Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs

Chung-Wei Lee
University of Southern California
leechung@usc.edu

Haipeng Luo
University of Southern California
haipengl@usc.edu

Chen-Yu Wei
University of Southern California
chenyu.wei@usc.edu

Mengxiao Zhang
University of Southern California
mengxiao.zhang@usc.edu

Abstract

We develop a new approach to obtaining high probability regret bounds for online learning with bandit feedback against an adaptive adversary. While existing approaches all require carefully constructing optimistic and biased loss estimators, our approach uses standard unbiased estimators and relies on a simple increasing learning rate schedule, together with the help of logarithmically homogeneous self-concordant barriers and a strengthened Freedman's inequality.

Besides its simplicity, our approach enjoys several advantages. First, the obtained high-probability regret bounds are data-dependent and could be much smaller than the worst-case bounds, which resolves an open problem asked by Neu [31]. Second, resolving another open problem of Bartlett et al. [12] and Abernethy and Rakhlin [1], our approach leads to the first general and efficient algorithm with a high-probability regret bound for adversarial linear bandits, while previous methods are either inefficient or only applicable to specific action sets. Finally, our approach can also be applied to learning adversarial Markov Decision Processes and provides the first algorithm with a high-probability small-loss bound for this problem.

1   Introduction

Online learning with partial information in an adversarial environment, such as the non-stochastic Multi-armed Bandit (MAB) problem [10], is by now a well-studied topic. However, the majority of work in this area has been focusing on obtaining algorithms with sublinear *expected* regret bounds, and these algorithms can in fact be highly unstable and suffer a huge variance. For example, it is known that the classic EXP3 algorithm [10] for MAB suffers *linear regret with a constant probability* (over its internal randomness), despite having nearly optimal expected regret (see [26, Section 11.5, Note 1]), making it a clearly undesirable choice in practice.

To address this issue, a few works develop algorithms with regret bounds that hold with *high probability*, including those for MAB [10, 8, 31], linear bandits [12, 1], and even adversarial Markov Decision Processes (MDPs) [25]. Getting high-probability regret bounds is also the standard way of deriving guarantees against an *adaptive adversary* whose decisions can depend on learner's previous actions. This is especially important for problems such as routing in wireless networks (modeled as linear bandits in [11]) where adversarial attacks can indeed adapt to algorithm's decisions on the fly.

As far as we know, all existing high-probability methods (listed above) are based on carefully constructing *biased* loss estimators that enjoy smaller variance compared to standard unbiased ones. While this principle is widely applicable, the actual execution can be cumbersome; for example, the

scheme proposed in [1] for linear bandits needs to satisfy seven conditions (see their Theorem 4), and other than two examples with specific action sets, no general algorithm satisfying these conditions was provided.

In this work, we develop a new and simple approach to obtaining high-probability regret bounds that works for a wide range of bandit problems with an adaptive adversary (including MAB, linear bandits, MDP, and more). Somewhat surprisingly, in contrast to all previous methods, our approach uses standard *unbiased* loss estimators. More specifically, our algorithms are based on Online Mirror Descent with a self-concordant barrier regularizer [2], a standard approach with expected regret guarantees. The key difference is that we adopt an increasing learning rate schedule, inspired by several recent works using similar ideas for completely different purposes (e.g., [5]). At a high level, the effect of this schedule magically cancels the potentially large variance of the unbiased estimators.

Apart from its simplicity, there are several important advantages of our approach. First of all, our algorithms all enjoy *data-dependent regret bounds*, which could be much smaller than the majority of existing high-probability bounds in the form of $\tilde{\mathcal{O}}(\sqrt{T})$ where T is the number of rounds. As a key example, we provide details for obtaining a particular kind of such bounds called "small-loss" bounds in the form $\tilde{\mathcal{O}}(\sqrt{L^\star})$, where $L^\star \leq T$ is the loss of the benchmark in the regret definition. For MAB and linear bandits, our approach also obtains bounds in terms of the variation of the environment in the vein of [23, 33, 37, 17], resolving an open problem asked by Neu [31].

Second, our approach provides the *first general and efficient* algorithm for adversarial linear bandits (also known as bandit linear optimization) with a high-probability regret guarantee. As mentioned, Abernethy and Rakhlin [1] provide a general recipe for this task but in the end only show concrete examples for two specific action sets. The problem of obtaining a general and efficient approach with regret $\tilde{\mathcal{O}}(\sqrt{T})$ was left open since then. The work of [12] proposes an inefficient but general approach, while the work of [22, 13] develop efficient algorithms for polytopes but with $\tilde{\mathcal{O}}(T^{2/3})$ regret. We not only resolve this long-standing open problem, but also provide improved data-dependent bounds.

Third, our approach is also applicable to learning episodic MDPs with unknown transition, adversarial losses, and bandit feedback. The algorithm is largely based on a recent work [25] on the same problem where a high-probability $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound is obtained. We again develop the first algorithm with a high-probability small-loss bound $\tilde{\mathcal{O}}(\sqrt{L^\star})$ in this setting. The problem in fact shares great similarity with the simple MAB problem. However, none of the existing methods for obtaining small-loss bounds for MAB can be generalized to the MDP setting (at least not in a direct manner) as we argue in Section 4. Our approach, on the other hand, generalizes directly without much effort.

