term is at most $\frac{K\mathcal{E}}{\mu T}$, but the second term is just $\mathbb{V}_{x_t,m_t,\ell_t}[\ell_t(\pi(x_t))]$ and is not related to $\mathcal{E}$. Simply bounding the second term by 1 leads to $\Omega(\sqrt{T})$ regret already.

We propose to address this issue by first shifting the variance-reduced estimator by $m_t(a_t^*)$, where $a_t^* = \operatorname{argmin}_{a\in[K]} m_t(a)$ is again the action with the smallest predicted loss. In other words, we use a new biased estimator: $\widetilde{\ell}_t(a) = \widehat{\ell}_t(a) - m_t(a_t^*) = \frac{\ell_t(a)-m_t(a)}{p_t(a)}\mathbb{1}[a_t = a] + m_t(a) - m_t(a_t^*)$. Moreover, we apply the same action remapping technique using the mapping $\phi_t : [K] \to \mathcal{A}_t$ as in the adversarial case (Eq. (3)). To see why this is useful, note that the variance term now becomes

$$\begin{aligned}
\mathbb{V}_t\left[\widetilde{\ell}_t(\phi_t(\pi(x_t)))\right] &\le \mathbb{E}_{x_t,m_t,\ell_t,a_t}\left[\left(\widehat{\ell}_t(\phi_t(\pi(x_t))) - m_t(a_t^*)\right)^2\right] \\
&\le 2\mathbb{E}_{x_t,m_t,\ell_t,a_t}\left[\frac{(\ell_t(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))))^2}{p_t^2(\phi_t(\pi(x_t)))}\mathbb{1}\left[a_t = \phi_t(\pi(x_t))\right]\right] \\
&\quad + 2\mathbb{E}_{x_t,m_t}\left[\left(m_t(\phi_t(\pi(x_t))) - m_t(a_t^*)\right)^2\right] \qquad \text{(using } (a+b)^2 \le 2a^2+2b^2\text{)} \\
&\le \frac{2K}{\mu}\mathbb{E}_{x_t,m_t,\ell_t}\left[(\ell_t(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))))^2\right] + 2\sigma^2 \le \frac{2K\mathcal{E}}{\mu T} + 2\sigma^2, \quad (4)
\end{aligned}$$

which improves over the variance term $\mathbb{V}_t[\widehat{\ell}_t(\pi(x_t))]$ if σ is small. Also note that with action remapping, we only explore actions in $\mathcal{A}_t$, and thus by the exact same arguments as in the adversarial case, the exploration term also becomes $\mu T\sigma + 2\mu\sqrt{\mathcal{E}T}$, again better than the naive approach as long as σ is small. Therefore, remapping improves both the variance and the exploration term.

It remains to analyze the bias from both the shifted estimator and the remapping. The former in fact does not introduce any bias for the regret since the shift $m_t(a_t^*)$ is the same for all actions. The latter introduces total bias $\mathcal{O}(\mathcal{E}/\sigma)$, again by the same analysis as in the adversarial case. With these modifications, we achieve $\widetilde{\mathcal{O}}((\mathcal{E}T)^{\frac{1}{3}})$ regret already (even with the presence of the lower-order term). This is summarized in the following theorem (see Appendix D.1 for the proof).

Theorem 6 ϵ-GREEDY.AR *(Algorithm 2 Option I) ensures* $\text{Reg} = \widetilde{\mathcal{O}}\big(\sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{d}{\mu} + \mu T\sigma + \mu\sqrt{\mathcal{E}T} + \frac{\mathcal{E}}{\sigma}\big)$. *For* $\mathcal{E} \le \sqrt{T}$, *picking* $\mu = \min\big\{\big(\frac{d^2}{\mathcal{E}T}\big)^{\frac{1}{3}}, 1\big\}$ *and* $\sigma = \big(\frac{\mathcal{E}^2}{dT}\big)^{\frac{1}{3}}$ *gives* $\text{Reg} = \mathcal{O}\big((d\mathcal{E}T)^{\frac{1}{3}} + \sqrt{d\mathcal{E}} + d\big)$.

Removing the lower-order term via Catoni's estimator. To further improve the regret bound to $\widetilde{\mathcal{O}}(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$, we need to improve the lower-order term as well. Fortunately, it turns out that this

Algorithm 2 ϵ-Greedy with Action Remapping (and Catoni's estimator)

Parameters: threshold $\sigma > 0$, exploration probability $\mu \in [0, 1]$.

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

Receive x_t and m_t. Define a_t^*, $\mathcal{A}_t$ and ϕ_t as in Eq. (3).

Find π_t $\begin{cases} = \operatorname{argmin}_{\pi \in \Pi} \text{ ERM}\left(\{x_s, \widetilde{\ell}_s \circ \phi_s\}_{s<t}\right), & \text{(Option I, termed } \epsilon\text{-GREEDY.AR)} \\ \approx \operatorname{argmin}_{\pi \in \Pi} \text{ Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s<t}\right) & \\ \quad \text{using Algorithm 3 with } \alpha = \sqrt{\frac{2\ln(TN)}{(\sigma^2 t + K\mathcal{E}/\mu)}}. & \text{(Option II, termed } \epsilon\text{-GREEDY.ARC)} \end{cases}$

Calculate $p_t \in \Delta_K$: $\quad p_t(a) = (1-\mu)\mathbb{1}[a = \phi_t(\pi_t(x_t))] + \frac{\mu}{|\mathcal{A}_t|}\mathbb{1}[a \in \mathcal{A}_t]$.

Sample $a_t \sim p_t$ and receive $\ell_t(a_t)$.

Construct estimator: $\quad \widetilde{\ell}_t(a) = \frac{\ell_t(a) - m_t(a)}{p_t(a)}\mathbb{1}[a_t = a] + m_t(a) - m_t(a_t^*)$ for all $a \in \mathcal{A}_t$.

Algorithm 3 Finding the Policy with the Smallest Catoni's Mean

Input: context x_s, loss estimator $\widetilde{\ell}_s$, remapping function ϕ_s, for $s = 1, \ldots, t-1$, and parameter α.

Define: $\psi(y) = \begin{cases} \ln(1 + y + y^2/2), & \text{if } y \geq 0, \\ -\ln(1 - y + y^2/2), & \text{else.} \end{cases}$ **Initialize**: $z_{\text{right}} = \frac{K}{\mu} + 1$, $z_{\text{left}} = -z_{\text{right}}$.

while $z_{\text{right}} - z_{\text{left}} \geq 1/T$ **do**

Let $z_{\text{mid}} = (z_{\text{left}} + z_{\text{right}})/2$.

Construct $c_s \in \mathbb{R}^K$ for all $s < t$ such that $c_s(a) = \psi\left(\alpha\left(\widetilde{\ell}_s(\phi_s(a)) - z_{\text{mid}}\right)\right)$.

Invoke oracle $\pi = \text{ERM}\left(\{x_s, c_s\}_{s<t}\right)$.

if $\sum_{s<t} c_s(\pi(x_s)) \geq 0$ **then** $z_{\text{left}} = z_{\text{mid}}$, **else** $z_{\text{right}} = z_{\text{mid}}$.

Construct $c_s \in \mathbb{R}^K$ for all $s < t$ such that $c_s(a) = \psi\left(\alpha\left(\widetilde{\ell}_s(\phi_s(a)) - z_{\text{right}}\right)\right)$.

Return $\pi_t = \text{ERM}\left(\{x_s, c_s\}_{s<t}\right)$.

lower-order term can be completely removed using *robust mean estimators* for heavy-tailed distributions, such as median of means, trimmed-mean, and Catoni's estimator (see the survey (Lugosi and Mendelson, 2019)). In particular, we use Catoni's estimator, as we show that it can be implemented efficiently via the ERM oracle.

More specifically, instead of following the policy with the smallest average estimated loss, we follow the policy with the smallest Catoni's mean: $\operatorname{argmin}_{\pi \in \Pi} \text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s<t}\right)$ where $\text{Catoni}_\alpha(y_1, \ldots, y_n)$ is the root of the function $f(z) = \sum_{j=1}^n \psi(\alpha(y_j - z))$ for some increasing function ψ (defined in Algorithm 3) and coefficient $\alpha > 0$. Generalizing the proof of Theorem 5 in Lugosi and Mendelson (2019) for i.i.d. random variables to a martingale sequence, we obtain a concentration result without the lower-order term (see Lemma 13 in Appendix A). Furthermore, we prove that a close approximation of this policy can be found efficiently via a binary search invoking $\mathcal{O}(\ln(TK/\mu))$ calls of the ERM oracle, detailed in Algorithm 3.

Lemma 7 *Algorithm 3 invokes the ERM oracle at most $\mathcal{O}(\ln(TK/\mu))$ times and returns a policy π_t such that:* $\text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi_t(x_s)))\right\}_{s<t}\right) \leq \min_{\pi \in \Pi} \text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s<t}\right) + \frac{1}{T}$.

The proof is based on the monotonicity of ψ (Appendix D.1). We remark that this result might be of independent interest and useful for developing oracle-efficient algorithms for other problems.

Combining the two key techniques above, we improve all the three terms and prove the following theorem (see Algorithm 2 for the pseudocode and Appendix D.1 for the complete proof).

Theorem 8 ϵ-GREEDY.ARC *(Algorithm 2 Option II) ensures* $\mathrm{Reg} = \widetilde{\mathcal{O}}\big(\sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \mu T\sigma + \mu\sqrt{\mathcal{E}T} + \frac{\mathcal{E}}{\sigma}\big)$. *Picking* $\mu = \min\big\{\sqrt{\frac{d}{T}}, 1\big\}$ *and* $\sigma = \sqrt{\mathcal{E}}(dT)^{\frac{-1}{4}}$ *gives* $\mathrm{Reg} = \mathcal{O}\big(\sqrt{\mathcal{E}}(dT)^{\frac{1}{4}} + \sqrt{d\mathcal{E}}\big)$.

Similarly, this requires setting σ in terms of $\mathcal{E}$, and when $\mathcal{E} = \Omega(\sqrt{T})$, one could switch to the optimal algorithm (Agarwal et al., 2014) and achieve $\mathcal{O}(\sqrt{T})$ regret. Therefore, our bound again matches the lower bound in Theorem 1. In fact, it also enjoys a better dependence on d compared to the adversarial case (Theorem 4).

4.1. Adaptive version with unknown $\mathcal{E}$

When $\mathcal{E}$ is unknown, the same bound is not achievable (Theorem 2) and we relax our goal to achieve a bound that is robust and always sublinear, and at the same time improves over $\mathcal{O}(\sqrt{T})$ when $\mathcal{E}$ is relatively small. We achieve this goal with a different approach compared to the adversarial case, by exploiting the stochasticity of the environment so we can directly estimate $\mathcal{E}$ in the early rounds. Specifically, we spend the first B rounds for pure exploration (i.e., pick a_t uniformly at random) to collect a set of data $\{a_t, \ell_t(a_t), m_t(a_t)\}_{t\le B}$. Then we design a novel *underestimator* $\widehat{\mathcal{E}}$ defined as

$$\widehat{\mathcal{E}} = T\sum_{i=0}^{\lceil \log_2 T\rceil}\left[\widehat{\alpha}_i - \frac{30\log T}{B}\right]_+ 2^{-2i}, \quad \text{where } \widehat{\alpha}_i = \frac{1}{B}\sum_{t=1}^{B}\mathbb{1}\left[|\ell_t(a_t) - m_t(a_t)| \in \left(2^{-i-1}, 2^{-i}\right]\right],$$

and $[\cdot]_+ = \max\{\cdot, 0\}$. For the rest of the game, we simply run Algorithm 2 with Option I, $\sigma = \sqrt{\widehat{\mathcal{E}}}(dT)^{-\frac{1}{3}}$, and $\mu = \min\left\{d^{2/3}T^{-1/3}, 1\right\}$. Note that here we use the simpler Option I, as it can be verified that even without the lower-order term the regret would still be the same in this case (moreover, Option II requires setting α in terms of $\mathcal{E}$).

The idea behind this estimator is as follows. First, $\widehat{\alpha}_i$ is clearly an unbiased estimator of $\alpha_i = \frac{1}{K}\sum_{a=1}^{K}\Pr\left[|\ell_t(a) - m_t(a)| \in (2^{-i-1}, 2^{-i}]\right]$, and is thus basically estimating the histogram of the distribution of $\ell_t - m_t$ with bins of exponentially increasing size. Therefore, $\sum_i \widehat{\alpha}_i 2^{-2i}$ is an approximation of $\frac{1}{K}\mathbb{E}\left[\|\ell_t - m_t\|_2^2\right]$. In the definition of $\widehat{\mathcal{E}}$, we subtract a deviation term $30\log T/B$ from $\widehat{\alpha}_i$ to make sure that $\widehat{\mathcal{E}}$ is an underestimator. It turns out that both the idea of underestimating and that of estimating the histogram are critical for the analysis, allowing us to prove the following guarantee (see Appendix D.2 for the complete pseudocode and proof). Note that this is the same bound as in the adversarial case (Theorem 5), ignoring the dependence on d.

Theorem 9 ϵ-GREEDY.VAR *(Algorithm 6 in Appendix D.2) guarantees* $\mathrm{Reg} = \widetilde{\mathcal{O}}\Big(B + K\sqrt{\frac{\mathcal{E}T}{B}} + K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}}\Big)$. *Setting* $B = T^{\frac{1}{3}}$ *gives* $\mathrm{Reg} = \widetilde{\mathcal{O}}(K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}})$.

5. Algorithms for Multiple Predictors

Finally, we extend our setting and consider learning with multiple predictors. That is, the learner receives M predictors $m_t^1, \ldots, m_t^M$ before choosing a_t at each round. Recall that in the adversarial

setting, the total error of the best predictor is measured by $\mathcal{E}^* = \min_{i\in[M]} \sum_{t=1}^{T} \|\ell_t - m_t^i\|_\infty^2$, while in the stochastic setting, it is measured by $\mathcal{E}^* = T \min_{i\in[M]} \mathbb{E}_{(x,\ell,m^{1:M})\sim\mathcal{D}} \left[\|\ell - m^i\|_\infty^2\right]$.

Our goal is to improve over $\mathcal{O}(\sqrt{T})$ regret whenever $\mathcal{E}^*$ and M are relatively small, assuming $\mathcal{E}^*$ is known for simplicity. In both cases, we deploy a natural idea: maintain an active set of predictors $\mathcal{P}_t \subseteq [M]$ (starting from $\mathcal{P}_1 = [M]$), and eliminate a predictor from this set whenever its observed total error exceeds $\mathcal{E}^*$. We define $m_t(a) = \min_{i\in\mathcal{P}_t} m_t^i(a)$ to be the smallest predicted loss for action a among the active predictors, and follow similar ideas of the single predictor case with m_t serving the role of the single predictor, which can be seen as a form of optimism. In addition to this basic idea, however, extra new techniques are required for the two settings as described below.

Adversarial Environments. The only extra difference compared to Algorithm 1 is in the construction of $\mathcal{A}_t$, the set of actions with predicted loss not larger than that of a baseline by σ. In Algorithm 1, the baseline is simply the action with the smallest predicted loss $a_t^* = \operatorname{argmin}_a m_t(a)$. However, with multiple predictors, we propose to (essentially) use a_t, the action *to be chosen* by the algorithm, as the baseline. Before explaining why this is a good idea, we first point out that this can indeed be efficiently implemented, even though the scheme appears self-referential as a_t itself depends on $\mathcal{A}_t$. Indeed, this resembles the idea of sleeping experts (Freund et al., 1997), if we treat actions outside $\mathcal{A}_t$ as asleep experts. For implementation details, see Algorithm 5 in Appendix C.3.

The ideas of the analysis are as follows. Using a_t as the baseline gives that the exploration overhead and the bias introduced by remapping are both in terms of $\sum_{t=1}^{T}(\ell_t(a_t) - m_t(a_t))^2$, which is of order $\mathcal{O}(M\mathcal{E}^*)$ because each predictor can contribute at most $\mathcal{E}^* + 1$ before being eliminated (Lemmas 17, 18 and 19). Second, note that we only need to refer to Eq. (2) (instead of the standard analysis) when $\ell_t(a_t) \le m_t(a_t)$, in which case we have $(\ell_t(a_t) - m_t(a_t))^2/p_t(a_t) \le (\ell_t(a_t) - m_t^{i^*}(a_t))^2/p_t(a_t)$ by the definition of m_t (i^* is the best predictor). This allows us to relate the expectation of this term to $\mathcal{E}^*$ as well. Put together, we prove the following theorem.

Theorem 10 *With the optimal parameters and* $M' = \max\{M, K\}$, EXP4.MOAR *(Algorithm 5 in Appendix C.3) ensures* $\mathrm{Reg} = \mathcal{O}\left(\sqrt{M'\mathcal{E}^*}(dT)^{\frac{1}{4}} + \sqrt{dM'\mathcal{E}^*} + d\right)$.

Stochastic Environments. There are two extra differences compared to Algorithm 2 in this case. First, at the beginning of each round, we check if all predictors are consistent to some extent. If not, that is, if there exist two predictors who disagree with each other by a large amount on some action, then we simply choose this action deterministically, since this guarantees to reveal which predictor makes a large error for this round. Second, in the case when all predictors are consistent, instead of doing ϵ-greedy as in Algorithm 2 (which we already show is optimal for single predictor), we find that we need to resort to the minimax optimal algorithm ILOVETOCONBANDITS (Agarwal et al., 2014) to better control the variance of the estimator. We develop a version of it with action remapping and Catoni's estimators. See Algorithm 7 for details. Combining everything, we prove:

Theorem 11 ILTCB.MARC *(Algorithm 7 in Appendix D.3) guarantees* $\mathrm{Reg} = \mathcal{O}\left(M^{\frac{2}{3}} d^{\frac{2}{5}} (\mathcal{E}^* T)^{\frac{1}{3}}\right)$.

As a final remark, we remind the reader of the lower bound $\Omega(\sqrt{\mathcal{E}^*}T^{\frac{1}{4}} + M)$ for $M \le \sqrt{T}$ given by Theorem 3. Our upper bound for the adversarial case has matching dependence on $\mathcal{E}^*$ and T, but not M, while our bound for the stochastic case is even looser. Closing the gap and generalizing the results to unknown $\mathcal{E}^*$ are two main future directions.

Acknowledgments

The authors would like to thank Akshay Krishnamurthy and Chicheng Zhang for introducing the idea of robust mean estimator. Part of this work was done when CYW was an intern at Microsoft Research. HL and CYW are supported by NSF Awards IIS-1755781 and IIS-1943607.

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 *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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, 2002.

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, 2011.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Proceedings of the 32nd Conference On Learning Theory*, 2019.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726, 2019.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, 2012.

Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *Conference on Learning Theory*, pages 210–227, 2013.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411, 2015.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Yoav Freund, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 334–343, 1997.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(Apr):1287–1311, 2011.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 128–141, 2016.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Conference on Learning Theory*, 2019.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems 21*, 2008.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Haipeng Luo. Lecture notes 21 of introduction to online learning. https://haipeng-luo.net/courses/CSCI699/lecture21.pdf, 2017.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26*, 2013b.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28*, 2015.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016a.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems 29*, 2016b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, 2018.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in neural information processing systems*, pages 3972–3980, 2016.