Techniques. Most new techniques of our work is in the algorithm for linear bandits (Section 3), which is based on the SCRIBLE algorithm from the seminal work [2, 3]. The first difference is that we propose to lift the problem from $\mathbb{R}^d$ to $\mathbb{R}^{d+1}$ (where d is the dimension of the problem) and use a *logarithmically homogeneous* self-concordant barrier of the conic hull of the action set (which always exists) as the regularizer for Online Mirror Descent. The nice properties of such a regularizer lead to a smaller variance of the loss estimators. Equivalently, this can be viewed as introducing a new sampling scheme for the original SCRIBLE algorithm in the space of $\mathbb{R}^d$. The second difference is the aforementioned new learning rate schedule, where we increase the learning rate by a small factor whenever the Hessian of the regularizer at the current point is "large" in some sense.

In addition, we also provide a strengthened version of the Freedman's concentration inequality for martingales [21], which is crucial to all of our analysis and might be of independent interest.

Related work. In online learning, there are subtle but important differences and connections between the concept of pseudo-regret, expected regret, and the actual regret, in the context of either oblivious or adaptive adversary. We refer the readers to [8] for detailed related discussions.

While getting expected small-loss regret is common [7, 32, 20, 6, 4, 27], most existing high-probability bounds are of order $\tilde{\mathcal{O}}(\sqrt{T})$. Although not mentioned in the original paper, the idea of implicit exploration from [31] can lead to high-probability small-loss bounds for MAB (see [26, Section 12.3, Note 4]). Lykouris et al. [29] adopt this idea together with a clipping trick to derive small-loss bounds for more general bandit problems with graph feedback. We are not aware of other works with high-probability small-loss bounds in the bandit literature. Note that in [8, Section 6], some high-probability "small-reward" bounds are derived, and they are very different in nature from small-loss

bounds (specifically, the former is equivalent to $\tilde{\mathcal{O}}(\sqrt{T - L^\star})$ in our notation). We are also not aware of high-probability version of other data-dependent regret bounds such as those from [23, 33, 37, 17].

The idea of increasing learning rate was first used in the seminal work of Bubeck et al. [15] for convex bandits. Inspired by this work, Agarwal et al. [5] first combined this idea with the log-barrier regularizer for the problem of "corralling bandits". Since then, this particular combination has proven fruitful for many other problems [37, 28, 27]. We also use it for MAB and MDP, but our algorithm for linear bandits greatly generalizes this idea to any self-concordant barrier.

Structure and notation. In Section 2, we start with a warm-up example on MAB, which is the cleanest illustration on the idea of using increasing learning rates to control the variance of unbiased estimators. Then in Section 3 and Section 4, we greatly generalize the idea to linear bandits and MDPs respectively. We focus on showing small-loss bounds as the main example, and only briefly discuss how to obtain other data-dependent regret bounds, since the ideas are very similar.

We introduce the notation for each setting in the corresponding section, but will use the following general notation throughout the paper: for a positive integer n, $[n]$ represents the set $\{1, \ldots, n\}$ and Δ_n represents the $(n-1)$-dimensional simplex; e_i stands for the i-th standard basis vector and $\mathbf{1}$ stands for the all-one vector (both in an appropriate dimension depending on the context); for a convex function ψ, the associated Bregman divergence is $D_\psi(u, w) = \psi(u) - \psi(w) - \nabla\psi(w)^\top(u - w)$; for a positive definite matrix $M \in \mathbb{R}^{d \times d}$ and a vector $u \in \mathbb{R}^d$, $\|u\|_M \triangleq \sqrt{u^\top M u}$ is the quadratic norm of u with respect to M; $\lambda_{\max}(M)$ denotes the largest eigenvalue of M; $\mathbb{E}_t[\cdot]$ is a shorthand for the conditional expectation given the history before round t; $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms.

2 Multi-armed bandits: an illustrating example

We start with the most basic bandit problem, namely adversarial MAB [10], to demonstrate the core idea of using increasing learning rate to reduce the variance of standard algorithms. The MAB problem proceeds in rounds between a learner and an adversary. For each round $t = 1, \ldots, T$, the learner selects one of the d available actions $i_t \in [d]$, while simultaneously the adversary decides a loss vector $\ell_t \in [0, 1]^d$ with $\ell_{t,i}$ being the loss for arm i. An adaptive adversary can choose ℓ_t based on the learner's previous actions $i_1, \ldots, i_{t-1}$ in an arbitrary way, while an oblivious adversary cannot and essentially decides all ℓ_t's ahead of time (knowing the learner's algorithm). At the end of round t, the learner observes the loss of the chosen arm ℓ_{t,i_t} and nothing else. The standard measure of the learner's performance is the regret, defined as $\text{Reg} = \sum_{t=1}^T \ell_{t,i_t} - \min_{i \in [d]} \sum_{t=1}^T \ell_{t,i}$, that is, the difference between the total loss of the learner and that of the best fixed arm in hindsight.

A standard framework to solve this problem is Online Mirror Descent (OMD), which at time t samples i_t from a distribution w_t, updated in the following recursive form: $w_{t+1} = \operatorname{argmin}_{w \in \Delta_d} \langle w, \widehat{\ell}_t \rangle + D_{\psi_t}(w, w_t)$, where ψ_t is the regularizer and $\widehat{\ell}_t$ is an estimator for ℓ_t. The standard estimator is the importance-weighted estimator: $\widehat{\ell}_{t,i} = \ell_{t,i} \mathbb{1}\{i_t = i\}/w_{t,i}$, which is clearly unbiased. Together with many possible choices of the regularizer (e.g., the entropy regularizer recovering EXP3 [10]), this ensures (nearly) optimal expected regret bound $\mathbb{E}[\text{Reg}] = \tilde{\mathcal{O}}(\sqrt{dT})$ against an oblivious adversary.