Kai Zheng, Haipeng Luo, Ilias Diakonikolas, and Liwei Wang. Equipping experts/bandits with long-term memory. In *Advances in Neural Information Processing Systems*, pages 5927–5937, 2019.

Appendix A. Concentration Inequalities

Lemma 12 (Freedman's inequality, *cf.* Theorem 1 of (Beygelzimer et al., 2011)) *Let $\mathcal{F}_0 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$, $|X_i| \le b$, and $\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}] \le V$ for some fixed $b \ge 0$ and $V \ge 0$. Then for any $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\sum_{i=1}^n X_i \le 2\sqrt{V_n \log(1/\delta)} + b\log(1/\delta).$$

Lemma 13 (Concentration inequality for Catoni's estimator) *Let $\mathcal{F}_0 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = \mu_i$ for some fixed μ_i, and $\sum_{i=1}^n \mathbb{E}[(X_i-\mu_i)^2|\mathcal{F}_{i-1}] \le V$ for some fixed V. Denote $\mu \triangleq \frac{1}{n}\sum_{i=1}^n \mu_i$ and let $\widehat{\mu}_{n,\alpha}$ be the Catoni's robust mean estimator of $X_1, \ldots, X_n$ with a fixed parameter $\alpha > 0$, that is, $\widehat{\mu}_{n,\alpha}$ is the unique root of the function*

$$f(z) = \sum_{i=1}^n \psi(\alpha(X_i - z))$$

where

$$\psi(y) = \begin{cases} \ln(1+y+y^2/2), & \text{if } y \ge 0, \\ -\ln(1-y+y^2/2), & \text{else.} \end{cases}$$

Then for any $\delta \in (0,1)$, as long as n is large enough such that $n \ge \alpha^2(V + \sum_{i=1}^n(\mu_i-\mu)^2) + 2\log(1/\delta)$, we have with probability at least $1-2\delta$,

$$|\widehat{\mu}_{n,\alpha} - \mu| \le \frac{\alpha(V + \sum_{i=1}^n(\mu_i-\mu)^2)}{n} + \frac{2\log(1/\delta)}{\alpha n}.$$

In particular, if $\mu_1 = \cdots = \mu_n = \mu$, we have[3]

$$|\widehat{\mu}_{n,\alpha} - \mu| \le \frac{\alpha V}{n} + \frac{2\log(1/\delta)}{\alpha n}.$$

3. In all our applications of this lemma, we have $\mu_1 = \cdots = \mu_n = \mu$.

Proof The proof generalizes that of (Lugosi and Mendelson, 2019, Theorem 5) for i.i.d. random variables, following similar ideas used in (Beygelzimer et al., 2011, Theorem 1). First, one can verify that $\psi(y) \leq \ln(1 + y + y^2/2)$ for all $y \in \mathbb{R}$. Therefore, for any fixed $z \in \mathbb{R}$ and any i, we have

$$\begin{aligned}
&\mathbb{E}_i\left[\exp\left(\psi(\alpha(X_i - z))\right)\right] && (\mathbb{E}_i[\cdot] \triangleq \mathbb{E}[\cdot \mid \mathcal{F}_{i-1}]) \\
&\leq \mathbb{E}_i\left[1 + \alpha(X_i - z) + \frac{\alpha^2(X_i - z)^2}{2}\right] \\
&= 1 + \alpha(\mu_i - z) + \frac{\alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2} \\
&\leq \exp\left(\alpha(\mu_i - z) + \frac{\alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2}\right). && (1 + y \leq e^y)
\end{aligned}$$

Define random variables $Z_0 = 1$, and for $i \geq 1$,

$$Z_i = Z_{i-1}\exp\left(\psi(\alpha(X_i - z))\right)\exp\left(-\left(\alpha(\mu_i - z) + \frac{\alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \alpha^2(\mu_i - z)^2}{2}\right)\right).$$

Then the last calculation shows $\mathbb{E}_i[Z_i] \leq Z_{i-1}$. Therefore, taking expectation over all random variables $X_1, \ldots, X_n$, we have

$$\mathbb{E}[Z_n] \leq \mathbb{E}[Z_{n-1}] \leq \cdots \leq \mathbb{E}[Z_0] = 1.$$

Further define

$$g(z) \triangleq n\alpha(\mu - z) + \frac{1}{2}\alpha^2\sum_{i=1}^n(\mu_i - z)^2 + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right)$$

and note that $f(z) \geq g(z)$ implies

$$\begin{aligned}
&\sum_{i=1}^n \psi(\alpha(X_i - z)) \geq n\alpha(\mu - z) + \frac{1}{2}\alpha^2\sum_{i=1}^n(\mu_i - z)^2 + \frac{1}{2}\alpha^2\sum_{i=1}^n\mathbb{E}_i\left[(X_i - \mu_i)^2\right] + \log\left(\frac{1}{\delta}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by the condition of } V\text{)} \\
&= \sum_{i=1}^n\left(\alpha(\mu_i - z) + \frac{\alpha^2(\mu_i - z)^2 + \alpha^2\mathbb{E}_i\left[(X_i - \mu_i)^2\right]}{2}\right) + \log\left(\frac{1}{\delta}\right),
\end{aligned}$$

which further implies $Z_n \geq 1/\delta$. By Markov's inequality, we then have $\Pr[f(z) \geq g(z)] \leq \Pr[Z_n \geq 1/\delta] \leq \Pr[Z_n \geq \mathbb{E}[Z_n]/\delta] \leq \delta$. Note further that we can rewrite $g(z)$ as

$$\begin{aligned}
g(z) &= n\alpha(\mu - z) + \frac{1}{2}\alpha^2(nz^2 - 2n\mu z + \sum_{i=1}^n\mu_i^2) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right) \\
&= n\alpha(\mu - z) + \frac{1}{2}\alpha^2(n(z - \mu)^2 - n\mu^2 + \sum_{i=1}^n\mu_i^2) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right)
\end{aligned}$$

$$= n\alpha(\mu - z) + \frac{1}{2}n\alpha^2(z-\mu)^2 + \frac{1}{2}\alpha^2\left(\sum_{i=1}^n \mu_i^2 - n\mu^2\right) + \frac{1}{2}\alpha^2 V + \log\left(\frac{1}{\delta}\right)$$

Now we pick z to be the smaller root z_0 of the quadratic function $g(z)$, that is,

$$z_0 = \mu + \frac{1}{\alpha}\left(1 - \sqrt{1 - \frac{\alpha^2(V + \sum_{i=1}^n(\mu_i - \mu)^2)}{n} - \frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right)$$

(which exists due to the condition on n). By the monotonicity of f and the fact $f(\widehat{\mu}_{n,\alpha}) = 0$ we then have

$$\Pr\left[\widehat{\mu}_{n,\alpha} \geq z_0\right] = \Pr\left[f(z_0) \geq 0\right] = \Pr\left[f(z_0) \geq g(z_0)\right] \leq \delta.$$

In other words, with probability at least $1-\delta$, we have

$$\begin{aligned}
\widehat{\mu}_{n,\alpha} - \mu &\leq \frac{1}{\alpha}\left(1 - \sqrt{1 - \frac{\alpha^2(V + \sum_{i=1}^n(\mu_i - \mu)^2)}{n} - \frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \\
&\leq \frac{1}{\alpha}\left(\frac{\alpha^2(V + \sum_{i=1}^n(\mu_i - \mu)^2)}{n} + \frac{2}{n}\log\left(\frac{1}{\delta}\right)\right) \qquad (1 - \sqrt{1-x} \leq x \text{ for } x \in [0,1]) \\
&= \frac{\alpha(V + \sum_{i=1}^n(\mu_i - \mu)^2)}{n} + \frac{2\log(1/\delta)}{\alpha n}.
\end{aligned}$$

Finally, via a symmetric argument one can show that $\mu - \widehat{\mu}_{n,\alpha} \leq \frac{\alpha(V+\sum_{i=1}^n(\mu_i-\mu)^2)}{n} + \frac{2\log(1/\delta)}{\alpha n}$ holds with probability at least $1-\delta$ as well. Applying a union bound then finishes the proof. ■

Appendix B. Proofs for Lower Bounds

In this section, we provide proofs for all the lower bounds discussed in Section 2. The techniques we use are reminiscent of those in several previous works on bandit problems that prove lower bounds for adaptive regret (Daniely et al., 2015, Theorem 3), switching regret (Wei et al., 2016, Theorem 4.1), and regret bounds in terms of the sparsity of the losses (Zheng et al., 2019, Theorem 6). While their constructions are for adversarial environments, ours are for the i.i.d. case (which is stronger for lower bounds). To make the proofs concise, we assume that numbers such as $\sqrt{T/K}$ are integers without rounding them.

Proof for Theorem 1. We first prove that for any algorithm, any $K \geq 2$, any $T \geq 8 \times 10^4$, and any value $V \in [0, T]$, there exists a stochastic environment with $\mathcal{E} \leq V$ and $N = (K-1)\sqrt{T/K} + 1$ such that $\mathrm{Reg} = \widetilde{\Omega}\big(\min\big\{\sqrt{V}(KT)^{\frac{1}{4}}, \sqrt{KT}\big\}\big)$. The construction is as follows. There are $\sqrt{T/K}$ possible context-predictor-loss tuples $\{(x^{(i)}, m^{(i)}, \ell^{(i)})\}_{i=1}^{\sqrt{T/K}}$, and in each round, (x_t, m_t, ℓ_t) is uniformly randomly drawn from this set. The policy set Π contains $(K-1)\sqrt{T/K}+1$ policies such that: there is a policy $\pi^{(0)}$ that always chooses action 1 given any context; other policies are indexed by $(i,k) \in [\sqrt{T/K}] \times \{2, \ldots, K\}$ such that

$$\pi^{(i,k)}(x) = \begin{cases} k & \text{if } x = x^{(i)}, \\ 1 & \text{otherwise.} \end{cases}$$

Now first consider an environment with $m^{(i)} = \ell^{(i)} = (\frac{1}{2}, \frac{1}{2}+\sigma, \ldots, \frac{1}{2}+\sigma)$ for all i, where $\sigma = \min\left\{\frac{1}{2}, \frac{\sqrt{V}}{2(KT)^{1/4}}\right\}$. Note that $\mathcal{E} = 0 \le V$. Under this environment and the given algorithm, if for all $(i,k) \in [\sqrt{T/K}] \times \{2,\ldots,K\}$, the expected total number of times where $(x_t, a_t) = (x^{(i)}, k)$ is larger than $\frac{1}{2}$, then the algorithm's regret against $\pi^{(0)}$ is

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(\pi^{(0)}(x_t))\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{\sqrt{T/K}}\sum_{k=2}^{K} \mathbb{1}[(x_t,a_t) = (x^{(i)},k)]\left(\ell_t(k) - \ell_t(1)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{\sqrt{T/K}}\sum_{k=2}^{K} \mathbb{1}[(x_t,a_t) = (x^{(i)},k)]\sigma\right] \\
&\ge \sqrt{\frac{T}{K}} \times (K-1) \times \frac{1}{2} \times \sigma \ge \frac{1}{4}\sqrt{KT}\sigma.
\end{aligned}$$

On the other hand, if there exists a pair $(i^*, k^*) \in [\sqrt{T/K}] \times \{2,\ldots,K\}$ such that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[(x_t,a_t) = (x^{(i^*)},k^*)]\right] \le \frac{1}{2},$$

then by Markov's inequality,

$$\begin{aligned}
\Pr\left[\sum_{t=1}^{T} \mathbb{1}[(x_t,a_t) = (x^{(i^*)},k^*)] = 0\right] &= \Pr\left[\sum_{t=1}^{T} \mathbb{1}[(x_t,a_t) = (x^{(i^*)},k^*)] < 1\right] \\
&= 1 - \Pr\left[\sum_{t=1}^{T} \mathbb{1}[(x_t,a_t) = (x^{(i^*)},k^*)] \ge 1\right] \ge \frac{1}{2}.
\end{aligned}$$

That is, with probability at least $\frac{1}{2}$, the learner never chooses action k^* when she sees context $x^{(i^*)}$. In this case, consider another environment where all $m^{(i)}$ and $\ell^{(i)}$ remain the same except that $\ell^{(i^*)}$ is changed to $(\frac{1}{2}, \frac{1}{2}+\sigma, \ldots, \frac{1}{2}-\sigma, \ldots, \frac{1}{2}+\sigma)$, where $\frac{1}{2}-\sigma$ appears in the k^*-th coordinate. Note that in this new environment we again have $\mathcal{E} = T\mathbb{E}_{(x,\ell,m)}\left[\|\ell - m\|_\infty^2\right] = \sqrt{TK} \times 4\sigma^2 \le V$. Moreover, with probability at least $\frac{1}{2}$ the learner never realizes the change of the environment and behaves exactly the same, since the only way to distinguish the two environments is to pick k^* under context $x^{(i^*)}$.

It remains to calculate the regret of the learner under this new environment. First, by Freedman's inequality (Lemma 12), we have with probability at least $1 - \frac{1}{T}$,

$$\sum_{t=1}^{T} \mathbb{1}[x_t = x^{(i^*)}] \ge \sqrt{KT} - 2\sqrt{\sqrt{KT}\log T} - \log T \ge \frac{\sqrt{KT}}{3} \tag{5}$$

where the last step uses the condition $K \ge 2$ and $T \ge 8 \times 10^4$. Define events

$$E_1 = \left\{\sum_{t=1}^{T} \mathbb{1}[(x_t,a_t) = (x^{(i^*)},k^*)] = 0\right\}, E_2 = \left\{\sum_{t=1}^{T} \mathbb{1}[x_t = x^{(i^*)}] \ge \frac{\sqrt{KT}}{3}\right\},$$

and use $\mathbb{E}', \Pr'$ to denote the expectation and probability under the new environment. Now we lower bound the regret against $\pi^{(i^*,k^*)}$ in this environment as

$$\begin{aligned}
&\mathbb{E}'\left[\sum_{t=1}^{T}\sum_{i=1}^{\sqrt{T/K}} \mathbb{1}[x_t = x^{(i)}]\left(\ell_t(a_t) - \ell_t(\pi^{(i^*,k^*)}(x^{(i)}))\right)\right] \\
&\geq \Pr'[E_1 \cap E_2] \times \mathbb{E}'\left[\sum_{t=1}^{T} \mathbb{1}[x_t = x^{(i^*)}]\left(\ell_t(a_t) - \ell_t(\pi^{(i^*,k^*)}(x^{(i^*)}))\right) \;\middle|\; E_1, E_2\right] \\
&= \Pr'[E_1 \cap E_2] \times \mathbb{E}'\left[\sum_{t=1}^{T} \mathbb{1}[x_t = x^{(i^*)}]\sigma \;\middle|\; E_1, E_2\right] \\
&\geq \left(\frac{1}{2} - \frac{1}{T}\right) \times \frac{\sqrt{KT}\sigma}{3} \geq \frac{\sqrt{KT}\sigma}{12}.
\end{aligned}$$

To summarize, in at least one of these two environments, the learner's regret is

$$\Omega(\sqrt{KT}\sigma) = \widetilde{\Omega}\big(\min\big\{\sqrt{V}(KT)^{\frac{1}{4}}, \sqrt{KT}\big\}\big),$$

finishing the lower bound proof for stochastic environments. For adversarial environments, the only change is to let each tuple $(x^{(i)}, m^{(i)}, \ell^{(i)})$ appear for exactly $\sqrt{T/K}$ times, so that $\mathcal{E} \leq V$ still holds in these two constructions under the slightly different definition for $\mathcal{E}$ (which is $\sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$). It is clear that the same lower bound holds. ■

Proof for Theorem 2. The idea of the proof is similar to that of Theorem 1. Assume there is an algorithm that guarantees for some $R = o(\sqrt{KT})$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t) - \sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] \leq R,$$

whenever $\mathcal{E} = 0$. Below we show that there is an environment with $\mathcal{E} = o(\sqrt{KT})$ where the algorithm suffers $\omega(\sqrt{KT})$ regret.

The construction is as follows. First, let C be a universal constant such that $R + K \leq \sqrt{CKT}$. Further define $\rho = \frac{R+K}{\sqrt{CKT}} \leq 1$, $\sigma = \frac{1}{2}\rho^{\frac{2}{5}}$, and $L_0 = \rho^{-\frac{3}{5}}$. There are $\sqrt{64CT/K}$ context-predictor-loss tuples $\{(x^{(i)}, m^{(i)}, \ell^{(i)})\}_{i=1}^{\sqrt{64CT/K}}$, and in each round, (x_t, m_t, ℓ_t) is uniformly randomly drawn from this set. The policy set contains $N = \Theta(T)$ policies such that: there is a policy $\pi^{(0)}$ that always chooses action 1 given any contexts; other policies are indexed by $(i, j, k) \in [\sqrt{64CT/K}] \times [\sqrt{64CT/K}] \times \{2, \ldots, K\}$ with $i \leq j$ such that

$$\pi^{(i,j,k)}(x) = \begin{cases} k, & \text{if } x \in \{x^{(i)}, x^{(i+1)}, \ldots, x^{(j)}\}. \\ 1, & \text{else.} \end{cases}$$

We first consider the algorithm's behavior under the environment with $m^{(i)} = \ell^{(i)} = (\frac{1}{2}, \frac{1}{2} + \sigma, \ldots, \frac{1}{2} + \sigma)$ for all i. In this environment, since $\mathcal{E} = 0$, the algorithm must guarantee that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}[a_t \neq 1]\right] \leq \frac{R}{\sigma}. \tag{6}$$

This is because every time $a_t \neq 1$, the learner incurs regret σ against $\pi^{(0)}$. Next we prove the following fact: there exists $i, j \in [\sqrt{64CT/K}]$ and $k \in \{2, \ldots, K\}$ such that $j - i + 1 = L_0$ and $\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s=i}^{j} \mathbb{1}[x_t = x^{(s)}, a_t = k]\right] \leq \frac{1}{2}$. We prove it by contradiction. Assume that for all (i, j) with $j = i - 1 + L_0$ and all $k \in \{2, \ldots, K\}$, $\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s=i}^{j} \mathbb{1}[x_t = x^{(s)}, a_t = k]\right] \geq \frac{1}{2}$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a_t \neq 1]\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{s=1}^{\sqrt{64CT/K}} \sum_{k=2}^{K} \mathbb{1}[x_t = x^{(s)}, a_t = k]\right] \\
&\geq \frac{\sqrt{64CT/K}}{L_0} \times \frac{1}{2} \times (K-1) \\
&\geq \frac{\sqrt{64CKT}}{4L_0} \geq \frac{2R}{\rho L_0} = \frac{R}{\sigma},
\end{aligned}$$

which leads to a contradiction (here, we also use the fact $1 \leq L_0 = \rho^{-\frac{3}{5}} \leq \rho^{-1} \leq \sqrt{\frac{64CT}{K}}$). Therefore, we have shown that there exist (i^*, j^*) with $j^* - i^* + 1 = L_0$ and $k^* \in \{2, \ldots, K\}$ such that $\mathbb{E}\left[\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}, a_t = k^*]\right] \leq \frac{1}{2}$. By Markov's inequality, we thus have

$$\Pr\left[\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}, a_t = k^*] = 0\right] = \Pr\left[\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}, a_t = k^*] < 1\right] \geq \frac{1}{2}.$$

Now we consider another environment, which is the same as the one above, except that for all $s = i^*, i^* + 1, \ldots, j^*$, the k^*-th coordinate of $\ell^{(s)}$ is changed from $\frac{1}{2} + \sigma$ to $\frac{1}{2} - \sigma$. Note that in this environment,

$$\mathcal{E} = \mathcal{O}\left(L_0 \sqrt{KT}\sigma^2\right) = \mathcal{O}(\sqrt{KT}\rho^{\frac{1}{5}}) = o(\sqrt{KT}),$$

where we use the fact $\rho = o(1)$. Moreover, with probability at least $1/2$, the algorithm never realizes the change and behaves exactly the same, since the only way to distinguish the two environments is to pick k^* under one of the contexts $x^{(i^*)}, x^{(i^*+1)}, \ldots, x^{(j^*)}$.