To obtain high-probability regret bounds (and also as a means to deal with adaptive adversary), various more sophisticated loss estimators have been proposed. Indeed, the key challenge in obtaining high-probability bounds lies in the potentially large variance of the unbiased estimators: $\mathbb{E}_t[\widehat{\ell}_{t,i}^2] = \ell_{t,i}^2/w_{t,i}$ is huge if $w_{t,i}$ is small. The idea of all existing approaches to addressing this issue is to introduce a slight bias to the estimator, making it an optimistic underestimator of ℓ_t with lower variance (see e.g., [10, 8, 31]). Carefully balancing the bias and variance, these algorithms achieve $\text{Reg} = \tilde{\mathcal{O}}(\sqrt{dT \ln(d/\delta)})$ with probability at least $1 - \delta$ against an adaptive adversary.

Our algorithm. In contrast to all these existing approaches, we next show that, perhaps surprisingly, using the standard unbiased estimator can also lead to the same (in fact, an even better) high-probability regret bound. We start by choosing a particular regularizer called log-barrier with time-varying and individual learning rate $\eta_{t,i}$: $\psi_t(w) = \sum_{i=1}^d \frac{1}{\eta_{t,i}} \ln \frac{1}{w_i}$, which is a self-concordant barrier for the positive orthant [30] and has been used for MAB in several recent works [20, 5, 16, 37, 17]. As mentioned in Section 1, the combination of log-barrier and a particular increasing learning rate

3

Algorithm 1 OMD with log-barrier and increasing learning rates for Multi-armed Bandits

---

Input: initial learning rate η.

Define: increase factor $\kappa = e^{\frac{1}{\ln T}}$, truncated simplex $\Omega = \left\{ w \in \Delta_d : w_i \geq \frac{1}{T}, \forall i \in [d] \right\}$.

Initialize: for all $i \in [d]$, $w_{1,i} = 1/d, \rho_{1,i} = 2d, \eta_{1,i} = \eta$.

for $t = 1, 2, \ldots, T$ **do**

> Sample $i_t \sim w_t$, observe ℓ_{t,i_t}, and construct estimator $\widehat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{i_t=i\}}{w_{t,i}}$ for all $i \in [d]$.
>
> Compute $w_{t+1} = \operatorname{argmin}_{w \in \Omega} \left\langle w, \widehat{\ell}_t \right\rangle + D_{\psi_t}(w, w_t)$ where $\psi_t(w) = \sum_{i=1}^d \frac{1}{\eta_{t,i}} \ln \frac{1}{w_i}$.
>
> **for** $i \in [d]$ **do**
>
> > **if** $\frac{1}{w_{t+1,i}} > \rho_{t,i}$ **then** set $\rho_{t+1,i} = \frac{2}{w_{t+1,i}}, \eta_{t+1,i} = \eta_{t,i}\kappa$;
> >
> > **else** set $\rho_{t+1,i} = \rho_{t,i}, \eta_{t+1,i} = \eta_{t,i}$.

---

schedule has been proven powerful for many different problems since the work of [5], which we also apply here. Specifically, the learning rates start with a fixed value $\eta_{1,i} = \eta$ for all arm $i \in [d]$, and every time the probability of selecting an arm i is too small, in the sense that $1/w_{t+1,i} > \rho_{t,i}$ for some threshold $\rho_{t,i}$ (starting with $2d$), we set the new threshold to be $2/w_{t+1,i}$ and increase the corresponding learning rate $\eta_{t,i}$ by a small factor κ.

The complete pseudocode is shown in Algorithm 1. The only slight difference compared to the algorithm of [5] is that instead of enforcing a $1/T$ amount of uniform exploration explicitly (which makes sure that each learning rate is increased by a most $\mathcal{O}(\ln T)$ times), we directly perform OMD over a truncated simplex $\Omega = \{w \in \Delta_d : w_i \geq 1/T, \forall i \in [d]\}$, making the analysis cleaner.

As explained in [5], increasing the learning rate in this way allows the algorithm to quickly realize that some arms start to catch up even though they were underperforming in earlier rounds, which is also the hardest case in our context of obtaining high-probability bounds because these arms have low-quality estimators at some point. At a technical level, this effect is neatly presented through a *negative* term in the regret bound, which we summarize below.

Lemma 2.1. *Algorithm 1 ensures* $\sum_{t=1}^T \ell_{t,i_t} - \sum_{t=1}^T \left\langle u, \widehat{\ell}_t \right\rangle \leq \mathcal{O}\left(\frac{d \ln T}{\eta} + \eta \sum_{t=1}^T \ell_{t,i_t} \right) - \frac{\langle \rho_T, u \rangle}{10 \eta \ln T}$ *for any* $u \in \Omega$.

The important part is the last negative term involving the last threshold ρ_T whose magnitude is large whenever an arm has a small sampling probability at some point over the T rounds. This bound has been proven in previous works such as [5] (see a proof in Appendix A.2), and next we use it to show that the algorithm in fact enjoys a high-probability regret bound, which is not discovered before.