It remains to calculate the regret of the learner under this environment. Define events

$$E_1 = \left\{\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[(x_t, a_t) = (x^{(s)}, k^*)] = 0\right\},$$

and

$$E_2 = \left\{\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}] \geq \frac{(j^* - i^* + 1)\sqrt{KT/64C}}{3}\right\}.$$

Note that in expectation, each context appears $\frac{T}{\sqrt{64CT/K}} = \sqrt{KT/64C}$ times. By Freedman's inequality (Lemma 12), with probability at least $1 - \frac{1}{T}$,

$$\sum_{t=1}^{T} \sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}] \geq (j^* - i^* + 1)\sqrt{\frac{KT}{64C}} - 2\sqrt{(j^* - i^* + 1)\sqrt{\frac{KT}{64C}} \log T} - \log T$$

$$\geq \frac{(j^* - i^* + 1)\sqrt{\frac{KT}{64C}}}{3}.$$

when $K \geq 2$ and $T > 6 \times 10^6 C$. That is, $\Pr[E_2] \geq 1 - 1/T$. Therefore, the expected regret against $\pi^{(i^*,j^*,k^*)}$ in this new environment is lower bounded by

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=i^*}^{j^*} \mathbb{1}[x_t = x^{(s)}]\left(\ell_t(a_t) - \ell_t(\pi^{(i^*,j^*,k^*)}(x^{(s)}))\right)\right] \\
&\geq \Pr[E_1 \cap E_2] \times \frac{(j^* - i^* + 1)\sqrt{KT/64C}}{3}\sigma \\
&\geq \left(\frac{1}{2} - \frac{1}{T}\right) \times \frac{\sqrt{KT/64C}}{3}\sigma L_0 = \Omega\left(\sqrt{\frac{KT}{C}}\rho^{-\frac{1}{5}}\right) = \omega(\sqrt{KT}).
\end{aligned}$$

This finishes the proof. ■

Proof for Theorem 3. When $\sqrt{V^*}(KT)^{\frac{1}{4}} \geq M$, we only need to prove a lower bound of $\widetilde{\Omega}(\sqrt{V^*}(KT)^{\frac{1}{4}})$, which is shown by Theorem 1 already. When $\sqrt{V^*}(KT)^{\frac{1}{4}} \leq M \leq \sqrt{T}$, we construct an stochastic environment below with $\mathcal{E}^* = 0$, $N = M$, and $K = 2$, where the regret of the algorithm is $\Omega(M)$.

The construction is as follows (and is again similar to those in the proofs of Theorems 1 and 2). There are $M - 1$ different context-predictor-loss tuples $\{x^{(i)}, m^{(0,i)}, \ldots, m^{(M-1,i)}, \ell^{(i)}\}_{i=1}^{M-1}$, and in every round, $(x_t, m_t^0, \ldots, m_t^{M-1}, \ell_t)$ is uniformly randomly sampled from this set. The policy set contains $N = M$ policies $\pi^{(0)}, \ldots, \pi^{(M-1)}$ such that: $\pi^{(0)}$ always chooses action 1 given any contexts; for $i \in [M-1]$, $\pi^{(i)}(x) = 2$ if $x = x^{(i)}$, and otherwise $\pi^{(i)}(x) = 1$.

Now consider an environment where $\ell^{(i)} = m^{(0,i)} = (\frac{1}{2}, 1)$ for all $i \in [M-1]$, $m^{(j,i)} = (\frac{1}{2}, 1)$ for all $i, j \in [M-1]$ with $i \neq j$, and $m^{(i,i)} = (\frac{1}{2}, 0)$ for all $i \in [M-1]$. Clearly, the predictor m^0 is a perfect predictor in this environment and thus $\mathcal{E}^* = 0$.

In this environment, if the expected number of times the learner chooses action 2 is larger than $\frac{M-1}{2}$, then she already suffers an expected regret of $\frac{M-1}{2} \times \frac{1}{2}$ compared to policy $\pi^{(0)}$, which always chooses action 1. On the other hand, if the expected number of times the learner chooses action 2 is smaller than $\frac{M-1}{2}$, then there exists an $i^* \in [M-1]$ such that the expected number of times the learner chooses action 2 upon seeing $x^{(i^*)}$ is less than $\frac{1}{2}$. By Markov's inequality, $\sum_{t=1}^{T} \mathbb{1}[(x_t, a_t) = (x^{(i^*)}, 2)] = 0$ holds with probability at least $\frac{1}{2}$. That is, with probability at least $\frac{1}{2}$, the learner never picks action 2 when the context is $x^{(i^*)}$.

Now consider a different environment where the only difference is that the $\ell^{(i^*)}(2)$ is changed from 1 to 0. With probability at least $\frac{1}{2}$, the learner does not realizes the change and behaves exactly the same. The expected regret compared to policy $\pi^{(i^*)}$ is thus $\Omega\left(\frac{T}{M-1} \times \frac{1}{2}\right)$ in this new environment. Moreover, notice that in this new environment, $\mathcal{E}^* = 0$ still holds because m^{i^*} now becomes the perfect predictor.

To sum up, we have shown that when there are $M > 1$ predictors, even if $\mathcal{E}^* = 0$, the learner has to suffer $\Omega\left(\min\left\{M - 1, \frac{T}{M-1}\right\}\right) = \Omega(M)$ regret. ■

Appendix C. Omitted Details for Adversarial Environments

In this section, we provide omitted details for the adversarial case, including the proof of Theorem 4 on the guarantee of Algorithm 1 for the case with single predictor and known $\mathcal{E}$ (Section C.1), the adaptive version of Algorithm 1 and its analysis when $\mathcal{E}$ is unknown (Section C.2), and the algorithm and analysis for multiple predictors (Section C.3).

C.1. Proof of Theorem 4

We first prove a lemma showing a somewhat non-conventional analysis of the optimistic EXP4 update. We denote the KL divergence of two distributions Q and P by $D(Q,P)=\sum_{\pi\in\Pi}Q(\pi)\ln\frac{Q(\pi)}{P(\pi)}$.

Lemma 14 *For any $\eta>0$, $\mathcal{M}_t,\mathcal{L}_t\in\mathbb{R}^N$, and distribution $Q'_t\in\Delta_\Pi$, define two distributions $Q_t,Q'_{t+1}\in\Delta_\Pi$ such that*

$$\begin{aligned} Q_t(\pi) &\propto Q'_t(\pi)\exp\left(-\eta\mathcal{M}_t(\pi)\right), \\ Q'_{t+1}(\pi) &\propto Q'_t(\pi)\exp\left(-\eta\mathcal{L}_t(\pi)\right). \end{aligned} \tag{7}$$

Then there exists $\xi_t\in\Delta_\Pi$ such that for any $Q^*\in\Delta_\Pi$, we have

$$\langle Q_t-Q^*,\mathcal{L}_t\rangle\le\frac{D(Q^*,Q'_t)-D(Q^*,Q'_{t+1})}{\eta}+2\eta\sum_{\pi\in\Pi}\xi_t(\pi)\left(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)\right)^2. \tag{8}$$

Moreover, if $\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)\ge-\frac{1}{\eta}$ holds for all π, then we have for any $Q^*\in\Delta_\Pi$,

$$\langle Q_t-Q^*,\mathcal{L}_t\rangle\le\frac{D(Q^*,Q'_t)-D(Q^*,Q'_{t+1})}{\eta}+\eta\sum_{\pi\in\Pi}Q_t(\pi)\left(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)\right)^2. \tag{9}$$

Proof First, we rewrite the updates in the standard optimistic online mirror descent framework: $Q_t=\operatorname{argmin}_{Q\in\Delta_\Pi}F_t(Q)$ and $Q'_{t+1}=\operatorname{argmin}_{Q\in\Delta_\Pi}F'_t(Q)$ where

$$\begin{aligned} F_t(Q) &= \eta\langle Q,\mathcal{M}_t\rangle+D(Q,Q'_t), \\ F'_t(Q) &= \eta\langle Q,\mathcal{L}_t\rangle+D(Q,Q'_t). \end{aligned} \tag{10}$$

Applying Lemma 6 of (Wei and Luo, 2018) shows

$$\langle Q_t-Q^*,\mathcal{L}_t\rangle\le\frac{D(Q^*,Q'_t)-D(Q^*,Q'_{t+1})}{\eta}+\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle-\frac{1}{\eta}D(Q'_{t+1},Q_t).$$

Next, we prove Eq. (8). By Taylor expansion, there exists some convex combination of Q_t and Q_{t+1}, denoted by ξ_t, such that

$$\begin{aligned} F'_t(Q_t)-F'_t(Q'_{t+1}) &= \nabla F'_t(Q'_{t+1})(Q_t-Q'_{t+1})+\frac{1}{2}(Q_t-Q'_{t+1})^\top\nabla^2F'_t(\xi_t)(Q_t-Q'_{t+1}) \\ &= \nabla F'_t(Q'_{t+1})(Q_t-Q'_{t+1})+\frac{1}{2}\sum_{\pi\in\Pi}\frac{(Q_t(\pi)-Q'_{t+1}(\pi))^2}{\xi_t(\pi)} \end{aligned}$$

$$\geq \frac{1}{2}\sum_{\pi\in\Pi}\frac{(Q_t(\pi)-Q'_{t+1}(\pi))^2}{\xi_t(\pi)},$$

where the last step is due to the optimality of Q'_{t+1}. On the other hand, we also have

$$\begin{aligned}
F'_t(Q_t)-F'_t(Q'_{t+1}) &= F_t(Q_t)-F_t(Q'_{t+1})+\eta\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle \\
&\leq \eta\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle && \text{(by optimality of } Q_t\text{)} \\
&\leq \eta\left(\sum_{\pi\in\Pi}\frac{(Q_t(\pi)-Q'_{t+1}(\pi))^2}{\xi_t(\pi)}\right)^{1/2}\left(\sum_{\pi\in\Pi}\xi_t(\pi)(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))^2\right)^{1/2}. \\
& && \text{(Cauchy-Schwarz inequality)}
\end{aligned}$$

Combining the two inequalities shows

$$\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle \leq 2\eta\sum_{\pi\in\Pi}\xi_t(\pi)(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))^2,$$

which proves Eq. (8) (since $D(Q'_{t+1},Q_t)$ is non-negative).

To proves Eq. (9), note that $Q'_{t+1}(\pi)=\frac{1}{Z}Q_t(\pi)\exp(-\eta(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)))$ where

$$Z=\sum_{\pi\in\Pi}Q_t(\pi)\exp(-\eta(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)))$$

is the normalization factor. Direct calculation shows

$$\begin{aligned}
&\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle-\frac{1}{\eta}D(Q'_{t+1},Q_t) \\
&=\sum_{\pi\in\Pi}\left\langle Q_t-Q'_{t+1},\mathcal{L}_t-\mathcal{M}_t\right\rangle-\frac{1}{\eta}\sum_{\pi\in\Pi}Q'_{t+1}(\pi)\ln Q'_{t+1}(\pi)+\frac{1}{\eta}\sum_{\pi\in\Pi}Q'_{t+1}(\pi)\ln Q_t(\pi) \\
&=\sum_{\pi\in\Pi}\left\langle Q_t,\mathcal{L}_t-\mathcal{M}_t\right\rangle+\frac{1}{\eta}\ln Z \\
&\leq\sum_{\pi\in\Pi}\left\langle Q_t,\mathcal{L}_t-\mathcal{M}_t\right\rangle+\frac{1}{\eta}\ln\sum_{\pi\in\Pi}Q_t(\pi)\left(1-\eta(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))+\eta^2(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))^2\right) \\
& && \text{(by } e^{-z}\leq 1-z+z^2 \text{ for } z\geq -1 \text{ and the condition } \eta(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))\geq -1\text{)} \\
&=\sum_{\pi\in\Pi}\left\langle Q_t,\mathcal{L}_t-\mathcal{M}_t\right\rangle+\frac{1}{\eta}\ln\left(1-\eta\left\langle Q_t,\mathcal{L}_t-\mathcal{M}_t\right\rangle+\eta^2\sum_{\pi\in\Pi}Q_t(\pi)(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi))^2\right) \\
&\leq\eta\sum_{\pi\in\Pi}Q_t(\pi)\left(\mathcal{L}_t(\pi)-\mathcal{M}_t(\pi)\right)^2. && \text{(by } \ln(1+z)\leq z\text{)}
\end{aligned}$$

This finishes the proof. ■

Proof [of Theorem 4] We directly apply Lemma 14 with $\mathcal{M}_t(\pi)=m_t(\phi_t(\pi(x_t)))$ and $\mathcal{L}_t(\pi)=\widehat{\ell}_t(\phi_t(\pi(x_t)))$ and use Eq. (8) with Q^* concentrating on the best policy π^*. Summing over t gives

$$\sum_{t=1}^{T}\sum_{\pi\in\Pi}Q_t(\pi)\widehat{\ell}_t(\phi_t(\pi(x_t)))-\sum_{t=1}^{T}\widehat{\ell}_t(\phi_t(\pi^*(x_t)))$$

$$\leq \frac{D(Q^*, Q_1')}{\eta} + 2\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \sum_{\pi : \phi_t(\pi(x_t))=a} \xi_t(\pi) \left(\widehat{\ell}_t(a) - m_t(a)\right)^2.$$

$$\leq \frac{\ln N}{\eta} + 2\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_t} \left(\widehat{\ell}_t(a) - m_t(a)\right)^2$$

$$= \frac{\ln N}{\eta} + 2\eta \sum_{t=1}^{T} \left(\widehat{\ell}_t(a_t) - m_t(a_t)\right)^2,$$

where in the last step we use the fact that $\widehat{\ell}_t(a) - m_t(a)$ is non-zero only if $a = a_t$. Note that this basically proves Eq. (2) (with remapping). The rest of the proof follows the analysis sketch in Section 3. First, we plug in the definition of $\widehat{\ell}_t$ and continue to bound the last expression by

$$\frac{\ln N}{\eta} + 2\eta \sum_{t=1}^{T} \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2} \leq \frac{\ln N}{\eta} + \frac{2\eta K}{\mu} \sum_{t=1}^{T} \frac{\|\ell_t - m_t\|_\infty^2}{p_t(a_t)},$$

where the last step uses the fact $p_t(a_t) \geq \mu/|\mathcal{A}_t| \geq \mu/K$. Taking expectation on both sides leads to

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{\pi \in \Pi} Q_t(\pi) \ell_t(\phi_t(\pi(x_t)))\right] - \sum_{t=1}^{T} \ell_t(\phi_t(\pi^*(x_t))) \leq \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu}. \tag{11}$$

Next, consider the expected loss of the algorithm at time t:

$$\begin{aligned}\sum_{a \in \mathcal{A}_t} p_t(a) \ell_t(a) &= (1-\mu) \sum_{a \in \mathcal{A}_t} \left(\sum_{\pi : \phi_t(\pi(x_t))=a} Q_t(\pi)\right) \ell_t(a) + \frac{\mu}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} \ell_t(a) \\ &= (1-\mu) \sum_{\pi \in \Pi} Q_t(\pi) \ell_t(\phi_t(\pi(x_t))) + \frac{\mu}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} \ell_t(a).\end{aligned}$$

Combining with Eq. (11) shows

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] &\leq (1-\mu) \sum_{t=1}^{T} \ell_t(\phi_t(\pi^*(x_t))) + \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu} + \sum_{t=1}^{T} \frac{\mu}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} \ell_t(a) \\ &= \sum_{t=1}^{T} \ell_t(\phi_t(\pi^*(x_t))) + \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu} + \sum_{t=1}^{T} \frac{\mu}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} \left(\ell_t(a) - \ell_t(\phi_t(\pi^*(x_t)))\right),\end{aligned}$$

where the last term can be further bounded as (by the definition of $\mathcal{A}_t$):

$$\begin{aligned}&\ell_t(a) - \ell_t(\phi_t(\pi^*(x_t)) \\ &= \ell_t(a) - m_t(a) + m_t(a) - m_t(\phi_t(\pi^*(x_t)) + m_t(\phi_t(\pi^*(x_t)) - \ell_t(\phi_t(\pi^*(x_t)) \\ &\leq 2\|\ell_t - m_t\|_\infty + \sigma.\end{aligned}$$

This shows

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] \leq \sum_{t=1}^{T} \ell_t(\phi_t(\pi^*(x_t))) + \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu} + \mu T \sigma + 2\mu \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty$$

$$\leq \sum_{t=1}^{T} \ell_t(\phi_t(\pi^*(x_t))) + \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu} + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T}.$$

(Cauchy-Schwarz inequality)

It remains to bound the bias due to remapping: when $\pi^*(x_t) \neq \phi_t(\pi^*(x_t))$ we have $\phi_t(\pi^*(x_t)) = a_t^*$, $m_t(a_t^*) \leq m_t(\pi^*(x_t)) - \sigma$, and

$$\begin{aligned}
&\ell_t(\phi_t(\pi^*(x_t))) - \ell_t(\pi^*(x_t)) \\
&= \ell_t(a_t^*) - m_t(a_t^*) + m_t(a_t^*) - m_t(\pi^*(x_t)) + m_t(\pi^*(x_t)) - \ell_t(\pi^*(x_t)), \\
&\leq 2\|\ell_t - m_t\|_\infty - \sigma \leq \frac{\|\ell_t - m_t\|_\infty^2}{\sigma}, \qquad (12)
\end{aligned}$$

where the last step is by the AM-GM inequality. When $\pi^*(x_t) = \phi_t(\pi^*(x_t))$, the above holds trivially. Summing over t we have thus shown

$$\text{Reg} \leq \frac{\ln N}{\eta} + \frac{2\eta K^2 \mathcal{E}}{\mu} + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T} + \frac{\mathcal{E}}{\sigma},$$

finishing the proof. ■

C.2. Adaptive Version of Algorithm 1

The adaptive version of Algorithm 1 is shown in Algorithm 4. We observe that when $\mathcal{E}$ is unknown, choosing actions only from $\mathcal{A}_t$ is problematic, because in the case when the predictors are highly inaccurate (that is, large $\mathcal{E}$), the environment can be such that the good actions are always outside $\mathcal{A}_t$ but the learner can never realize that. Based on this intuition, we remove the action remapping component in this case, implemented by simply setting $\sigma = 1$.

For the exploration parameter μ, note that its optimal choice is independent of $\mathcal{E}$ already in the known $\mathcal{E}$ case (see Theorem 4), which turns out to be also the case here (albeit with a different value).

Also note that standard Optimistic Online Mirror Descent analysis requires using the same learning rate in Lines 2 and 6 (see (Wei and Luo, 2018) for example). However, using η_t in both places is invalid since a_t and $\ell_t(a_t)$ are unknown when executing Line 2, while using η_{t-1} in both places also leads to some technical issue due to the large magnitude of loss estimators. Instead, we use η_{t-1} in Line 2 and η_t in Line 6, and carefully bound the bias introduced by this learning rate mismatch. Analyzing this learning rate mismatch is the key of our analysis, as we will show later.

A minor but also necessary difference with Algorithm 1 is that we also enforce Q_t and Q'_t to be in the clipped simplex $\overline{\Delta}_\Pi = \left\{Q \in \Delta_\Pi : Q(\pi) \geq \frac{1}{NT}, \forall \pi \in \Pi\right\}$, by writing the updates of Q_t and Q'_t in the Optimistic Online Mirror Descent form over $\overline{\Delta}_\Pi$.

Proof [of Theorem 5] Define $m'_t = \frac{\eta_{t-1}}{\eta_t} m_t$. Note that the update in Line 2 and Line 6 in Algorithm 4 is the same as Eq. (10) with $\eta = \eta_t$, $\mathcal{M}_t(\pi) = m'_t(\pi(x_t))$, and $\mathcal{L}_t(\pi) = \widehat{\ell}_t(\pi(x_t))$, except that the constraint set becomes $\overline{\Delta}_\Pi$. By the exact same arguments as the proof of Lemma 14, we conclude that Eq. (8) holds for any $Q^* \in \overline{\Delta}_\Pi$. In particular, we pick $Q^* = \left(1 - \frac{1}{T}\right) \mathbf{e}_{\pi^*} + \frac{1}{NT}\mathbf{1} \in \overline{\Delta}_\Pi$, where

Algorithm 4 EXP4.OVAR: Optimistic EXP4 with Variance-adaptivity and Action Remapping

Parameter: exploration probability $\mu \in [0,1]$.

Define: $\overline{\Delta}_\Pi = \left\{Q \in \Delta_\Pi : Q(\pi) \geq \frac{1}{NT},\ \forall \pi \in \Pi\right\}$ and $D(Q,P) = \sum_{\pi\in\Pi} Q(\pi) \ln \frac{Q(\pi)}{P(\pi)}$.

Initialize: $Q'_1(\pi) = \frac{1}{N}$ for all $\pi \in \Pi$ and $\eta_0 = \sqrt{\log(NT)}$.

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

Receive x_t and m_t.

Calculate

$$Q_t = \operatorname*{argmin}_{Q\in\overline{\Delta}_\Pi} \left\{ \eta_{t-1} \sum_{\pi\in\Pi} Q(\pi) m_t(\pi(x_t)) + D(Q, Q'_t) \right\}.$$

Calculate $p_t \in \Delta_K$: $\quad p_t(a) = (1-\mu) \sum_{\pi:\pi(x_t)=a} Q_t(\pi) + \frac{\mu}{K}$.

Sample $a_t \sim p_t$ and receive $\ell_t(a_t)$.

Construct estimator: $\quad \widehat{\ell}_t(a) = \frac{\ell_t(a) - m_t(a)}{p_t(a)} \mathbb{1}[a_t = a] + m_t(a)$ for all $a \in [K]$.

Calculate

$$Q'_{t+1} = \operatorname*{argmin}_{Q\in\overline{\Delta}_\Pi} \left\{ \eta_t \sum_{\pi\in\Pi} Q(\pi) \widehat{\ell}_t(\pi(x_t)) + D(Q, Q'_t) \right\}$$

where

$$\eta_t = \sqrt{\log(NT)} \left(1 + \sum_{s=1}^{t} \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2} \right)^{-\frac{1}{2}}. \tag{13}$$

$\mathbf{e}_{\pi^*}$ is the distribution that concentrates on π^* and $\frac{1}{N}\mathbf{1}$ is the uniform distribution over Π. With this Q^*, summing Eq. (8) over t, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{\pi\in\Pi} Q_t(\pi) \widehat{\ell}_t(\pi(x_t)) - \left(1 - \frac{1}{T}\right) \sum_{t=1}^{T} \widehat{\ell}_t(\pi^*(x_t)) - \frac{1}{NT} \sum_{t=1}^{T} \sum_{\pi\in\Pi} \widehat{\ell}(\pi(x_t)) \\
&\leq \sum_{t=1}^{T} \left(\frac{D(Q^*, Q'_t) - D(Q^*, Q'_{t+1})}{\eta_t} \right) + 2 \sum_{t=1}^{T} \eta_t \sum_{\pi\in\Pi} \xi_t(\pi) \left(\widehat{\ell}_t(\pi(x_t)) - m'_t(\pi(x_t)) \right)^2. \qquad (14)
\end{aligned}$$

The first term on the right-hand side of Eq. (14) is equal to

$$\frac{D(Q^*, Q'_1)}{\eta_1} + \sum_{t=2}^{T} D(Q^*, Q'_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) - \frac{D(Q^*, Q'_{T+1})}{\eta_T}. \tag{15}$$

Note that for all $Q \in \overline{\Delta}_\Pi$, we have $D(Q^*, Q) = \sum_{\pi\in\Pi} Q^*(\pi) \log \frac{Q^*(\pi)}{Q(\pi)} \le \sum_{\pi\in\Pi} Q^*(\pi) \log \frac{1}{1/(NT)} = \log(NT)$. Since $\frac{1}{\eta_t} \ge \frac{1}{\eta_{t-1}}$, we can thus upper bound Eq. (15) by

$$\frac{\log(NT)}{\eta_1} + \sum_{t=2}^{T} \log(NT) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) = \frac{\log(NT)}{\eta_T}.$$

We continue to show that the second term on the right-hand side of Eq. (14) is in fact also of order $\mathcal{O}\left(\frac{\log(NT)}{\eta_T}\right)$. First, by direct calculation we have

$$\begin{aligned}
&\sum_{t=1}^{T} \eta_t \sum_{\pi\in\Pi} \xi_t(\pi) \left(\widehat{\ell}_t(\pi(x_t)) - m'_t(\pi(x_t))\right)^2 \\
&= \sum_{t=1}^{T} \eta_t \sum_{\pi\in\Pi} \xi_t(\pi) \left(\frac{(\ell_t(a_t) - m_t(a_t))\mathbb{1}[\pi_t(x_t) = a_t]}{p_t(a_t)} + m_t(\pi(x_t)) - \frac{\eta_{t-1}}{\eta_t} m_t(\pi(x_t))\right)^2 \\
&\le \sum_{t=1}^{T} 2\eta_t \left(\frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2}\right) + \sum_{t=1}^{T} 2\eta_t \left(1 - \frac{\eta_{t-1}}{\eta_t}\right)^2.
\end{aligned}$$

To deal with the first term in the last expression, we define $b_t = \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2}$ so that

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \left(\frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2}\right) &= (\log NT)^{\frac{1}{2}} \sum_{t=1}^{T} \frac{b_t}{\sqrt{1 + \sum_{s=1}^{t} b_t}} \le (\log NT)^{\frac{1}{2}} \int_0^{\sum_{s=1}^{T} b_t} \frac{dx}{\sqrt{1+x}} \\
&= \mathcal{O}\left((\log NT)^{\frac{1}{2}} \left(1 + \sum_{t=1}^{T} b_t\right)^{\frac{1}{2}}\right) = \mathcal{O}\left(\frac{\log(NT)}{\eta_T}\right).
\end{aligned}$$

To deal with the second term, simply note that

$$\sum_{t=1}^{T} \eta_t \left(1 - \frac{\eta_{t-1}}{\eta_t}\right)^2 = \sum_{t=1}^{T} \frac{1}{\eta_t} (\eta_t - \eta_{t-1})^2 \le \frac{(\log NT)^{\frac{1}{2}}}{\eta_T} \sum_{t=1}^{T} (\eta_{t-1} - \eta_t) \le \frac{\log(NT)}{\eta_T}.$$

Combining everything above, we conclude that the right-hand side of Eq. (14) is upper bounded by

$$\begin{aligned}
\mathcal{O}\left(\frac{\log(NT)}{\eta_T}\right) &= \mathcal{O}\left(\left(\log(NT) + \log(NT) \sum_{t=1}^{T} \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)^2}\right)^{\frac{1}{2}}\right) \\
&= \mathcal{O}\left(\left(\log(NT) + \frac{K\log(NT)}{\mu} \sum_{t=1}^{T} \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)}\right)^{\frac{1}{2}}\right),
\end{aligned}$$

whose expectation is upper bounded by (using Jensen's inequality)

$$\mathcal{O}\left(\left(\log(NT) + \frac{K\log(NT)}{\mu} \mathbb{E}\left[\sum_{t=1}^{T} \frac{(\ell_t(a_t) - m_t(a_t))^2}{p_t(a_t)}\right]\right)^{\frac{1}{2}}\right)$$

$$= \mathcal{O}\left(\left(\log(NT) + \frac{K^2 \log(NT)\mathcal{E}}{\mu}\right)^{\frac{1}{2}}\right) = \widetilde{\mathcal{O}}\left(d\sqrt{\frac{\mathcal{E}}{\mu}}\right).$$

Now we lower bound the expectation of the left-hand side of Eq. (14):

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t)) - \left(1-\frac{1}{T}\right)\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t)) - \frac{1}{NT}\sum_{t=1}^{T}\sum_{\pi\in\Pi}\widehat{\ell}(\pi(x_t))\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\ell_t(\pi(x_t)) - \sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] - 1 \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\left(p_t(a) + \mu\sum_{\pi:\pi(x_t)=a} Q_t(\pi) - \frac{\mu}{K}\right)\ell_t(a) - \sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] - 1 \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\ell_t(a) - \sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] - 1 - \mu T \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t) - \sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] - 1 - \mu T.
\end{aligned}$$

Combining the bounds for both sides of Eq. (14) finishes the proof. ■

C.3. Algorithms and Analysis for Multiple Predictors

The pseudocode of our algorithm for multiple predictors is in Algorithm 5. As discussed in Section 5, there are several extra ingredients compared to Algorithm 1 in this case. First, we maintain an active set $\mathcal{P}_t$ of predictors that are still plausibly the best predictor:

$$i^* = \operatorname*{argmin}_{i\in[M]} \sum_{t=1}^{T} \|\ell_t - m_t^i\|_\infty^2.$$

Specifically, the variable V^i maintains the remaining error "budget" for each predictor (starting from $\mathcal{E}^*$), and is decreased by $(\ell_t(a_t) - m_t^i(a_t))^2$ at the end of each round. Then $\mathcal{P}_t$ is simply the set of predictors with a non-negative error budget. Second, for each action a, we let

$$m_t(a) = \min_{i\in\mathcal{P}_t} m_t^i(a),$$

to be the smallest prediction among all active predictors, and treat m_t as if it was the only prediction similar to the single predictor case, which can be seen as a form of optimism.

Finally and perhaps most importantly, we construct the set $\mathcal{A}_t$ using a different baseline. Essentially, the baseline is a_t, the action to be chosen by the algorithm, which is of course not available before constructing $\mathcal{A}_t$. However, instead of using $m_t(a_t)$ as the baseline in the definition of $\mathcal{A}_t$, we use the expected prediction $\langle p_t, m_t\rangle$, and instead of explicitly remapping an action $a \notin \mathcal{A}_t$ to be a_t, we change the values of $\widehat{\ell}_t(a)$ and $m_t(a)$ to $\langle p_t, \widehat{\ell}_t\rangle$ and $\langle p_t, m_t\rangle$ respectively for these actions. While this is still a self-referential scheme since the construction of p_t depends on $\mathcal{A}_t$,

Algorithm 5 EXP4.MOAR: Optimistic EXP4 with Action Remapping for Multiple predictors

Parameters: learning rate $\eta > 0$, threshold $\sigma > 0$, exploration probability $\mu \in [0, \frac{1}{2}]$, best error $\mathcal{E}^*$.
Initialize: $Q'_1(\pi) = \frac{1}{N}$ for all $\pi \in \Pi$, budget $V^i = \mathcal{E}^*$ for all $i \in [M]$, active set $\mathcal{P}_1 = [M]$.
for $t = 1, \ldots, T$ **do**

Receive x_t and m_t^i for all $i \in [M]$. Let $m_t(a) = \min_{i \in \mathcal{P}_t} m_t^i(a)$ for all $a \in [K]$.

Step 1. Jointly decide the awake action set $\mathcal{A}_t$ and the action distribution p_t.
Let $b_1, b_2, \ldots, b_K$ be a permutation of $[K]$ such that

$$m_t(b_1) \leq m_t(b_2) \leq \ldots \leq m_t(b_K).$$

for $j = 1, 2, \ldots, K$ **do**
Set $\mathcal{A} = \{b_1, \ldots, b_j\}$.
Calculate $p \in \Delta_K$:

$$p(a) = \begin{cases} (1-\mu)\frac{\sum_{\pi:\pi(x_t)=a} Q'_t(\pi) \exp\left(-\eta m_t(\pi(x_t))\right)}{\sum_{\pi:\pi(x_t)\in\mathcal{A}} Q'_t(\pi) \exp\left(-\eta m_t(\pi(x_t))\right)} + \frac{\mu}{|\mathcal{A}|}, & \text{for } a \in \mathcal{A}, \\ 0, & \text{for } a \notin \mathcal{A}. \end{cases}$$

if $j = K$ or $m_t(b_j) \leq \langle p, m_t \rangle + \sigma \leq m_t(b_{j+1})$ **then**
$\mathcal{A}_t = \mathcal{A}$, $p_t = p$, **break**.

Step 2. Choose an action and construct loss estimators.
Sample $a_t \sim p_t$ and receive $\ell_t(a_t)$.
Construct estimator:

$$\widehat{\ell}_t(a) = \begin{cases} \frac{(\ell_t(a) - m_t(a))\mathbb{1}[a_t = a]}{p_t(a)} + m_t(a), & \text{for } a \in \mathcal{A}_t, \\ \sum_{a \in \mathcal{A}_t} p_t(a)\widehat{\ell}_t(a), & \text{for } a \notin \mathcal{A}_t. \end{cases}$$

Step 3. Make updates.
Calculate $Q'_{t+1} \in \Delta_\Pi$: $\quad Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp\left(-\eta \widehat{\ell}_t(\pi(x_t))\right)$.
for $i \in \mathcal{P}_t$ **do**
$V^i \leftarrow V^i - (\ell_t(a_t) - m_t^i(a_t))^2$.
Update active set $\mathcal{P}_{t+1} = \left\{i \in \mathcal{P}_t : V^i \geq 0\right\}$.

we show that this can in fact be implemented efficiently by trying all the K possibilities for $\mathcal{A}_t$: $\{b_1\}, \{b_1, b_2\}, \ldots, \{b_1, b_2, \ldots, b_K\}$, where $b_1, \ldots, b_K$ are such that $m_t(b_1) \leq m_t(b_2) \leq \ldots \leq m_t(b_K)$. The concrete procedure is detailed in Step 1 of Algorithm 5, and we prove in the following lemma that it does exactly what we want.

Lemma 15 *Define $\mathcal{M}_t, \mathcal{L}_t \in \mathbb{R}^K$, $Q_t \in \Delta_\Pi$, and $q_t \in \Delta_{\mathcal{A}_t}$ as*

$$\mathcal{M}_t(a) = \begin{cases} m_t(\pi(x_t)), & \textit{if } \pi(x_t) \in \mathcal{A}_t, \\ \langle p_t, m_t \rangle, & \textit{otherwise} \end{cases}, \quad \textit{and} \quad \mathcal{L}_t(a) = \begin{cases} \widehat{\ell}_t(\pi(x_t)), & \textit{if } \pi(x_t) \in \mathcal{A}_t, \\ \left\langle p_t, \widehat{\ell}_t \right\rangle, & \textit{otherwise} \end{cases},$$

$$Q_t(\pi) \propto Q'_t(\pi)\exp\left(-\eta \mathcal{M}_t(\pi)\right), \quad \textit{and}$$
$$q_t(a) = \frac{\sum_{\pi:\pi(x_t)=a} Q_t(\pi)}{\sum_{\pi:\pi(x_t)\in\mathcal{A}_t} Q_t(\pi)}.$$

Then Algorithm 5 ensures the following properties:

$$Q'_{t+1}(\pi) \propto Q'_t(\pi)\exp\left(-\eta \mathcal{L}_t(\pi)\right), \tag{16}$$
$$p_t(a) = \begin{cases} (1-\mu)q_t(a) + \frac{\mu}{|\mathcal{A}_t|}, & \textit{if } a \in \mathcal{A}_t, \\ 0, & \textit{else}, \end{cases} \tag{17}$$
$$\mathcal{A}_t = \{a \in [K] : m_t(a) \le \langle p_t, m_t\rangle + \sigma\}. \tag{18}$$

Proof The first property on Q'_{t+1} is simply by the definition of $\mathcal{L}_t$ and $\left\langle p_t, \widehat{\ell}_t \right\rangle = \sum_{a\in\mathcal{A}_t} p_t(a)\widehat{\ell}_t(a)$. The second equation is also clear by the definition of Q_t:

$$q_t(a) = \frac{\sum_{\pi:\pi(x_t)=a} Q_t(\pi)}{\sum_{\pi:\pi(x_t)\in\mathcal{A}_t} Q_t(\pi)} = \frac{\sum_{\pi:\pi(x_t)=a} Q'_t(\pi)\exp\big(-\eta m_t(\pi(x_t))\big)}{\sum_{\pi:\pi(x_t)\in\mathcal{A}} Q'_t(\pi)\exp\big(-\eta m_t(\pi(x_t))\big)}.$$

The last equation clearly holds when $j < K$ and the condition $m_t(b_j) \le \langle p, m_t\rangle + \sigma \le m_t(b_{j+1})$ holds and triggers the "break" statement, so it remains to prove Eq. (18) if the "break" statement is triggered in the last iteration when $j = K$, in which case we have for all $j < K$,

$$\langle p^j, m_t\rangle + \sigma < m_t(b_j) \quad \text{or} \quad \langle p^j, m_t\rangle + \sigma > m_t(b_{j+1}) \tag{19}$$

where p^j is the value of p in the j-th iteration.