Indeed, comparing Lemma 2.1 with the definition of regret, one sees that as long as we can relate the estimated loss of the benchmark $\sum_t \left\langle u, \widehat{\ell}_t \right\rangle$ with its true loss $\sum_t \left\langle u, \ell_t \right\rangle$, then we immediately obtain a regret bound by setting $u = (1 - \frac{d}{T})e_{i^\star} + \frac{1}{T}\mathbf{1} \in \Omega$ where $i^\star = \operatorname{argmin}_i \sum_t \ell_{t,i}$ is the best arm. A natural approach is to apply standard concentration inequality, in particular Freedman's inequality [21], to the martingale difference sequence $\left\langle u, \widehat{\ell}_t - \ell_t \right\rangle$. The deviation from Freedman's inequality is in terms of the variance of $\left\langle u, \widehat{\ell}_t \right\rangle$, which in turn depends on $\sum_i u_i / w_{t,i}$. As explained earlier, the negative term is exactly related to this and can thus cancel the potentially large variance!

One caveat, however, is that the deviation from Freedman's inequality also depends on a *fixed* upper bound of the random variable $\left\langle u, \widehat{\ell}_t \right\rangle \leq \sum_i u_i / w_{t,i}$, which could be as large as T (since $w_{t,i} \geq 1/T$) and ruin the bound. If the dependence on such a fixed upper bound could be replaced with the (random) upper bound $\sum_i u_i / w_{t,i}$, then we could again use the negative term to cancel this dependence. Fortunately, since $\sum_i u_i / w_{t,i}$ is measurable with respect to the σ-algebra generated by everything *before* round t, we are indeed able to do so. Specifically, we develop the following strengthened version of Freedman's inequality, which might be of independent interest.

Theorem 2.2. *Let $X_1, \ldots, X_T$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$ such that $\mathbb{E}[X_t|\mathcal{F}_t] = 0$. Suppose $B_t \in [1, b]$ for a fixed constant b is $\mathcal{F}_t$-measurable and such that $X_t \leq B_t$ holds almost surely. Then with probability at least $1 - \delta$ we have $\sum_{t=1}^T X_t \leq C\left(\sqrt{8V \ln(C/\delta)} + 2B^\star \ln(C/\delta) \right)$, where $V = \max\left\{ 1, \sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_t] \right\}$, $B^\star = \max_{t \in [T]} B_t$, and $C = \lceil \log(b) \rceil \lceil \log(b^2 T) \rceil$.*

4

This strengthened Freedman's inequality essentially recovers the standard one when B_t is a fixed quantity. In our application, B_t is exactly $\langle \rho_t, u \rangle$ which is $\mathcal{F}_t$-measurable. With the help of this concentration result, we are now ready to show the high-probability guarantee of Algorithm 1.

Theorem 2.3. *Algorithm 1 with a suitable choice of η ensures that with probability at least $1 - \delta$, $\mathrm{Reg} = \widetilde{\mathcal{O}}\big(\sqrt{dL^\star \ln(d/\delta)} + d\ln(d/\delta)\big)$, where $L^\star = \min_i \sum_{t=1}^T \ell_{t,i}$ is the loss of the best arm.*

The proof is a direct combination of Lemma 2.1 and Theorem 2.2 and can be found in Appendix A.3. Our high-probability guarantee is of the same order $\tilde{\mathcal{O}}(\sqrt{dT \ln(d/\delta)})$ as in previous works [10, 8] since $L^\star = \mathcal{O}(T)$. However, as long as $L^\star = o(T)$ (that is, the best arm is of high quality), our bound becomes much better. This kind of high-probability small-loss bounds appears before (e.g., [29]). Nevertheless, in Section 4 we argue that only our approach can directly generalize to learning MDPs.

Finally, we remark that the same algorithm can also obtain other data-dependent regret bounds by changing the estimator to $\widehat{\ell}_{t,i} = (\ell_{t,i} - m_{t,i})\mathbb{1}\{i_t = i\}/w_{t,i} + m_{t,i}$ for some optimistic prediction m_t. We refer the reader to [37] for details on how to set m_t in terms of observed data and what kind of bounds this leads to, but the idea of getting the high-probability version is completely the same as what we have illustrated here. This resolves an open problem mentioned in [31, Section 5].

3 Generalization to adversarial linear bandits

Next, we significantly generalize our approach to adversarial linear bandits, which is the main algorithmic contribution of this work. Linear bandits generalize MAB from the simplex decision set Δ_d to an arbitrary convex body $\Omega \subseteq \mathbb{R}^d$. For each round $t = 1, \ldots, T$, the learner selects an action $\widetilde{w}_t \in \Omega$ while simultaneously the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$, assumed to be normalized such that $\max_{w \in \Omega} |\langle w, \ell_t \rangle| \leq 1$. Again, an adaptive adversary can choose ℓ_t based on the learner's previous actions, while an oblivious adversary cannot. At the end of round t, the learner suffers and only observes loss $\langle \widetilde{w}_t, \ell_t \rangle$. The regret of the learner is defined as $\mathrm{Reg} = \max_{u \in \Omega} \sum_{t=1}^T \langle \widetilde{w}_t - u, \ell_t \rangle$, which is the difference between the total loss of the learner and that of the best fixed action within Ω. Linear bandits subsume many other well-studied problems such as online shortest path for network routing, online matching, and other combinatorial bandit problems (see e.g., [9, 18]).

The seminal work of Abernethy et al. [2] develops the first general and efficient linear bandit algorithm (called SCRIBLE in its journal version [3]) with expected regret $\tilde{\mathcal{O}}(d\sqrt{\nu T})$ (against an oblivious adversary), which uses a ν-self-concordant barrier as the regularizer for OMD. It is known that any convex body in $\mathbb{R}^d$ admits a ν-self-concordant barrier with $\nu = \mathcal{O}(d)$ [30]. The minimax regret of this problem is known to be of order $\tilde{\mathcal{O}}(d\sqrt{T})$ [19, 14], but efficiently achieving this bound (in expectation) requires a log-concave sampler and a volumetric spanner of Ω [24].