Note that for all $k \le j$, we have $p^j(b_k) \ge p^{j+1}(b_k)$ by the definition of p, and also $p^{j+1}(b_{j+1}) = \sum_{k\le j}(p^j(b_k) - p^{j+1}(b_k))$. With these facts we prove $\langle p^{j+1}, m_t\rangle \ge \langle p^j, m_t\rangle$ below:

$$\begin{aligned}
&\langle p^{j+1}, m_t\rangle \\
&= p^{j+1}(b_{j+1})m_t(b_{j+1}) + \sum_{k\le j} p^{j+1}(b_k)m_t(b_k) \\
&= \sum_{k\le j}\left(p^j(b_k) - p^{j+1}(b_k)\right)m_t(b_{j+1}) + p^{j+1}(b_k)m_t(b_k) \\
&\ge \sum_{k\le j}\left(p^j(b_k) - p^{j+1}(b_k)\right)m_t(b_k) + p^{j+1}(b_k)m_t(b_k) \qquad (m_t(b_{j+1}) \ge m_t(b_k),\ \forall k \le j) \\
&= \langle p^j, m_t\rangle .
\end{aligned}$$

Therefore, realizing $\langle p^1, m_t\rangle + \sigma = m_t(b_1) + \sigma > m_t(b_1)$ and thus $\langle p^1, m_t\rangle + \sigma > m_t(b_2)$ by Eq. (19), we have

$$\langle p^2, m_t\rangle + \sigma \ge \langle p^1, m_t\rangle + \sigma > m_t(b_2),$$

which in turn further implies (by repeatedly using Eq. (19) and $\langle p^{j+1}, m_t\rangle \ge \langle p^j, m_t\rangle$)

$$\langle p^3, m_t\rangle + \sigma \ge \langle p^2, m_t\rangle + \sigma > m_t(b_3)$$

$$\cdots,$$

$$\left\langle p^K, m_t \right\rangle + \sigma \geq \left\langle p^{K-1}, m_t \right\rangle + \sigma > m_t(b_K).$$

The last statement proves Eq. (18) again. ■

With this fact, the analysis of the algorithm follows similar steps as in the proof of Theorem 4. First, we apply Lemma 14 to prove the following.

Lemma 16 *Algorithm 5 ensures for any* $\pi^* \in \Pi$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t))\right] + \mathcal{O}\left(\frac{\ln N}{\eta} + \frac{\eta K^2\mathcal{E}^*}{\mu} + \frac{\eta K M\mathcal{E}^*}{\mu}\right).$$

Proof We apply Lemma 14 with $\mathcal{M}_t$ and $\mathcal{L}_t$ defined in Lemma 15. First note that

$$\mathcal{L}_t(\pi) - \mathcal{M}_t(\pi) = \begin{cases} \ell_t(a_t) - m_t(a_t), & \text{if } \pi(x_t) \notin \mathcal{A}_t, \\ \frac{\ell_t(a_t) - m_t(a_t)}{p_t(a_t)} & \text{if } \pi(x_t) = a_t, \\ 0, & \text{if } a_t \neq \pi(x_t) \in \mathcal{A}_t. \end{cases}$$

Therefore, when $\ell_t(a_t) \geq m_t(a_t)$, the condition $\mathcal{L}(\pi) - \mathcal{M}_t(\pi) \geq -1/\eta$ holds and we apply Eq. (9) and bound the last term by

$$\begin{aligned}
&\eta \sum_{\pi\in\Pi} Q_t(\pi)\left(\mathcal{L}_t(\pi) - \mathcal{M}_t(\pi)\right)^2 \\
&= \eta\left(\sum_{\pi:\pi(x_t)=a_t} Q_t(\pi)\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t^2(a_t)} + \sum_{\pi:\pi(x_t)\notin\mathcal{A}_t} Q_t(\pi)\left(\ell_t(a_t) - m_t(a_t)\right)^2\right) \\
&\leq \eta\left(q_t(a_t)\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t^2(a_t)} + \left(\ell_t(a_t) - m_t(a_t)\right)^2\right) \\
&\leq \eta\left(\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{(1-\mu)p_t(a_t)} + \left(\ell_t(a_t) - m_t(a_t)\right)^2\right) && \text{(by Eq. (17))} \\
&\leq \eta\cdot\mathcal{O}\left(\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t(a_t)}\right) && (\mu \leq 1/2) \\
&\leq \frac{\eta K}{\mu}\cdot\mathcal{O}\left(\left(\ell_t(a_t) - m_t(a_t)\right)^2\right) && (p_t(a_t) \geq \mu/K)
\end{aligned}$$

On the other hand, if $\ell_t(a_t) \leq m_t(a_t)$, we apply Eq. (8) and bound the last term by

$$\begin{aligned}
&2\eta \sum_{\pi\in\Pi} \xi_t(\pi)\left(\mathcal{L}_t(\pi) - \mathcal{M}_t(\pi)\right)^2 \\
&= 2\eta\left(\sum_{\pi:\pi(x_t)=a_t} \xi_t(\pi)\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t^2(a_t)} + \sum_{\pi:\pi(x_t)\notin\mathcal{A}_t} \xi_t(\pi)\left(\ell_t(a_t) - m_t(a_t)\right)^2\right) \\
&\leq 2\eta\left(\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t^2(a_t)} + \left(\ell_t(a_t) - m_t(a_t)\right)^2\right)
\end{aligned}$$

$$\leq \frac{\eta K}{\mu} \cdot \mathcal{O}\left(\frac{\left(\ell_t(a_t) - m_t(a_t)\right)^2}{p_t(a_t)}\right), \qquad (p_t(a_t) \geq \mu/K)$$

$$\leq \frac{\eta K}{\mu} \cdot \mathcal{O}\left(\frac{\left(\ell_t(a_t) - m_t^{i^*}(a_t)\right)^2}{p_t(a_t)}\right). \qquad (\ell_t(a_t) \leq m_t(a_t) \leq m_t^{i^*}(a_t))$$

Combining the two situations, setting Q^* to concentrate on π^*, and summing over t show:

$$\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t)) \leq \sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t)) + \frac{\ln N}{\eta} + \frac{\eta K}{\mu} \cdot \mathcal{O}\left(\sum_{t=1}^{T}\frac{\left(\ell_t(a_t) - m_t^{i^*}(a_t)\right)^2}{p_t(a_t)} + \left(\ell_t(a_t) - m_t(a_t)\right)^2\right).$$

Taking expectation on both sides we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t))\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t))\right] + \frac{\ln N}{\eta} \\
&\quad + \frac{\eta K}{\mu} \cdot \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in[K]}\left(\ell_t(a) - m_t^{i^*}(a)\right)^2 + \left(\ell_t(a_t) - m_t(a_t)\right)^2\right]\right) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t))\right] + \frac{\ln N}{\eta} + \mathcal{O}\left(\frac{\eta K^2\mathcal{E}^*}{\mu} + \frac{\eta K M\mathcal{E}^*}{\mu}\right),
\end{aligned}$$

where in the last step we use Lemma 19. This finishes the proof. ■

Next, we relate the term $\mathbb{E}[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t))]$ to the loss of the algorithm, and the term $\mathbb{E}[\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t))]$ to the loss of the best policy, in the following two lemmas respectively.

Lemma 17 *Algorithm 5 ensures*

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t))\right] + \mathcal{O}\left(\sqrt{\mu M\mathcal{E}^* T} + \mu T\sigma + \mu^2 T\right).$$

Proof With $Z_t = \sum_{\pi:\pi(x_t)\in\mathcal{A}_t} Q_t(\pi)$ so that $Z_t q_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$, we rewrite the expected loss of the algorithm as

$$\begin{aligned}
&\mathbb{E}\left[\ell_t(a_t)\right] \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}_t} p_t(a)\ell_t(a)\right] \\
&= \mathbb{E}\left[(1-\mu)\sum_{a\in\mathcal{A}_t} q_t(a)\ell_t(a) + \frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)\right] && \text{(by Eq. (17))} \\
&= \mathbb{E}\left[(1-\mu)\sum_{a\in\mathcal{A}_t}\left(Z_t q_t(a) + (1-Z_t)\, q_t(a)\right)\ell_t(a) + \frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)\right]
\end{aligned}$$

$$= \mathbb{E}\left[(1-\mu)Z_t \sum_{a\in\mathcal{A}_t} q_t(a)\ell_t(a) + (1-Z_t)\sum_{a\in\mathcal{A}_t}\left(p_t(a) - \frac{\mu}{|\mathcal{A}_t|}\right)\ell_t(a) + \frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)\right]$$

(by Eq. (17))

$$= \mathbb{E}\left[Z_t \sum_{a\in\mathcal{A}_t} q_t(a)\ell_t(a) + (1-Z_t)\langle p_t, \ell_t\rangle + \mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{\ell_t(a)}{|\mathcal{A}_t|} - q_t(a)\ell_t(a)\right)\right]$$

$$= \mathbb{E}\left[\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t)) + \mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{\ell_t(a)}{|\mathcal{A}_t|} - q_t(a)\ell_t(a)\right)\right],$$

where in the last step we use the fact

$$\begin{aligned}\sum_{\pi\in\Pi} Q_t(\pi)\widehat{\ell}_t(\pi(x_t)) &= \sum_{a\in\mathcal{A}_t}\sum_{\pi:\pi(x_t)=a} Q_t(\pi)\widehat{\ell}_t(a) + \sum_{\pi:\pi(x_t)\notin\mathcal{A}_t} Q_t(\pi)\left\langle p_t, \widehat{\ell}_t\right\rangle \\ &= Z_t\sum_{a\in\mathcal{A}_t} q_t(a)\widehat{\ell}_t(a) + (1-Z_t)\left\langle p_t, \widehat{\ell}_t\right\rangle\end{aligned}$$

by the definition of $\widehat{\ell}_t$. It thus remains to bound $\mathbb{E}\left[\sum_{t=1}^T \mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{\ell_t(a)}{|\mathcal{A}_t|} - q_t(a)\ell_t(a)\right)\right]$, which we decompose into four terms:

$$\mathbb{E}\left[\sum_{t=1}^T \frac{\mu Z_t}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}(\ell_t(a) - m_t(a))\right],$$

$$\mathbb{E}\left[\sum_{t=1}^T \mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{m_t(a)}{|\mathcal{A}_t|} - q_t(a)m_t(a)\right)\right],$$

$$\mathbb{E}\left[\sum_{t=1}^T \mu Z_t \sum_{a\in\mathcal{A}_t}\left(q_t(a)m_t(a) - q_t(a)m_t^{i^*}(a)\right)\right],$$

$$\mathbb{E}\left[\sum_{t=1}^T \mu Z_t \sum_{a\in\mathcal{A}_t}\left(q_t(a)m_t^{i^*}(a) - q_t(a)\ell_t(a)\right)\right].$$

The first term can be bounded as

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T \frac{\mu Z_t}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}(\ell_t(a) - m_t(a))\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^T \frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}|\ell_t(a) - m_t(a)|\right] \\ &\le \mathbb{E}\left[\sqrt{\sum_{t=1}^T\sum_{a\in\mathcal{A}_t}\frac{\mu}{|\mathcal{A}_t|}}\sqrt{\sum_{t=1}^T\sum_{a\in\mathcal{A}_t}\frac{\mu}{|\mathcal{A}_t|}(\ell_t(a) - m_t(a))^2}\right] && \text{(Cauchy-Schwarz inequality)} \\ &\le \sqrt{\mu T}\cdot\mathbb{E}\left[\sqrt{\sum_{t=1}^T\sum_{a\in\mathcal{A}_t} p_t(a)(\ell_t(a) - m_t(a))^2}\right] && (p_t(a)\ge\mu/|\mathcal{A}_t|)\end{aligned}$$

$$\leq \sqrt{\mu T} \cdot \sqrt{\mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(a_t) - m_t(a_t))^2\right]} \qquad \text{(Jensen's inequality)}$$

$$\leq \mathcal{O}\left(\sqrt{\mu M \mathcal{E}^* T}\right). \qquad \text{(by Lemma 19)}$$

The second term can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{m_t(a)}{|\mathcal{A}_t|} - q_t(a)m_t(a)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mu Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{\langle p_t, m_t\rangle + \sigma}{|\mathcal{A}_t|} - q_t(a)m_t(a)\right)\right] && \text{(by Eq. (18))} \\
&\leq \mu T\sigma + \mathbb{E}\left[\sum_{t=1}^{T}\mu Z_t \sum_{a\in\mathcal{A}_t}\left(p_t(a) - q_t(a)\right)m_t(a)\right] \\
&= \mu T\sigma + \mu^2\mathbb{E}\left[\sum_{t=1}^{T} Z_t \sum_{a\in\mathcal{A}_t}\left(\frac{1}{|\mathcal{A}_t|} - q_t(a)\right)m_t(a)\right] && \text{(by Eq. (17))} \\
&\leq \mu T\sigma + \mu^2 T.
\end{aligned}$$

The third term is simply non-positive by the definition of m_t, and finally the four term can be bounded by (using Cauchy-Schwarz inequality again)

$$\mathbb{E}\left[\sum_{t=1}^{T}\mu Z_t \sum_{a\in\mathcal{A}_t}\left(q_t(a)m_t^{i^*}(a) - q_t(a)\ell_t(a)\right)\right] \leq \mu\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t - m_t^{i^*}\|_\infty\right] \leq \mu\sqrt{\mathcal{E}^* T},$$

which can be absorbed by the bound of the first term since $\mu \leq 1$. Combining all the bounds proves the lemma. ■

Lemma 18 *Algorithm 5 ensures*

$$\mathbb{E}\left[\sum_{t=1}^{T}\widehat{\ell}_t(\pi^*(x_t))\right] \leq \sum_{t=1}^{T}\ell_t(\pi^*(x_t)) + \mathcal{O}\left(\frac{M\mathcal{E}^*}{\sigma}\right).$$

Proof Note that

$$\mathbb{E}\left[\widehat{\ell}_t(\pi^*(x_t))\right] = \ell_t(\pi^*(x_t)) + \mathbb{E}\left[\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(\ell_t(a_t) - \ell_t(\pi^*(x_t))\right)\right],$$

where the second term is bounded as

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(\ell_t(a_t) - m_t(a_t) + m_t(a_t) - m_t(\pi^*(x_t)) + m_t^{i*}(\pi^*(x_t)) - \ell_t(\pi^*(x_t))\right)\right] \\
&\qquad\qquad (m_t(a) \leq m_t^{i^*}(a)) \\
&= \mathbb{E}\left[\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(\ell_t(a_t) - m_t(a_t) + \langle p_t, m_t\rangle - m_t(\pi^*(x_t)) + m_t^{i*}(\pi^*(x_t)) - \ell_t(\pi^*(x_t))\right)\right] \\
&\leq \mathbb{E}\left[|\ell_t(a_t) - m_t(a_t)| + \|\ell_t - m_t^{i*}\|_\infty - \sigma\right] && \text{(by Eq. (18))}
\end{aligned}$$

$$= \mathbb{E}\left[\left(|\ell_t(a_t) - m_t(a_t)| - \frac{\sigma}{2}\right) + \left(\|\ell_t - m_t^{i*}\|_\infty - \frac{\sigma}{2}\right)\right]$$
$$\leq \mathbb{E}\left[\frac{(\ell_t(a_t) - m_t(a_t)|)^2}{2\sigma} + \frac{\|\ell_t - m_t^{i*}\|_\infty^2}{2\sigma}\right]. \qquad \text{(AM-GM inequality)}$$

Summing over t and using the fact $\sum_{t=1}^T \|\ell_t - m_t^{i*}\|_\infty^2 = \mathcal{E}^*$ and Lemma 19 complete the proof. ■

In the proofs of all the three lemmas above, we have used the following fact:

Lemma 19 *Algorithm 5 ensures* $\sum_{t=1}^T (\ell_t(a_t) - m_t(a_t))^2 \leq M(\mathcal{E}^* + 1)$.

Proof Let $\mathcal{T}_i = \{t \in T : i \in \mathcal{P}_t\}$ be the time steps where predictor i is active. Then

$$\sum_{t=1}^T (\ell_t(a_t) - m_t(a_t))^2 \leq \sum_{t=1}^T \sum_{i \in \mathcal{P}_t} (\ell_t(a_t) - m_t^i(a_t))^2$$
$$= \sum_{i \in [M]} \sum_{t \in \mathcal{T}_i} (\ell_t(a_t) - m_t^i(a_t))^2 \leq M(\mathcal{E}^* + 1),$$

where the last step uses the fact $\sum_{t \in \mathcal{T}_i} (\ell_t(a_t) - m_t^i(a_t))^2 \leq \mathcal{E}^* + 1$ since the last term in the summation is bounded by one, while the rest cannot exceed $\mathcal{E}^*$ because i has not been removed from the active set yet. ■

Finally, we are ready to prove Theorem 10.

Proof [of Theorem 10] Combining Lemmas 16, 17, and 18, we have

$$\text{Reg} = \mathcal{O}\left(\frac{\ln N}{\eta} + \frac{\eta K^2 \mathcal{E}^*}{\mu} + \frac{\eta K M \mathcal{E}^*}{\mu} + \sqrt{\mu M \mathcal{E}^* T} + \mu T \sigma + \mu^2 T + \frac{M\mathcal{E}^*}{\sigma}\right).$$

With $M' = \max\{K, M\}$, setting

$$\mu = \min\left\{\frac{1}{2}, \sqrt{\frac{d}{T}}\right\}, \sigma = \sqrt{\frac{M\mathcal{E}^*}{\mu T}}, \eta = \sqrt{\frac{\mu \ln N}{K M' \mathcal{E}^*}},$$

gives $\text{Reg} = \mathcal{O}\left(\sqrt{M'\mathcal{E}^*}(dT)^{\frac{1}{4}} + \sqrt{dM'\mathcal{E}^*} + d\right)$. ■

Appendix D. Omitted Details for Stochastic Environments

In this section, we provide omitted details for the stochastic case, including proofs for results with known $\mathcal{E}$ and a single predictor (Section D.1), the adaptive version of Algorithm 2 and its analysis when $\mathcal{E}$ is unknown (Section D.2), and the algorithm and analysis for multiple predictors (Section D.3).

D.1. Proofs of Lemma 7 and Theorems 6 and 8

First, we prove Lemma 7 which certifies the efficiency and (approximate) correctness of the binary search procedure for finding the policy with the smallest Catoni's mean (Algorithm 3).

Proof [of Lemma 7] The fact that the algorithm stops after $\log_2\left(2T\left(\frac{K}{\mu}+1\right)\right) = \mathcal{O}(\ln(KT/\mu))$ iterations is clear due to the initial value of z_{left} and z_{right}, and the precision $1/T$.

To prove the approximate optimality of the output π_t, note that the algorithm maintains the following loop invariants:

$$\min_{\pi\in\Pi}\sum_{s<t}\psi\left(\alpha\left(\widetilde{\ell}_s(\phi_s(\pi(x_s))) - z_{\text{left}}\right)\right) \geq 0$$

and

$$\min_{\pi\in\Pi}\sum_{s<t}\psi\left(\alpha\left(\widetilde{\ell}_s(\phi_s(\pi(x_s))) - z_{\text{right}}\right)\right) \leq 0.$$

Therefore, by the monotonicity of ψ, all policies have Catoni's mean larger than z_{left}, and there exists a policy

$$\operatorname*{argmin}_{\pi\in\Pi}\sum_{s<t}\psi\left(\alpha\left(\widetilde{\ell}_s(\phi_s(\pi(x_s))) - z_{\text{right}}\right)\right)$$

with Catoni's mean smaller than z_{right}. These two facts imply that both $\text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi_t(x_s)))\right\}_{s<t}\right)$ and $\min_{\pi\in\Pi}\text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s<t}\right)$ are between z_{left} and z_{right}, and are thus $1/T$ away from each other since we have $z_{\text{right}} - z_{\text{left}} \leq 1/T$ after the algorithm stops. ■

To prove both Theorem 6 and Theorem 8, we introduce the following notation.

Definition 20 *Denote by* $\mathcal{L}(\pi) \triangleq \mathbb{E}_{(x_t,m_t,\ell_t)\sim\mathcal{D}}[\ell_t(\pi(x_t))]$ *the expected loss of policy* π*, and by* $\overline{\mathcal{L}}(\pi) \triangleq \mathbb{E}_{(x_t,m_t,\ell_t)\sim\mathcal{D}}[\ell_t(\phi_t(\pi(x_t)))]$ *the expected loss of policy* π *after remapping.*

For both theorems we make use of the following lemmas.

Lemma 21 *Algorithm 2 (with either Option I or Option II) ensures*

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\overline{\mathcal{L}}(\pi_t)\right] + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T}.$$

Proof Denote the conditional expectation given the history up to the beginning of time t by $\mathbb{E}_t[\cdot]$. By the choice of a_t we have

$$\begin{aligned}
\mathbb{E}_t\left[\ell_t(a_t)\right] &= (1-\mu)\mathbb{E}_t\left[\ell(\phi_t(\pi_t(x_t))\right] + \mathbb{E}_t\left[\frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)\right] \\
&= \overline{\mathcal{L}}(\pi_t) + \mathbb{E}_t\left[\frac{\mu}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\left(\ell_t(a) - \ell_t(\phi_t(\pi_t(x_t)))\right)\right] \\
&\leq \overline{\mathcal{L}}(\pi_t) + \mu\mathbb{E}_t\left[\sup_{a,a'\in\mathcal{A}_t}|\ell_t(a) - \ell_t(a')|\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \overline{\mathcal{L}}(\pi_t) + \mu\mathbb{E}_t\left[\sup_{a,a'\in\mathcal{A}_t} |\ell_t(a) - m_t(a)| + |m_t(a) - m_t(a')| + |m_t(a') - \ell_t(a')|\right] \\
&\leq \overline{\mathcal{L}}(\pi_t) + \mu\mathbb{E}_t\left[\sigma + 2\|\ell_t - m_t\|_\infty\right] && \text{(by the definition of } \mathcal{A}_t\text{)} \\
&= \overline{\mathcal{L}}(\pi_t) + \mu\sigma + 2\mu\mathbb{E}_t\left[\|\ell_t - m_t\|_\infty\right].
\end{aligned}
$$

Summing over T and applying Cauchy-Schwarz inequality:

$$
\mathbb{E}\left[\sum_{t=1}^T \|\ell_t - m_t\|_\infty\right] \leq \sqrt{T\mathbb{E}\left[\sum_{t=1}^T \|\ell_t - m_t\|_\infty^2\right]} = \sqrt{\mathcal{E}T}
$$

finish the proof. ■

Lemma 22 *Algorithm 2 (with either Option I or Option II) ensures*

$$
T(\overline{\mathcal{L}}(\pi^*) - \mathcal{L}(\pi^*)) \leq \frac{\mathcal{E}}{\sigma}.
$$

Proof The proof is exactly the same as the adversarial case (*cf.* Eq. (12)). First rewrite $\overline{\mathcal{L}}(\pi^*) - \mathcal{L}(\pi^*)$ as $\mathbb{E}\left[\ell_t(\phi_t(\pi^*(x_t))) - \ell_t(\pi^*(x_t))\right]$. When $\pi^*(x_t) \neq \phi_t(\pi^*(x_t))$ we have $\phi_t(\pi^*(x_t)) = a_t^*$, $m_t(a_t^*) \leq m_t(\pi^*(x_t)) - \sigma$, and

$$
\begin{aligned}
&\ell_t(\phi_t(\pi^*(x_t))) - \ell_t(\pi^*(x_t)) \\
&= \ell_t(a_t^*) - m_t(a_t^*) + m_t(a_t^*) - m_t(\pi^*(x_t)) + m_t(\pi^*(x_t)) - \ell_t(\pi^*(x_t)), \\
&\leq 2\|\ell_t - m_t\|_\infty - \sigma \leq \frac{\|\ell_t - m_t\|_\infty^2}{\sigma},
\end{aligned}
$$

where the last step is by the AM-GM inequality. When $\pi^*(x_t) = \phi_t(\pi^*(x_t))$, the above holds trivially. Plugging the definition of $\mathcal{E}$ then finishes the proof. ■

We are now ready to prove Theorems 6 and 8, using different concentrations according to the two different ways of calculating π_t.

Proof [of Theorem 6] First, for any fix π and t, we invoke Lemma 12 with $X_s = \widetilde{\ell}_s(\phi_s(\pi(x_s))) - \overline{\mathcal{L}}(\pi) + \mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}[\min_a m(a)]$ for $s = 1, \ldots, t$, $b = \mathcal{O}(\frac{K}{\mu})$, and $V_t = \mathcal{O}(\frac{K\mathcal{E}t}{\mu T} + \sigma^2 t)$ (see Eq. (4)). Together with a union bound over all t and π, we have with probability at least $1 - 1/T$,

$$
\begin{aligned}
&\left|\frac{1}{t}\sum_{s=1}^t \widetilde{\ell}_s(\phi_s(\pi(x_s))) - \overline{\mathcal{L}}(\pi) + \mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}[\min_a m(a)]\right| \\
&\qquad = \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu T t} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{K\log(NT)}{\mu t}\right)
\end{aligned} \tag{20}
$$

for all $t \in [T]$ and $\pi \in \Pi$. Therefore, we have

$\overline{\mathcal{L}}(\pi_t)$

$$\leq \frac{1}{t}\sum_{s=1}^{t}\widetilde{\ell}_s(\phi_s(\pi_t(x_s))) + \mathbb{E}[\min_a m(a)] + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu T t} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{K\log(NT)}{\mu t}\right)$$

(by Eq. (20))

$$\leq \frac{1}{t}\sum_{s=1}^{t}\widetilde{\ell}_s(\phi_s(\pi^*(x_s))) + \mathbb{E}[\min_a m(a)] + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu T t} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{K\log(NT)}{\mu t}\right)$$

(by the optimality of π_t)

$$\leq \overline{\mathcal{L}}(\pi^*) + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu T t} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{K\log(NT)}{\mu t}\right). \qquad \text{(by Eq. (20))}$$

Combining Lemma 21, the inequality above, and Lemma 22, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\overline{\mathcal{L}}(\pi_t)\right] + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T} \\
&\leq T\overline{\mathcal{L}}(\pi^*) + \mathcal{O}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sum_{t=1}^{T}\sqrt{\left(\frac{K\mathcal{E}}{\mu T t} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{K\log(NT)}{\mu t}\right) \\
&= T\overline{\mathcal{L}}(\pi^*) + \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{d}{\mu}\right) \qquad (21) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] + \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{d}{\mu} + \frac{\mathcal{E}}{\sigma}\right),
\end{aligned}$$

which finishes the proof. ■

Proof [of Theorem 8] First, for any fix π and t, we invoke Lemma 13 with $X_s = \widetilde{\ell}_s(\phi_s(\pi(x_s)))$ for $s = 1, \ldots, t$, $\mu_1 = \cdots = \mu_t = \mu = \overline{\mathcal{L}}(\pi) - \mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}[\min_a m(a)]$, and $V = \mathcal{O}(\frac{K\mathcal{E}}{\mu} + \sigma^2 t)$ (see Eq. (4) for the variance calculation). Together with a union bound over all t and π, and the value of α specified in Algorithm 2, we have with probability at least $1 - 2/T$,

$$\begin{aligned}
&\left|\text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s\leq t}\right) - \overline{\mathcal{L}}(\pi) + \mathbb{E}_{(x,\ell,m)\sim\mathcal{D}}[\min_a m(a)]\right| \\
&\quad = \frac{1}{t}\left(\alpha V + \frac{2\log(NT^2)}{\alpha}\right) = \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu t^2} + \frac{\sigma^2}{t}\right)\log(NT)}\right)
\end{aligned} \qquad (22)$$

for all $t \geq \alpha^2 V + 2\log(NT^2) = 4\log(NT^2)$ and $\pi \in \Pi$. Therefore, we have for $t \geq 4\ln(NT^2)$,

$$\begin{aligned}
&\overline{\mathcal{L}}(\pi_t) \\
&\leq \text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi_t(x_s)))\right\}_{s\leq t}\right) + \mathbb{E}[\min_a m(a)] + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu t^2} + \frac{\sigma^2}{t}\right)\log(NT)}\right) \\
&\qquad \text{(by Eq. (22))} \\
&\leq \text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s(\phi_s(\pi^*(x_s)))\right\}_{s\leq t}\right) + \mathbb{E}[\min_a m(a)] + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu t^2} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{1}{T}\right) \\
&\qquad \text{(by Lemma 7)}
\end{aligned}$$

$$\leq \overline{\mathcal{L}}(\pi^*) + \mathcal{O}\left(\sqrt{\left(\frac{K\mathcal{E}}{\mu t^2} + \frac{\sigma^2}{t}\right)\log(NT)} + \frac{1}{T}\right). \qquad \text{(by Eq. (22))}$$

Combining Lemma 21, the inequality above, and Lemma 22, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(a_t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\overline{\mathcal{L}}(\pi_t)\right] + \mu T\sigma + 2\mu\sqrt{\mathcal{E}T} \\
&\leq T\overline{\mathcal{L}}(\pi^*) + \mathcal{O}\left(4\ln(NT^2) + \mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sum_{t=1}^{T}\sqrt{\left(\frac{K\mathcal{E}}{\mu t^2} + \frac{\sigma^2}{t}\right)\log(NT)}\right) \\
&= T\overline{\mathcal{L}}(\pi^*) + \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT}\right) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\ell_t(\pi^*(x_t))\right] + \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{\mathcal{E}}{\sigma}\right),
\end{aligned}$$

which finishes the proof. ■

D.2. Adaptive Version of Algorithm 2

The pseudocode of the adaptive version of Algorithm 2 in shown in Algorithm 6. To prove its regret guarantee, we make use of the following useful lemmas. The first one shows the concentration of $\widehat{\alpha}_i$ around $\alpha_i = \frac{1}{K}\sum_{a=1}^{K}\Pr\left[|\ell_t(a) - m_t(a)| \in (2^{-i-1}, 2^{-i}]\right]$.

Lemma 23 *Algorithm 6 ensures:*

- *If* $\alpha_i > \frac{360\log T}{B}$*, then with probability at least* $1 - 1/T$*,*

$$\left[\widehat{\alpha}_i - \frac{30\log T}{B}\right]_+ \geq \frac{1}{3}\alpha_i; \tag{23}$$

- *With probability* $1 - 1/T$*,*

$$\left[\widehat{\alpha}_i - \frac{30\log T}{B}\right]_+ \leq \frac{3}{2}\alpha_i. \tag{24}$$

Proof Clearly, $\mathbb{E}[\widehat{\alpha}_i] = \alpha_i$. By Freedman's inequality (Lemma 12), with probability $1 - \frac{1}{T}$,

$$\begin{aligned}
|\widehat{\alpha}_i - \alpha_i| &\leq 2\sqrt{\frac{\alpha_i\log T}{B}} + \frac{\log T}{B} \\
&\leq \frac{\alpha_i}{2} + \frac{30\log T}{B}, \qquad \text{(AM-GM inequality)}
\end{aligned}$$

implying both $\widehat{\alpha}_i \leq \frac{3}{2}\alpha_i + \frac{30\log T}{B}$ and $\frac{\alpha_i}{2} \leq \widehat{\alpha}_i + \frac{30\log T}{B}$. The former implies Eq. (24). Rearranging the latter gives $\widehat{\alpha}_i - \frac{30\log T}{B} \geq \frac{\alpha_i}{2} - \frac{60\log T}{B}$. If $\alpha_i > \frac{360\log T}{B}$, then $\frac{\alpha_i}{2} - \frac{60\log T}{B}$ can further be lower bounded by $\frac{\alpha_i}{2} - \frac{\alpha_i}{6} = \frac{\alpha_i}{3}$, thus proving Eq. (23). ■

The next lemma shows that $\widehat{\mathcal{E}}$ is essentially an underestimator of $\mathcal{E}$.

Algorithm 6 ϵ-GREEDY.VAR: ϵ-Greedy with Variance-adaptivity and Action Remapping

for $t = 1, \dots, B$ **do**
  Draw $a_t \sim \text{Uniform}([K])$.
Let

$$\widehat{\alpha}_i = \frac{1}{B}\sum_{t=1}^{B} \mathbb{1}\left[|\ell_t(a_t) - m_t(a_t)| \in (2^{-i-1}, 2^{-i}]\right],$$
$$\widehat{\mathcal{E}} = T\sum_{i=0}^{\lceil \log_2 T\rceil}\left[\widehat{\alpha}_i - \frac{30\log T}{B}\right]_+ 2^{-2i}.$$

Run Algorithm 2 for the remaining rounds with Option I, $\sigma = \sqrt{\widehat{\mathcal{E}}}(dT)^{-\frac{1}{3}}$, and $\mu = \min\left\{d^{\frac{2}{3}}/T^{\frac{1}{3}}, 1\right\}$.

Lemma 24 *With probability* $1 - \frac{1}{T}$, $\widehat{\mathcal{E}} \le 6\mathcal{E}$.

Proof By Lemma 23 and the definition of $\widehat{\mathcal{E}}$, with probability $1 - \frac{1}{T}$ we have

$$\begin{aligned}
\widehat{\mathcal{E}} &\le T\sum_{i=0}^{\lceil \log_2 T\rceil}\frac{3}{2}\alpha_i 2^{-2i} = 6T\sum_{i=0}^{\lceil \log_2 T\rceil}\alpha_i 2^{-2i-2} \\
&= 6T\sum_{i=0}^{\lceil \log_2 T\rceil}\mathbb{E}\left[\frac{1}{K}\sum_{a=1}^{K}\mathbb{1}\left[|\ell_t(a) - m_t(a)| \in (2^{-i-1}, 2^{-i}]\right]\right]2^{-2i-2} \\
&= 6T\mathbb{E}\left[\frac{1}{K}\sum_{a=1}^{K}\sum_{i=0}^{\lceil \log_2 T\rceil}\mathbb{1}\left[|\ell_t(a) - m_t(a)| \in (2^{-i-1}, 2^{-i}]\right]2^{-2i-2}\right] \\
&\le 6T\mathbb{E}\left[\frac{1}{K}\sum_{a=1}^{K}(\ell_t(a) - m_t(a))^2\right] \\
&\le 6T\mathbb{E}\left[\|\ell_t - m_t\|_\infty^2\right] = 6\mathcal{E}.
\end{aligned}$$

■

The final lemma analyzes the bias due to remapping with the new value of σ, which replaces the role of Lemma 22 when analyzing Algorithm 6.

Lemma 25 *Algorithm 6 ensures:*

$$(T-B)(\overline{\mathcal{L}}(\pi^*) - \mathcal{L}(\pi^*)) = \widetilde{\mathcal{O}}\left(K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}} + K\sqrt{\frac{\mathcal{E}T}{B}}\right).$$

Proof First we bound $(T-B)(\overline{\mathcal{L}}(\pi^*) - \mathcal{L}(\pi^*))$ by $\mathbb{E}\left[\sum_{t=B+1}^{T}(2\|\ell_t - m_t\|_\infty - \sigma)\right]$, following the exact same argument as in the proof of Lemma 22. We then further bound the latter by

$$\mathbb{E}\left[\sum_{t=B+1}^{T}(2\|\ell_t - m_t\|_1 - \sigma)\right]$$

$$\leq \mathbb{E}\left[\sum_{t=B+1}^{T}\left(2\sum_{a=1}^{K}\left(\sum_{i=0}^{\lceil\log_2 T\rceil} 2^{-i}\mathbb{1}[|\ell_t(a)-m_t(a)|\in(2^{-i-1},2^{-i}]]+\mathcal{O}\left(\frac{1}{T}\right)\right)-\sigma\right)\right]$$

(the $\mathcal{O}\left(\frac{1}{T}\right)$ term incurs when all indicators are zero)

$$= \mathbb{E}\left[\sum_{t=B+1}^{T}\left(2K\sum_{i=0}^{\lceil\log_2 T\rceil} 2^{-i}F_t(i)-\sigma\right)\right]+\mathcal{O}(K),$$

where we define $F_t(i) \triangleq \frac{1}{K}\sum_{a=1}^{K}\mathbb{1}[|\ell_t(a)-m_t(a)|\in(2^{-i-1},2^{-i}]]$. We decompose the summation above into two parts:

$$\mathbb{E}\left[\sum_{t=B+1}^{T}\left(K\sum_{i\in\mathcal{I}}2^{-i}F_t(i)-\sigma\right)\right]+\mathbb{E}\left[\sum_{t=B+1}^{T}\left(K\sum_{i\in\overline{\mathcal{I}}}2^{-i}F_t(i)\right)\right]$$

where $\mathcal{I}\triangleq\{i\leq\lceil\log_2 T\rceil:\alpha_i>\frac{360\log T}{B}\}$ and $\overline{\mathcal{I}}\triangleq\{i\leq\lceil\log_2 T\rceil:\alpha_i\leq\frac{360\log T}{B}\}$. We bound the first term as:

$$\begin{aligned}
&\sum_{t=B+1}^{T}\left(K\sum_{i\in\mathcal{I}}2^{-i}F_t(i)-\sigma\right)\\
&\leq\sum_{t=B+1}^{T}\frac{\left(K\sum_{i\in\mathcal{I}}2^{-i}F_t(i)\right)^2}{4\sigma} && \text{(AM-GM inequality)}\\
&\leq\sum_{t=B+1}^{T}\frac{K^2(\log_2 T)\sum_{i\in\mathcal{I}}2^{-2i}F_t(i)^2}{4\sigma} && \text{(Cauchy-Schwarz)}\\
&=\widetilde{\mathcal{O}}\left(K^2\sum_{t=B+1}^{T}\frac{\sum_{i\in\mathcal{I}}2^{-2i}F_t(i)}{\sigma}\right). && (0\leq F_t(i)\leq 1)
\end{aligned}$$

Now we take the expectation conditioned on all history before time B and the high probability event in Lemma 23. Noting that $\mathbb{E}[F_t(i)]=\alpha_i$ and plugging the value of σ, we arrive at

$$\begin{aligned}
\widetilde{\mathcal{O}}\left(\frac{K^2T\sum_{i\in\mathcal{I}}\alpha_i 2^{-2i}}{\sqrt{\widehat{\mathcal{E}}}(dT)^{-\frac{1}{3}}}\right)&\leq\widetilde{\mathcal{O}}\left(\frac{K^2T\sum_{i\in\mathcal{I}}\alpha_i 2^{-2i}}{\sqrt{T\sum_{i\in\mathcal{I}}\alpha_i 2^{-2i}}(dT)^{-\frac{1}{3}}}\right) && \text{(Eq. (23))}\\
&=\widetilde{\mathcal{O}}\left(K^2\sqrt{T\sum_{i\in\mathcal{I}}\alpha_i 2^{-2i}}(dT)^{\frac{1}{3}}\right)\\
&=\widetilde{\mathcal{O}}\left(K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}}\right).
\end{aligned}$$