High-probability bounds for linear bandits are very scarce, especially for a general decision set Ω. In [12], an algorithm with high-probability regret $\tilde{\mathcal{O}}(\sqrt{d^3 T} \ln(1/\delta))$ was developed, but it cannot be implement efficiently. In [1], a general recipe was provided, but seven conditions need to be satisfied to arrive at a high-probability guarantee, and only two concrete examples were shown (when Ω is the simplex or the Euclidean ball). We propose a new algorithm based on SCRIBLE, which is the first general and efficient linear bandit algorithm with a high-probability regret guarantee, resolving the problem left open since the work of [12, 1].

Issues of SCRIBLE. To introduce our algorithm, we first review SCRIBLE. As mentioned, it is also based on OMD and maintains a sequence $w_1, \ldots, w_T \in \Omega$ updated as $w_{t+1} = \mathrm{argmin}_{w \in \Omega} \langle w, \widehat{\ell}_t \rangle + \frac{1}{\eta} D_\psi(w, w_t)$ where $\widehat{\ell}_t$ is an estimator for ℓ_t, η is some learning rate, and importantly, ψ is a ν-self-concordant barrier for Ω which, again, always exists. Due to space limit, we defer the definition and properties of self-concordant barriers to Appendix B.2. To incorporate exploration, the actual point played by the algorithm at time t is $\widetilde{w}_t = w_t + H_t^{-1/2} s_t$ where $H_t = \nabla^2 \psi(w_t)$ and s_t is uniformly randomly sampled from the d-dimensional unit sphere $\mathbb{S}^d$.[1] The point $\widetilde{w}_t$ is on the boundary of the *Dikin ellipsoid* centered at w_t (defined as $\{w : \|w - w_t\|_{H_t} \leq 1\}$)

---

[1]In fact, s_t can be sampled from any orthonormal basis of $\mathbb{R}^d$ together with their negation. For example, in the original SCRIBLE, the eigenbasis of H_t is used as this orthonormal basis. The version of sampling from a unit sphere first appears in [36], which works more generally for convex bandits.

Algorithm 2 SCRIBLE with lifting and increasing learning rates

---

Input: decision set $\Omega \subseteq \mathbb{R}^d$, a ν-self-concordant barrier ψ for Ω, initial learning rate η.

Define: increase factor $\kappa = e^{\frac{1}{100d\ln(\nu T)}}$, normal barrier $\Psi(\boldsymbol{w}) = \Psi(w, b) = 400\left(\psi\left(\frac{w}{b}\right) - 2\nu \ln b\right)$.

Initialize: $w_1 = \operatorname{argmin}_{w \in \Omega} \psi(w)$, $\boldsymbol{w}_1 = (w_1, 1)$, $\boldsymbol{H}_1 = \nabla^2\Psi(\boldsymbol{w_1})$, $\eta_1 = \eta$, $\mathcal{S} = \{1\}$.

Define: shrunk lifted decision set $\boldsymbol{\Omega}' = \{\boldsymbol{w} = (w, 1) : w \in \Omega, \pi_{w_1}(w) \leq 1 - \frac{1}{T}\}$.

1 **for** $t = 1, 2, \ldots, T$ **do**

2 $\quad$ Uniformly at random sample $\boldsymbol{s}_t$ from $\left(\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{e}_{d+1}\right)^{\perp} \cap \mathbb{S}^{d+1}$.

3 $\quad$ Compute $\widetilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{s}_t \triangleq (\widetilde{w}_t, 1)$.

4 $\quad$ Play $\widetilde{w}_t$, observe loss $\langle \widetilde{w}_t, \ell_t\rangle$, and construct loss estimator $\widehat{\boldsymbol{\ell}}_t = d\langle \widetilde{w}_t, \ell_t\rangle \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t$.

5 $\quad$ Compute $\boldsymbol{w}_{t+1} = \operatorname{argmin}_{\boldsymbol{w} \in \boldsymbol{\Omega}'} \langle \boldsymbol{w}, \widehat{\boldsymbol{\ell}}_t\rangle + D_{\Psi_t}(\boldsymbol{w}, \boldsymbol{w}_t)$, where $\Psi_t = \frac{1}{\eta_t}\Psi$.

6 $\quad$ Compute $\boldsymbol{H}_{t+1} = \nabla^2\Psi(\boldsymbol{w}_{t+1})$.

7 $\quad$ **if** $\lambda_{\max}(\boldsymbol{H}_{t+1} - \sum_{\tau \in \mathcal{S}}\boldsymbol{H}_\tau) > 0$ **then** $\mathcal{S} \leftarrow \mathcal{S} \cup \{t+1\}$ and set $\eta_{t+1} = \eta_t\kappa$;

8 $\quad$ **else** set $\eta_{t+1} = \eta_t$.

---

and is known to be always within Ω. Finally, the estimator $\widehat{\ell}_t$ is constructed as $d\langle \widetilde{w}_t, \ell_t\rangle H_t^{1/2}s_t$, which can be computed using only the feedback $\langle \widetilde{w}_t, \ell_t\rangle$ and is unbiased as one can verify.

The analysis of [2] shows the following bound related to the loss estimators: $\sum_{t=1}^T \langle w_t - u, \widehat{\ell}_t\rangle \leq \tilde{\mathcal{O}}(\frac{\nu}{\eta} + \eta d^2 T)$ for any $u \in \Omega$ (that is not too close to the boundary). Since $\mathbb{E}_t[\langle w_t - u, \widehat{\ell}_t\rangle] = \mathbb{E}_t[\langle \widetilde{w}_t - u, \ell_t\rangle]$, this immediately yields an expected regret bound (for an oblivious adversary). However, to obtain a high-probability bound, one needs to consider the deviation of $\sum_{t=1}^T \langle w_t - u, \widehat{\ell}_t\rangle$ from $\sum_{t=1}^T \langle \widetilde{w}_t - u, \ell_t\rangle$. Applying our strengthened Freedman's inequality (Theorem 2.2) with $X_t = \langle \widetilde{w}_t - u, \ell_t\rangle - \langle w_t - u, \widehat{\ell}_t\rangle$, with some direct calculations one can see that both the variance term V and the range term $B^\star$ from the theorem are related to $\max_t \|w_t\|_{H_t}$ and $\max_t \|u\|_{H_t}$, both of which can be prohibitively large. We next discuss how to control each of these two terms, leading to the two new ideas of our algorithm (see Algorithm 2).

Controlling $\|w_t\|_{H_t}$. Readers who are familiar with self-concordant functions would quickly realize that $\|w_t\|_{H_t} = \sqrt{w_t^\top \nabla^2 \psi(w_t)w_t}$ is simply $\sqrt{\nu}$ provided that ψ is also *logarithmically homogeneous*. A logarithmically homogeneous self-concordant barrier is also called a *normal barrier* (see Appendix B.2 for formal definitions and related properties). However, normal barriers are only defined for cones instead of convex bodies.

Inspired by this fact, we propose to *lift the problem to* $\mathbb{R}^{d+1}$. To make the notation clear, we use bold letters for vectors in $\mathbb{R}^{d+1}$ and matrices in $\mathbb{R}^{(d+1)\times(d+1)}$. The lifting is done by operating over a lifted decision set $\boldsymbol{\Omega} = \{\boldsymbol{w} = (w, 1) \in \mathbb{R}^{d+1} : w \in \Omega\}$, that is, we append a dummy coordinate with value 1 to all actions. The conic hull of this set is $\mathcal{K} = \{(w, b) : w \in \mathbb{R}^d, b \geq 0, \frac{1}{b}w \in \Omega\}$. We then perform OMD over the lifted decision set but with a normal barrier defined over the cone $\mathcal{K}$ as the regularizer to produce the sequence $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ (Line 5). In particular, using the original regularizer ψ we construct the normal barrier as: $\Psi(w, b) = 400\left(\psi\left(\frac{w}{b}\right) - 2\nu \ln b\right)$.[2] Indeed, Proposition 5.1.4 of [30] asserts that this is a normal barrier for $\mathcal{K}$ with self-concordant parameter $\mathcal{O}(\nu)$.

So far nothing really changes since $\Psi(w, 1) = 400\psi(w)$. However, the key difference is in the way we sample the point $\widetilde{\boldsymbol{w}}_t$. If we still follow SCRIBLE to sample from the Dikin ellipsoid centered at $\boldsymbol{w}_t$, it is possible that the sampled point leaves $\boldsymbol{\Omega}$. To avoid this, it is natural to sample only the intersection of the Dikin ellipsoid and $\boldsymbol{\Omega}$ (again an ellipsoid). Algebraically, this means setting $\widetilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \boldsymbol{H}_t^{-1/2}\boldsymbol{s}_t$ where $\boldsymbol{H}_t = \nabla^2\Psi(\boldsymbol{w}_t)$ and $\boldsymbol{s}_t$ is sampled uniformly at random from $(\boldsymbol{H}_t^{-1/2}\boldsymbol{e}_{d+1})^{\perp} \cap \mathbb{S}^{d+1}$ ($v^{\perp}$ is the space orthogonal to v). Indeed, since $\boldsymbol{s}_t$ is orthogonal to $\boldsymbol{H}_t^{-1/2}\boldsymbol{e}_{d+1}$, the last coordinate of $\boldsymbol{H}_t^{-1/2}\boldsymbol{s}_t$ is zero, making $\widetilde{\boldsymbol{w}}_t = (\widetilde{w}_t, 1)$ stay in $\boldsymbol{\Omega}$. See Line 2 and Line 3. To sample $\boldsymbol{s}_t$ efficiently, one can either sample a vector uniformly randomly from $\mathbb{S}^{d+1}$,