We continue to bound the second term:

$$\mathbb{E}\left[\sum_{t=B+1}^{T}\left(K\sum_{i\in\overline{\mathcal{I}}}2^{-i}F_t(i)\right)\right]\leq KT\sum_{i\in\overline{\mathcal{I}}}2^{-i}\alpha_i$$

$$\leq K\left(T\sum_{i\in\overline{\mathcal{I}}}\alpha_i\right)^{\frac{1}{2}}\left(T\sum_{i\in\overline{\mathcal{I}}}2^{-2i}\alpha_i\right)^{\frac{1}{2}} \qquad \text{(Cauchy-Schwarz)}$$

$$= \widetilde{\mathcal{O}}\left(K\sqrt{\frac{T}{B}}\times\sqrt{\mathcal{E}}\right). \qquad \text{(definition of } \overline{\mathcal{I}}\text{)}$$

Combining the two terms finishes the proof. ■

Proof [of Theorem 9] By the exact same argument as the proof of Theorem 6 (*cf.* Eq. (21)), we bound the expected loss of the second phase of the algorithm by

$$\mathbb{E}\left[\sum_{t=B+1}^{T}\ell_t(a_t)\right] = (T-B)\overline{\mathcal{L}}(\pi^*) + \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{d}{\mu}\right).$$

Further applying Lemma 25 and bounding the regret of the first phase of the algorithm trivially by B, we have

$$\begin{aligned}
\text{Reg} &= \widetilde{\mathcal{O}}\left(\mu T\sigma + \mu\sqrt{\mathcal{E}T} + \sqrt{\frac{d\mathcal{E}}{\mu}} + \sigma\sqrt{dT} + \frac{d}{\mu} + K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}} + K\sqrt{\frac{\mathcal{E}T}{B}} + B\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\widehat{\mathcal{E}}}(dT)^{\frac{1}{3}} + \sqrt{\mathcal{E}}(dT)^{\frac{1}{6}} + \sqrt{\widehat{\mathcal{E}}}(dT)^{\frac{1}{6}} + (dT)^{\frac{1}{3}} + K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}} + K\sqrt{\frac{\mathcal{E}T}{B}} + B\right) \\
&\qquad \text{(by our choices of } \mu \text{ and } \sigma \text{ defined in Algorithm 6)} \\
&= \widetilde{\mathcal{O}}\left(K^2\sqrt{\mathcal{E}}(dT)^{\frac{1}{3}} + K\sqrt{\frac{\mathcal{E}T}{B}} + B\right). \qquad \text{(Lemma 24)}
\end{aligned}$$

This finishes the proof. ■

D.3. Algorithms and Analysis for Multiple Predictors

In this section, we provide the complete pseudocode of our algorithm for learning with multiple predictors in the stochastic setting (Algorithm 7) and its analysis. As mentioned in Section 5, there are several extra ingredients compared to Algorithm 2. First, just as in Algorithm 5, we maintain an active set of predictors $\mathcal{P}_t$ by bookkeeping the remaining error budget $\widehat{V}_t^i$ for each predictor i. One difference is that the budget starts from $2\mathcal{E}^* + 8\log T$, which takes into account a direct deviation bound. Another difference is that whenever the set $\mathcal{P}_t$ is updated, we discard previous data and run the algorithm from scratch (see Step 3 of Algorithm 7). The reason to do so is to make sure that the data $\{x_s, \ell_s, m_s\}_{s=t_b}^{t}$ are i.i.d., where $m_t(a) = \min_{i\in\mathcal{P}_t} m_t^i(a)$ depends on $\mathcal{P}_t$.

Second, at the beginning of each round, we check if all predictors are consistent to some extent. If not, that is, if there exist two predictors who disagree with each other by $\sigma/3$ on some action, then we simply choose this action deterministically, since this guarantees to reveal which predictor makes a large error for this round. See Step 1 of Algorithm 7. In this case, we set the loss estimators to be zero.

Optimization Problem (to solve for Q_t)

Parameter: $0 < \alpha \leq \min\left\{\sqrt{\frac{\mu T}{65K\mathcal{E}^*}}, \sqrt{\frac{1}{325K\sigma^2}}, \sqrt{\frac{\mu^2}{1300\sigma^2}}, 1\right\}$.

Define: b is such that $t \in [t_b, t_{b+1})$, $\mathcal{P}_{[-1,1]}(X) \triangleq \max\{\min\{X, 1\}, -1\}$ is the projection onto $[-1, 1]$, and

$$\widehat{\mathcal{C}}_t(\pi) = \mathcal{P}_{[-1,1]}\left[\text{Catoni}_\alpha\left(\left\{\widetilde{\ell}_s\left(\phi_s(\pi(x_s))\right)\right\}_{s=t_b}^{t-1}\right)\right],$$
$$\widehat{\text{Reg}}_t(\pi) = \widehat{\mathcal{C}}_t(\pi) - \min_{\pi'}\widehat{\mathcal{C}}_t(\pi').$$

Let Q_t be a solution of Q that satisfies (25), (26), (27).

$$Q \in \Delta_\Pi, \tag{25}$$
$$\sum_{\pi\in\Pi} Q(\pi)\widehat{\text{Reg}}_t(\pi) \leq 240\alpha K\sigma^2 \log T, \tag{26}$$
$$\forall \pi \in \Pi, \qquad \frac{1}{t-t_b}\sum_{s=t_b}^{t-1}\frac{1}{Q^\mu\left(\phi_s(\pi(x_s)) \mid x_s, m_s\right)} \leq 2K + \frac{\widehat{\text{Reg}}_t(\pi)}{120\alpha\sigma^2\log T}, \tag{27}$$

where

$$Q^\mu(a \mid x_s, m_s) = (1-\mu)\sum_{\pi'\in\Pi} Q(\pi')\mathbb{1}[\phi_s(\pi'(x_s)) = a] + \frac{\mu}{|\mathcal{A}_s|}\mathbb{1}[a \in \mathcal{A}_s]. \tag{28}$$

Figure 1: An Optimization Problem for Algorithm 7

Finally, in the case when all predictors are consistent, instead of doing ϵ-greedy as in Algorithm 2, we deploy similar ideas as the minimax optimal algorithm ILOVETOCONBANDITS (Agarwal et al., 2014) to come up with a sparse distribution Q_t over the polices, computed by solving an optimization problem described in Figure 1. At a high level, the optimization problem tries to find a policy with low empirical regret (Eq. (26)) and low empirical variance (Eq. (27)) simultaneously. The difference compared to (Agarwal et al., 2014) is that we apply action remapping as well as (clipped) Catoni's estimators. The fact that this optimization problem can be solved efficiently is by the original arguments in (Agarwal et al., 2014) and the binary search procedure we develop in Algorithm 3 (details omitted).

To analyze the regret of Algorithm 7 and prove Theorem 11, we introduce some definitions and useful lemmas.

Definition 26 *For some epoch b of Algorithm 7 (with a corresponding fixed active set $\mathcal{P}_{t_b}$), define*

$$\mathcal{C}^{(b)}(\pi) \triangleq \mathbb{E}_{(x_t,m_t,\ell_t)\sim\mathcal{D}}\left[B_t\left(\ell_t(\phi_t(\pi(x_t))) - \min_{a\in[K]} m_t(a)\right)\right]$$

Algorithm 7 ILTCB.MARC: ILOVETOCONBANDITS with Action Remapping and Catoni's estimator for Multiple predictors

Parameters: $\mathcal{E}^*, \sigma \in [0,1], \mu \in [0,1]$
Initialization: $\widehat{V}_1^i = 2\mathcal{E}^* + 8\log T$ for all $i \in [M]$.
$\mathcal{P}_1 = [M]$.
$t_1 = 1$
for $b = 1, 2, \ldots$ **do**
  for $t = t_b, \ldots$ **do**
    Receive x_t and m_t^i for all $i \in [M]$.
    Let $m_t(a) = \min_{i \in \mathcal{P}_t} m_t^i(a)$ for all $a \in [K]$.
    Define $a_t^*, \mathcal{A}_t, \phi_t$ according to Eq. (3).

    Step 1. Check if the predictors are consistent, and calculate p_t
    Let $B_t = \mathbb{1}[\forall a \in [K], \forall i,j \in \mathcal{P}_t, |m_t^i(a) - m_t^j(a)| \le \frac{\sigma}{3}]$.
    Let Q_t be a solution of the **Optimization Problem** defined in Figure 1, and define

$$p_t(a) = \begin{cases} \mathbb{1}[a = a'] & \text{if } B_t = 0 \quad (a' \text{ is such that } \exists i,j \in \mathcal{P}_t, |m_t^i(a') - m_t^j(a')| > \frac{\sigma}{3}) \\ Q_t^\mu(a \mid x_t, m_t) & \text{if } B_t = 1 \quad (\text{see Eq.(28) for the definition of } Q_t^\mu(a \mid x_t, m_t)) \end{cases}$$

    Step 2. Choose an action and construct loss estimators
    Sample $a_t \sim p_t$ and receive $\ell_t(a_t)$.
    Define

$$\widetilde{\ell}_t(a) = \left[\frac{(\ell_t(a) - m_t(a))\mathbb{1}[a_t = a]}{p_t(a)} + m_t(a) - m_t(a_t^*) \right] B_t$$

    Step 3. Make updates
    for $i \in \mathcal{P}_t$ **do**
      $\widehat{V}_{t+1}^i \leftarrow \widehat{V}_t^i - (\ell_t(a_t) - m_t^i(a_t))^2$
    $\mathcal{P}_{t+1} = \left\{ i \in \mathcal{P}_t : \widehat{V}_{t+1}^i \ge 0 \right\}$.

    if $\mathcal{P}_{t+1} = \emptyset$ **then**
      $\mathcal{P}_{t+1} \leftarrow [M], \quad \widehat{V}_{t+1}^i \leftarrow 2\mathcal{E}^* + 8\log T, \ \forall i \in [M]$.

    if $\mathcal{P}_{t+1} \neq \mathcal{P}_t$ **then**
      $t_{b+1} = t + 1$
      break

where $t = t_b$, *and*

$$\mathrm{Reg}^{(b)}(\pi) \triangleq \mathcal{C}^{(b)}(\pi) - \min_{\pi' \in \Pi} \mathcal{C}^{(b)}(\pi').$$

Also, define constant $C_0 \triangleq \log(8T^4N^2)$.[4]

Lemma 27 *The Optimization problem defined in Figure 1 admits a solution.*

Proof The proof follows Lemma 1 of (Luo, 2017) (with $\beta = \frac{1}{120\alpha\sigma^2 \log T}$). ■

Lemma 28 *For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T} \|\ell_t - m_t^*\|_\infty^2 \le 2\mathcal{E}^* + 8\log(1/\delta).$$

Proof This is by the definition of $\mathcal{E}^*$ and a direct application of Bernstein's inequality:

$$\begin{aligned}\sum_{t=1}^{T} \|\ell_t - m_t^*\|_\infty^2 &\le \mathcal{E}^* + 4\sqrt{\log(1/\delta)\sum_{t=1}^{T}\mathbb{E}\left[\|\ell_t - m_t^*\|_\infty^4\right]} + 4\log(1/\delta)\\ &\le \mathcal{E}^* + 4\sqrt{\log(1/\delta)\mathcal{E}^*} + 4\log(1/\delta)\\ &\le 2\mathcal{E}^* + 8\log(1/\delta),\end{aligned}$$

where the last step uses AM-GM inequality. ■

Lemma 29 *With probability at least* $1-\frac{1}{T}$*, for all* j*,* $t \in [t_j, t_{j+1})$*, all* $Q \in \Delta_\Pi$*, and all* $\pi \in \Pi$*, the following holds*

$$\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{1}{Q^\mu(\phi_t(\pi(x_t)) \mid x_t, m_t)}\right] \le \frac{6.4}{t-t_j}\sum_{s=t_j}^{t-1}\frac{1}{Q^\mu(\phi_s(\pi(x_s)) \mid x_s, m_s)} + \frac{80C_0}{(t-t_j)\mu^2},$$

where $C_0 = \log\left(8T^4N^2\right)$.

Proof This lemma has appeared several times in the literature such as (Dudik et al., 2011, Theorem 6), (Agarwal et al., 2014, Lemma 10), and (Chen et al., 2019, Lemma 13). Basically this is a consequence of the contexts being i.i.d. generated, and is not related to the algorithm. ■

Lemma 30 *With probability at least* $1-\frac{1}{T}$*, we have for any* π*,* j*, and* $t \in [t_j, t_{j+1})$*,*

$$\mathbb{V}_{(x_t,m_t,\ell_t,a_t)}\left[\widetilde{\ell}_t(\phi_t(\pi(x_t)))\right] \le \frac{4K\mathcal{E}^*}{\mu T} + 20K\sigma^2 + \frac{6.4\widehat{\mathrm{Reg}}_t(\pi)}{120\alpha\log T} + \frac{80\sigma^2 C_0}{(t-t_j)\mu^2}.$$

4. Recall that m_t does not depend on history once $\mathcal{P}_{t_b}$ is fixed, and hence can be treated as jointly i.i.d. along with x_t, ℓ_t over an epoch with a fixed active set.

Proof We prove the lemma by the following sequence of direct calculations:

$$
\begin{aligned}
&\mathbb{V}_{(x_t,m_t,\ell_t,a_t)}\left[\widetilde{\ell}_t(\phi_t(\pi(x_t)))\right] \\
&\le 2\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\left(\frac{(\ell_t(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t)))\mathbb{1}[a_t = \phi_t(\pi(x_t))]}{p_t(\phi_t(\pi(x_t)))}\right)^2 B_t\right] \\
&\qquad + 2\mathbb{E}_{(x_t,m_t)}\left[(m_t(\phi_t(\pi(x_t))) - m_t(a_t^*))^2\right] \\
&\le 4\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\left(\frac{(\ell_t(\phi_t(\pi(x_t))) - m_t^*(\phi_t(\pi(x_t))))\mathbb{1}[a_t = \phi_t(\pi(x_t))]}{p_t(\phi_t(\pi(x_t)))}\right)^2 B_t\right] \\
&\qquad + 4\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\left(\frac{(m_t^*(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))))\mathbb{1}[a_t = \phi_t(\pi(x_t))]}{p_t(\phi_t(\pi(x_t)))}\right)^2 B_t\right] \\
&\qquad + 2\sigma^2 \\
&\le 4\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{(\ell_t(\phi_t(\pi(x_t))) - m_t^*(\phi_t(\pi(x_t))))^2}{p_t(\phi_t(\pi(x_t)))} B_t\right] \\
&\qquad + 4\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{(m_t^*(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))))^2}{p_t(\phi_t(\pi(x_t)))} B_t\right] \\
&\qquad + 2\sigma^2 \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 4\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{(\frac{\sigma}{3})^2}{Q_t^\mu(\phi_t(\pi(x_t)) \mid x_t, m_t)}\right] + 2\sigma^2 \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 2\sigma^2 + \sigma^2\left(\frac{6.4}{t-t_j}\sum_{s=t_j}^{t-1}\frac{1}{Q_t^\mu(\phi_s(\pi(x_s)) \mid x_s, m_s)} + \frac{80C_0}{(t-t_j)\mu^2}\right) && \text{(Lemma 29)} \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 2\sigma^2 + \sigma^2\left(6.4 \times \left(2K + \frac{\widehat{\mathrm{Reg}}_t(\pi)}{120\alpha\sigma^2\log T}\right) + \frac{80C_0}{(t-t_j)\mu^2}\right) && \text{(Eq. (27))} \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 20K\sigma^2 + \frac{6.4\widehat{\mathrm{Reg}}_t(\pi)}{120\alpha\log T} + \frac{80\sigma^2 C_0}{(t-t_j)\mu^2}.
\end{aligned}
$$

■

Lemma 31 *For any π, j, and $t \in [t_j, t_{j+1})$, we have*

$$\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\widetilde{\ell}_t(\phi_t(\pi(x_t)))\right] = \mathcal{C}^{(j)}(\pi).$$

(Recall the definition of $\mathcal{C}^{(j)}(\pi)$ in Definition 26.)

Proof By direct calculation, we have

$$
\begin{aligned}
&\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\widetilde{\ell}_t(\phi_t(\pi(x_t)))\right] \\
&= \mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[\left(\frac{(\ell_t(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))))\mathbb{1}[a_t = \phi_t(\pi(x_t))]}{p_t(\phi_t(\pi(x_t)))} + m_t(\phi_t(\pi(x_t))) - m_t(a_t^*)\right) B_t\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{(x_t,m_t,\ell_t)}\left[(\ell_t(\phi_t(\pi(x_t))) - m_t(\phi_t(\pi(x_t))) + m_t(\phi_t(\pi(x_t))) - m_t(a_t^*))\, B_t\right] \\
&= \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t \ell_t(\phi_t(\pi(x_t))) - B_t \min_a m_t(a)\right] \\
&= \mathcal{C}^{(j)}(\pi),
\end{aligned}
$$

finishing the proof. ■

Lemma 32 *Recall the definition of* $\mathrm{Reg}^{(j)}(\pi)$ *in Definition 26. With probability at least* $1 - \frac{1}{T}$*, we have for any* j *and* $t \in [t_j, t_{j+1})$*,*

$$
\mathrm{Reg}^{(j)}(\pi) \le 2\widehat{\mathrm{Reg}}_t(\pi) + \frac{40\alpha K\mathcal{E}^*}{\mu T} + 200\alpha K\sigma^2 + \frac{800\alpha\sigma^2 C_0 \log T}{(t-t_j)\mu^2} + \frac{20\log(NT^2)\log T}{\alpha(t-t_j)}, \tag{29}
$$

$$
\widehat{\mathrm{Reg}}_t(\pi) \le 2\mathrm{Reg}^{(j)}(\pi) + \frac{40\alpha K\mathcal{E}^*}{\mu T} + 200\alpha K\sigma^2 + \frac{800\alpha\sigma^2 C_0 \log T}{(t-t_j)\mu^2} + \frac{20\log(NT^2)\log T}{\alpha(t-t_j)}. \tag{30}
$$

Proof We first notice that when $t - t_j \le 20\log(NT^2)\log T$, both inequalities hold trivially because the left-hand side is at most $1 \le \frac{20\log(NT^2)\log T}{t-t_j} \le \frac{20\log(NT^2)\log T}{\alpha(t-t_j)}$. Thus we only need to consider the case $t - t_j \ge 20\log(NT^2)\log T$.