---

[2]Our algorithm works with any normal barrier, not just this particular one. We use this particular form to showcase that we only require a self-concordant barrier of the original set Ω, exactly the same as SCRIBLE.
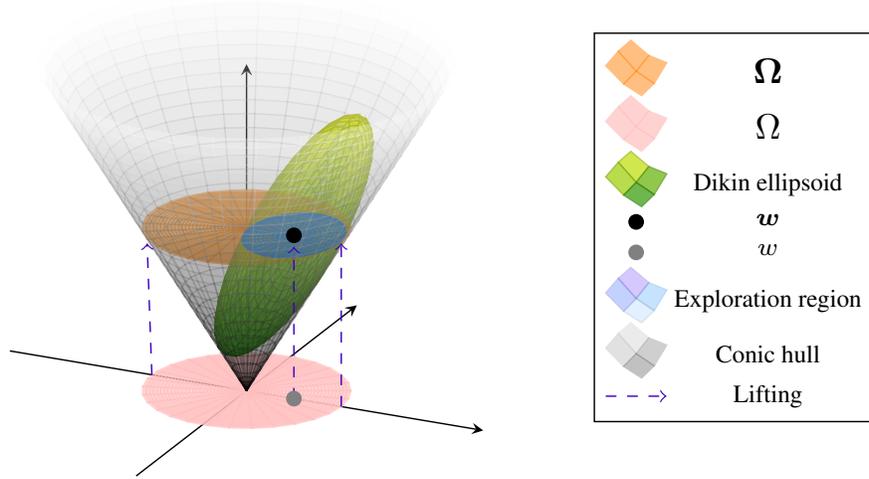
Figure 1: An illustration of the concept of lifting, the conic hull, and the Dikin ellipsoid. In this example d is 2, and the pink disk at the bottom is the original decision set Ω. The gray dot w is a point in Ω. In Algorithm 2, we lift the problem from $\mathbb{R}^2$ to $\mathbb{R}^3$, and obtain the lifted, orange, decision set $\boldsymbol{\Omega}$. For example, w is lifted to the black dot $\boldsymbol{w} = (w, 1)$. Then we construct the conic hull of the lifted decision set, that is, the gray cone, and construct a normal barrier for this conic hull. By Lemma B.1, the Dikin ellipsoid centered at $\boldsymbol{w}$ of this normal barrier (the green ellipsoid), is alway within the cone. In Algorithm 2, if $\boldsymbol{w}$ is the OMD iterate, we explore and play an action within the intersection of $\boldsymbol{\Omega}$ and the Dikin ellipsoid centered at $\boldsymbol{w}$, that is, the (boundary of) the blue ellipse.

project it onto the subspace perpendicular to $\boldsymbol{H}_t^{-1/2}\boldsymbol{e}_{d+1}$, and then normalize; or sample a vector s_t uniformly randomly from $\mathbb{S}^d$, then normalize $\boldsymbol{H}_t^{\frac{1}{2}}(s_t^\top, 0)^\top$ to obtain $\boldsymbol{s}_t$.

Finally, after playing $\widetilde{w}_t$ and observing $\langle \widetilde{w}_t, \ell_t \rangle$, we construct the loss estimator the same way as SCRIBLE: $\widehat{\boldsymbol{\ell}}_t = d \langle \widetilde{w}_t, \ell_t \rangle \boldsymbol{H}_t^{1/2} \boldsymbol{s}_t$ (Line 4). Lemma B.9 shows that the first d coordinates of $\widehat{\boldsymbol{\ell}}_t$ is indeed an unbiased estimator of ℓ_t. This makes the entire analysis of SCRIBLE hold in $\mathbb{R}^{d+1}$, but now the key term $\|\boldsymbol{w}_t\|_{\boldsymbol{H}_t}$ we want to control is *exactly* $20\sqrt{2\nu}$ (see Lemma B.5 and Lemma B.6)!

We provide an illustration of the lifting idea in Figure 1. One might ask whether this lifting is necessary; indeed, one can also spell out the algorithm in $\mathbb{R}^d$ (see Appendix B.1). Importantly, compared to SCRIBLE, the key difference is still that the sampling scheme has changed: the sampled point is not necessarily on the Dikin ellipsoid with respect to ψ. In other words, another view of our algorithm is that it is SCRIBLE with a new sampling scheme. We emphasize that, however, it is important (or at least much cleaner) to perform the analysis in $\mathbb{R}^{d+1}$. In fact, even in Algorithm 1 for MAB, similar lifting implicitly happens already since Δ_d is a convex body in dimension $d-1$ instead of d, and log-barrier is indeed a canonical normal barrier for the positive orthant.

Controlling $\|u\|_{H_t}$. Next, we discuss how to control the term $\|u\|_{H_t}$, or rather $\|\boldsymbol{u}\|_{\boldsymbol{H}_t}$ after the lifting. This term is the analogue of $\sum_i \frac{u_i}{w_{t,i}}$ for the case of MAB, and our goal is again to cancel it with the negative term introduced by increasing the learning rate. Indeed, a closer look at the OMD analysis reveals that increasing the learning rate at the end of time t brings a negative term involving $-D_\Psi(\boldsymbol{u}, \boldsymbol{w}_{t+1})$ in the regret bound. In Lemma B.13, we show that this negative term is upper bounded by $-\|\boldsymbol{u}\|_{\boldsymbol{H}_{t+1}} + 800\nu \ln(800\nu T + 1)$, making the canceling effect possible.

It just remains to figure out when to increase the learning rate and how to make sure we only increase it logarithmic (in T) times as in the case for MAB. Borrowing ideas from Algorithm 1, intuitively one should increase the learning rate only when $\boldsymbol{H}_t$ is "large" enough, but the challenge is how to

measure this quantitatively. Only looking at the eigenvalues of $\boldsymbol{H}_t$, a natural idea, does not work as it does not account for the fact that the directions of eigenvectors are changing over time.