We prove them by induction on t. Let $\pi^* = \mathrm{argmin}_{\pi'} \mathcal{C}^{(j)}(\pi')$. Assume (29) and (30) hold for $t_j, \dots, t-1$. By the induction hypothesis and Lemma 30, for any π, the conditional variance of $\widehat{\ell}_s(\phi_s(\pi(x_s)))$ can be upper bounded as follows:

$$
\begin{aligned}
&\mathbb{V}\left[\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right] \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 20K\sigma^2 + \frac{6.4\widehat{\mathrm{Reg}}_s(\pi)}{120\alpha\log T} + \frac{80\sigma^2 C_0}{(s-t_j)\mu^2} \\
&\le \frac{4K\mathcal{E}^*}{\mu T} + 20K\sigma^2 + \frac{80\sigma^2 C_0}{(s-t_j)\mu^2} \\
&\quad + \frac{6.4}{120\alpha\log T}\left(2\mathrm{Reg}^{(j)}(\pi) + \frac{40\alpha K\mathcal{E}^*}{\mu T} + 200\alpha K\sigma^2 + \frac{800\alpha\sigma^2 C_0\log T}{(s-t_j)\mu^2} + \frac{20\log(NT^2)\log T}{\alpha(s-t_j)}\right) \\
&\le \frac{\mathrm{Reg}^{(j)}(\pi)}{8\alpha} + \frac{6.5K\mathcal{E}^*}{\mu T} + 32.5K\sigma^2 + \frac{130\sigma^2 C_0}{(s-t_j)\mu^2} + \frac{3.25\log(NT^2)}{\alpha^2(s-t_j)} \triangleq V_s.
\end{aligned}
$$

Let $V = \sum_{s=t_j}^{t-1} V_s$. We first verify that $t - t_j \ge \alpha^2 V + 2\log(NT^2)$. This can be seen by the following:

$$
\begin{aligned}
&\alpha^2 V + 2\log(NT^2) \\
&\le \alpha^2 \sum_{s=t_j}^{t-1}\left(\frac{\mathrm{Reg}^{(j)}(\pi)}{8\alpha} + \frac{6.5K\mathcal{E}^*}{\mu T} + 32.5K\sigma^2 + \frac{130\sigma^2 C_0}{(s-t_j)\mu^2} + \frac{3.25\log(NT^2)}{\alpha^2(s-t_j)}\right) + 2\log(NT^2) \\
&\le \alpha^2(t-t_j)\left(\frac{1}{8\alpha} + \frac{6.5K\mathcal{E}^*}{\mu T} + 32.5K\sigma^2 + \frac{130\sigma^2 C_0\log T}{(t-t_j)\mu^2} + \frac{3.25\log(NT^2)\log T}{\alpha^2(t-t_j)}\right) + 2\log(NT^2)
\end{aligned}
$$

$$\leq (t-t_j)\left(\frac{\alpha}{8}+\frac{6.5\alpha^2K\mathcal{E}^*}{\mu T}+32.5\alpha^2K\sigma^2+\frac{130\alpha^2\sigma^2C_0\log T}{(t-t_j)\mu^2}+\frac{3.25\log(NT^2)\log T}{(t-t_j)}\right)+2\log(NT^2)$$

$$\leq (t-t_j)\left(\frac{1}{8}+0.1+0.1+\frac{C_0\log T}{10(t-t_j)}+\frac{3.25}{20}\right)+0.1(t-t_j) \qquad \text{(by the constraints on } \alpha\text{)}$$

$$\leq (t-t_j)\left(\frac{1}{8}+0.1+0.1+\frac{16\log(NT^2)\log T}{10\times 20\log(NT^2)\log T}+\frac{3.25}{20}\right)+0.1(t-t_j)$$

(by the definition of C_0)

$$\leq t-t_j. \tag{31}$$

Because of Eq. (31), we are now able to use Lemma 13 for the samples $\left\{\widetilde{\ell}_s(\phi_s(\pi(x_s)))\right\}_{s=t_j}^{t-1}$ with $\delta=\frac{1}{NT^2}$. By Lemmas 13 and 31, we have with probability $1-\delta$ that

$$\begin{aligned}
&\mathrm{Reg}^{(j)}(\pi)=\mathcal{C}^{(j)}(\pi)-\mathcal{C}^{(j)}(\pi^*)\\
&\leq \widehat{\mathcal{C}}_t(\pi)-\widehat{\mathcal{C}}_t(\pi^*)\\
&\quad+\frac{\alpha}{t-t_j}\sum_{s=t_j}^{t-1}\left(\frac{\mathrm{Reg}^{(j)}(\pi)}{8\alpha}+\frac{\mathrm{Reg}^{(j)}(\pi^*)}{8\alpha}+\frac{13K\mathcal{E}^*}{\mu T}+65K\sigma^2+\frac{260\sigma^2C_0}{(s-t_j)\mu^2}+\frac{6.5\log(NT^2)}{\alpha^2(s-t_j)}\right)\\
&\quad+\frac{4\log(NT^2)}{\alpha(t-t_j)}\\
&\leq \widehat{\mathrm{Reg}}_t(\pi)+\frac{1}{8}\mathrm{Reg}^{(j)}(\pi)+\frac{13\alpha K\mathcal{E}^*}{\mu T}+65\alpha K\sigma^2+\frac{260\alpha\sigma^2C_0\log T}{(t-t_j)\mu^2}+\frac{10.5\log(NT^2)\log T}{\alpha(t-t_j)}.
\end{aligned}$$

(using $\mathrm{Reg}^{(j)}(\pi^*)=0$)

Rearranging the above inequality gives

$$\mathrm{Reg}^{(j)}(\pi)\leq\frac{8}{7}\widehat{\mathrm{Reg}}_t(\pi)+\frac{15\alpha K\mathcal{E}^*}{\mu T}+75\alpha K\sigma^2+\frac{300\alpha\sigma^2C_0\log T}{(t-t_j)\mu^2}+\frac{12\log(NT^2)\log T}{\alpha(t-t_j)},$$

proving Eq. (29). Similarly,

$$\begin{aligned}
&\widehat{\mathrm{Reg}}_t(\pi)=\widehat{\mathcal{C}}_t(\pi)-\widehat{\mathcal{C}}_t(\widehat{\pi})\\
&\leq \mathcal{C}_t(\pi)-\mathcal{C}_t(\widehat{\pi})\\
&\quad+\frac{\alpha}{t-t_j}\sum_{s=t_j}^{t-1}\left(\frac{\mathrm{Reg}^{(j)}(\pi)}{8\alpha}+\frac{\mathrm{Reg}^{(j)}(\widehat{\pi})}{8\alpha}+\frac{13K\mathcal{E}^*}{\mu T}+65K\sigma^2+\frac{260\sigma^2C_0}{(s-t_j)\mu^2}+\frac{6.5\log(NT^2)}{\alpha^2(s-t_j)}\right)\\
&\quad+\frac{4\log(NT^2)}{\alpha(t-t_j)}\\
&\leq \frac{9}{8}\mathrm{Reg}^{(j)}(\pi)+\frac{13\alpha K\mathcal{E}^*}{\mu T}+65\alpha K\sigma^2+\frac{260\alpha\sigma^2C_0\log T}{(t-t_j)\mu^2}+\frac{10.5\log(NT^2)\log T}{\alpha(t-t_j)}\\
&\quad+\frac{1}{8}\left(2\widehat{\mathrm{Reg}}_t(\widehat{\pi})+\frac{40\alpha K\mathcal{E}^*}{\mu T}+200\alpha K\sigma^2+\frac{800\alpha\sigma^2C_0\log T}{(t-t_j)\mu^2}+\frac{20\log(NT^2)\log T}{\alpha(t-t_j)}\right)
\end{aligned}$$

(using (29), which we just proved above)

$$\leq \frac{9}{8}\mathrm{Reg}^{(j)}(\pi) + \frac{18\alpha K\mathcal{E}^*}{\mu T} + 90\alpha K\sigma^2 + \frac{360\alpha\sigma^2 C_0 \log T}{(t-t_j)\mu^2} + \frac{13\log(NT^2)\log T}{\alpha(t-t_j)}.$$

(using $\widehat{\mathrm{Reg}}_t(\widehat{\pi}) = 0$)

This proves Eq. (30) and finishes the induction. Recall that we pick $\delta = \frac{1}{NT^2}$. Thus the total failure probability is at most $\frac{1}{NT^2} \times TN \leq \frac{1}{T}$. ■

Lemma 33 *With probability at least* $1 - \frac{1}{T}$*, we have for any* π^**,* j*, and* $t \in [t_j, t_{j+1})$*,*

$$\begin{aligned}
&\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[B_t\ell_t(a_t)\right] \\
&\leq \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right] \\
&\qquad + \widetilde{\mathcal{O}}\left(\frac{\alpha K\mathcal{E}^*}{\mu T} + \alpha K\sigma^2 + \frac{\alpha\sigma^2\log N}{(t-t_j)\mu^2} + \frac{\log N}{\alpha(t-t_j)} + \mu\sqrt{\frac{\mathcal{E}^*}{T}} + \mu\sigma\right).
\end{aligned}$$

Proof By the way a_t is chosen when $B_t = 1$, we have

$$\begin{aligned}
&\mathbb{E}_{(x_t,m_t,\ell_t,a_t)}\left[B_t\ell_t(a_t)\right] \\
&= \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\sum_{a\in[K]} p_t(a)\ell_t(a)\right] \\
&= (1-\mu)\mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\sum_{\pi\in\Pi} Q_t(\pi)\ell_t(\phi_t(\pi(x_t)))\right] + \mu\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{B_t}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)\right]. \qquad (32)
\end{aligned}$$

We continue to bound the first term in Eq. (32) as:

$$\begin{aligned}
&\sum_{\pi\in\Pi} Q_t(\pi)\mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi(x_t)))\right] \\
&= \sum_{\pi\in\Pi} Q_t(\pi)\mathcal{C}^{(j)}(\pi) + \mathbb{E}_{m_t}\left[B_t\min_a m_t(a)\right] && \text{(Definition 26)} \\
&\leq \sum_{\pi\in\Pi} Q_t(\pi)\mathrm{Reg}^{(j)}(\pi) + \mathcal{C}^{(j)}(\pi^*) + \mathbb{E}_{m_t}\left[B_t\min_a m_t(a)\right] && \text{(Definition 26)} \\
&\leq \sum_{\pi\in\Pi} Q_t(\pi)\left(2\widehat{\mathrm{Reg}}_t(\pi) + \frac{40\alpha K\mathcal{E}^*}{\mu T} + 200\alpha K\sigma^2 + \frac{800\alpha\sigma^2 C_0\log T}{(t-t_j)\mu^2} + \frac{20\log(NT^2)\log T}{\alpha(t-t_j)}\right) \\
&\quad + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right] && \text{(Lemma 32 and Definition 26)} \\
&= \mathcal{O}\left(\frac{\alpha K\mathcal{E}^*}{\mu T} + \alpha K\sigma^2 + \frac{\alpha\sigma^2 C_0\log T}{(t-t_j)\mu^2} + \frac{\log(NT^2)\log T}{\alpha(t-t_j)}\right) + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right]. \\
& && \text{(Eq.(26))} \\
&= \widetilde{\mathcal{O}}\left(\frac{\alpha K\mathcal{E}^*}{\mu T} + \alpha K\sigma^2 + \frac{\alpha\sigma^2\log N}{(t-t_j)\mu^2} + \frac{\log N}{\alpha(t-t_j)}\right) + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right].
\end{aligned}$$

The second term in Eq. (32) can be bounded as follows (without the μ factor):

$$\mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{1}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\ell_t(a)B_t\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{1}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\left(\ell_t(a) - m_t(a) + m_t(a) - m_t(\phi_t(\pi^*(x_t)))\right)B_t\right] \\
&\quad + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[\frac{1}{|\mathcal{A}_t|}\sum_{a\in\mathcal{A}_t}\left(m_t(\phi_t(\pi^*(x_t))) - \ell_t(\phi_t(\pi^*(x_t)))\right)B_t\right] + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right] \\
&\le \mathbb{E}_{(x_t,m_t,\ell_t)}\left[2\max_a|\ell_t(a) - m_t(a)|B_t + \sigma\right] + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right] \\
&\le \mathbb{E}_{(x_t,m_t,\ell_t)}\left[2\max_a|\ell_t(a) - m_t^*(a)|B_t + \frac{2\sigma}{3} + \sigma\right] + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right] \\
&\hspace{28em}\text{(definition of } B_t\text{)} \\
&\le 2\left(\sqrt{\frac{\mathcal{E}^*}{T}} + \sigma\right) + \mathbb{E}_{(x_t,m_t,\ell_t)}\left[B_t\ell_t(\phi_t(\pi^*(x_t)))\right]. \qquad \text{(definition of } \mathcal{E}^* \text{ and Jensen's inequality)}
\end{aligned}
$$

Combining these two bounds finishes the proof. ∎

Lemma 34 *Algorithm 7 ensures for any π^*,*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} B_t\ell_t(a_t) - B_t\ell_t(\phi_t(\pi^*(x_t)))\right] \\
&\le \widetilde{\mathcal{O}}\left(\frac{\alpha K\mathcal{E}^*}{\mu} + \alpha TK\sigma^2 + \frac{M\alpha\sigma^2\log N}{\mu^2} + \frac{M\log N}{\alpha} + \mu\sqrt{T\mathcal{E}^*} + T\mu\sigma\right)
\end{aligned}
$$

Proof This is proven by summing the statement of Lemma 33 over t and noticing that with probability $1 - \frac{1}{T}$, there exists a predictor with $\sum_{t=1}^{T}\|\ell_t - m_t^i\|_\infty \le 2\mathcal{E}^* + 8\log T$ and thus there are at most M episodes (see Lemma 28). ∎

Lemma 35 *Algorithm 7 ensures*

$$
\mathbb{E}\left[\sum_{t=1}^{T} B_t\ell_t(\phi_t(\pi^*(x_t))) - B_t\ell_t(\pi^*(x_t))\right] \le \mathcal{O}\left(\frac{\mathcal{E}^*}{\sigma}\right).
$$

Proof The proof is similar to that of Lemma 22:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} B_t\ell_t(\phi_t(\pi^*(x_t))) - B_t\ell_t(\pi^*(x_t))\right] \\
&= \mathbb{E}\left[B_t\sum_{t=1}^{T}\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(\ell_t(a_t^*) - \ell_t(\pi^*(x_t))\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}B_t\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(\ell_t(a_t^*) - m_t(a_t^*) + m_t(a_t^*) - m_t(\pi^*(x_t)) + m_t(\pi^*(x_t)) - \ell_t(\pi^*(x_t))\right)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}B_t\mathbb{1}[\pi^*(x_t)\notin\mathcal{A}_t]\left(-\sigma + 2\|\ell_t - m_t^*\|_\infty + 2\|m_t - m_t^*\|_\infty\right)\right]
\end{aligned}
$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} B_t \mathbb{1}[\pi^*(x_t) \notin \mathcal{A}_t]\left(2\|\ell_t - m_t^*\|_\infty - \frac{\sigma}{3}\right)\right] \qquad \text{(Definition of } B_t\text{)}$$

$$\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\ell_t - m_t^*\|_\infty^2}{\sigma}\right]\right) \qquad \text{(AM-GM)}$$

$$= \mathcal{O}\left(\frac{\mathcal{E}^*}{\sigma}\right).$$

■

Lemma 36 *With probability* $1 - \frac{1}{T}$,

$$\sum_{t=1}^{T}(1 - B_t) \leq \widetilde{\mathcal{O}}\left(\frac{M(1+\mathcal{E}^*)}{\sigma^2}\right).$$

Proof With probability $1 - \frac{1}{T}$, there exists a predictor with $\sum_{t=1}^{T} \|\ell_t - m_t^i\|_\infty \leq 2\mathcal{E}^* + 8\log T$ and thus there are at most M episodes (see Lemma 28). Under this event, every time when $B_t = 0$, there exist $i, i' \in \mathcal{P}_t$ such that $|m_t^i(a_t) - m_t^{i'}(a_t)| \geq \frac{\sigma}{3}$. Therefore, the total budget $\sum_{i\in\mathcal{P}_t} \widehat{V}_i$ decreases by at least $\left(\ell_t(a_t) - m_t^i(a_t)\right)^2 + \left(\ell_t(a_t) - m_t^{i'}(a_t)\right)^2 \geq \frac{1}{2}\left(m_t^i(a_t) - m_t^{i'}(a_t)\right)^2 \geq \frac{\sigma^2}{18}$. Realizing that the initial total budget is $\widetilde{\mathcal{O}}(M(1+\mathcal{E}^*))$ finishes the proof. ■

Finally, we are ready to prove Theorem 11.

Proof [of Theorem 11] Combining Lemmas 34, 35, 36 and picking the optimal parameters in each step, we bound the regret as:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \ell_t(\pi^*(x_t))\right]$$

$$\leq \widetilde{\mathcal{O}}\left(\alpha T K\sigma^2 + \frac{\alpha K\mathcal{E}^*}{\mu} + \frac{M\alpha\sigma^2\log N}{\mu^2} + \frac{M\log N}{\alpha} + \mu\sqrt{T\mathcal{E}^*} + T\mu\sigma + \frac{M(1+\mathcal{E}^*)}{\sigma^2}\right)$$

$$= \widetilde{\mathcal{O}}\left(\sqrt{MK(\log N)T\sigma^2} + \sqrt{\frac{MK(\log N)\mathcal{E}^*}{\mu}} + \frac{M(\log N)\sigma}{\mu} + \mu\sqrt{T\mathcal{E}^*} + T\mu\sigma + \frac{M(1+\mathcal{E}^*)}{\sigma^2}\right)$$
$$+ \widetilde{\mathcal{O}}\left(M\log N\left(\sqrt{\frac{K\mathcal{E}^*}{\mu T}} + \sqrt{K\sigma^2} + \frac{\sigma}{\mu}\right)\right)$$
(picking the optimal α under the constraints of α)

$$= \widetilde{\mathcal{O}}\left(\sqrt{MdT\sigma^2} + \sqrt{\frac{Md\mathcal{E}^*}{\mu}} + \frac{Md\sigma}{\mu} + \mu\sqrt{T\mathcal{E}^*} + T\mu\sigma + \frac{M(1+\mathcal{E}^*)}{\sigma^2}\right)$$
(assume $T \geq M\log N$)

$$= \widetilde{\mathcal{O}}\left(\sqrt{MdT}\sigma + M^{\frac{1}{3}}d^{\frac{1}{3}}\sqrt{\mathcal{E}^*}T^{\frac{1}{6}} + M^{\frac{1}{3}}d^{\frac{1}{3}}\mathcal{E}^{*\frac{1}{3}}\sigma^{\frac{1}{3}}T^{\frac{1}{3}} + \sqrt{Md}\mathcal{E}^{*\frac{1}{4}}T^{\frac{1}{4}}\sqrt{\sigma} + \frac{M(1+\mathcal{E}^*)}{\sigma^2}\right)$$
(picking the optimal μ)

$$= \widetilde{\mathcal{O}}\left((M^2 d)^{\frac{1}{3}}(1+\mathcal{E}^*)^{\frac{1}{3}}T^{\frac{1}{3}} + (Md)^{\frac{1}{3}}\sqrt{\mathcal{E}^*}T^{\frac{1}{6}} + (M^3 d^2)^{\frac{1}{7}}(1+\mathcal{E}^*)^{\frac{3}{7}}T^{\frac{2}{7}} + M^{\frac{3}{5}}d^{\frac{2}{5}}(1+\mathcal{E}^*)^{\frac{2}{5}}T^{\frac{1}{5}}\right)$$

(picking the optimal σ)

$$= \widetilde{\mathcal{O}}\left(M^{\frac{2}{3}}d^{\frac{2}{5}}(1+\mathcal{E}^*)^{\frac{1}{3}}T^{\frac{1}{3}}\right),$$

where the last step uses the fact that we only care about the case when $1+\mathcal{E}^* \leq \sqrt{T}$ to simplify the bound. ■