Instead, we propose the following condition: at the end of time t, increase the learning rate by a factor of κ if $\lambda_{\max}(\boldsymbol{H}_{t+1} - \sum_{\tau \in \mathcal{S}} \boldsymbol{H}_\tau) > 0$, with $\mathcal{S}$ containing all the previous time steps prior to time t where the learning rate was increased (Line 7). First, note that this condition makes sure that we always have enough negative terms to cancel $\max_t \|\boldsymbol{u}\|_{\boldsymbol{H}_t}$. Indeed, suppose t is the time with the largest $\|\boldsymbol{u}\|_{\boldsymbol{H}_{t+1}}$. If we have increased the learning rate at time t, then the introduced negative term exactly matches $\|\boldsymbol{u}\|_{\boldsymbol{H}_{t+1}}$ as mentioned above; otherwise, the condition did not hold and by definition we have $\|\boldsymbol{u}\|_{\boldsymbol{H}_{t+1}} \leq \sqrt{\sum_{\tau \in \mathcal{S}} \|\boldsymbol{u}\|_{\boldsymbol{H}_s}^2} \leq \sum_{\tau \in \mathcal{S}} \|\boldsymbol{u}\|_{\boldsymbol{H}_\tau}$, meaning that the negative terms introduced in previous steps are already enough to cancel $\|\boldsymbol{u}\|_{\boldsymbol{H}_{t+1}}$.

Second, in Lemma B.12 we show that this schedule indeed makes sure that the learning rate is increased by only $\tilde{\mathcal{O}}(d)$ times. The key idea is to prove that $\det(\sum_{\tau \in \mathcal{S}} \boldsymbol{H}_\tau)$ is at least doubled each time we add one more time step to $\mathcal{S}$. Thus, if the eigenvalues of $\boldsymbol{H}_t$ are bounded, $|\mathcal{S}|$ cannot be too large. Ensuring the last fact requires a small tweak to the OMD update (Line 5), where we constrain the optimization over a slightly shrunk version of $\boldsymbol{\Omega}$ defined as $\boldsymbol{\Omega}' = \{\boldsymbol{w} \in \boldsymbol{\Omega} : \pi_{\boldsymbol{w}_1}(\boldsymbol{w}) \leq 1 - \frac{1}{T}\}$. Here, π is the Minkowsky function and we defer its formal definition to Appendix B.2, but intuitively $\boldsymbol{\Omega}'$ is simply obtained by shrinking the lifted decision set by a small amount of $1/T$ with respect to the center $\boldsymbol{w}_1$ (which is the analogue of the truncated simplex for MAB). This makes sure that $\boldsymbol{w}_t$ is never too close to the boundary, and in turn makes sure that the eigenvalues of $\boldsymbol{H}_t$ are bounded.

This concludes the two main new ideas of our algorithm; see Algorithm 2 for the complete pseudocode. Clearly, our algorithm can be implemented as efficiently as the original SCRIBLE. The regret guarantee is summarized below.

Theorem 3.1. *Algorithm 2 with a suitable choice of η ensures that with probability at least $1 - \delta$:*

$$\mathrm{Reg} = \begin{cases} \tilde{\mathcal{O}}\left(d^2 \nu \sqrt{T \ln \frac{1}{\delta}} + d^2 \nu \ln \frac{1}{\delta}\right), & \text{against an oblivious adversary;} \\ \tilde{\mathcal{O}}\left(d^2 \nu \sqrt{dT \ln \frac{1}{\delta}} + d^3 \nu \ln \frac{1}{\delta}\right), & \text{against an adaptive adversary.} \end{cases}$$

Moreover, if $\langle w, \ell_t \rangle \geq 0$ for all $w \in \Omega$ and all t, then T in the bounds above can be replaced by $L^\star = \min_{u \in \Omega} \sum_{t=1}^T \langle u, \ell_t \rangle$, that is, the total loss of the best action.

Our results are the first general high-probability regret guarantees achieved by an efficient algorithm (for either oblivious or adaptive adversary). We not only achieve $\sqrt{T}$-type bounds, but also improve it to $\sqrt{L^\star}$-type small-loss bounds, which does not exist before. Note that the latter holds only when losses are nonnegative, which is a standard setup for small-loss bounds and is true, for instance, for all combinatorial bandit problems where $\Omega \subseteq [0,1]^d$ lives in the positive orthant. Similarly to MAB, we can also obtain other data-dependent regret bounds by only changing the estimator to $d\langle \widetilde{w}_t, \ell_t - m_t \rangle H_t^{1/2} s_t + m_t$ for some predictor m_t (see [33, 17]).[3]

Ignoring lower order terms, our bound for oblivious adversaries is $d\sqrt{\nu}$ times worse than the expected regret of SCRIBLE. For adaptive adversary, we pay extra dependence on d, which is standard since an extra union bound over u is needed and is discussed in [1] as well. The minimax regret for adaptive adversary is still unknown. Reducing the dependence on d for both cases is a key future direction.

4 Generalization to adversarial MDPs

Finally, we briefly discuss how to generalize Algorithm 1 for MAB to learning adversarial Markov Decision Processes (MDPs), leading to the first algorithm with a high-probability small-loss regret guarantee for this problem. We consider an episodic MDP setting with finite horizon, unknown transition kernel, bandit feedback, and adversarial losses, the exact same setting as the recent work [25] (which is the state-of-the-art for adversarial tabular MDPs; see [25] for related work).

Specifically, the problem is parameterized by a state space X, an action space A, and an unknown transition kernel $P : X \times A \times X \to [0,1]$ with $P(x'|x,a)$ being the probability of reaching state x'

---

[3] One caveat is that this requires measuring the learner's loss in terms of $\langle w_t, \ell_t \rangle$, as opposed to $\langle \widetilde{w}_t, \ell_t \rangle$, since the deviation between these two is not related to m_t.

after taking action a at state x. Without loss of generality (see discussions in [25]), the state space is assumed to be partitioned into $J + 1$ layers $X_0, \ldots, X_J$ where $X_0 = \{x_0\}$ and $X_J = \{x_J\}$ contain only the start and end state respectively, and transitions are only possible between consecutive layers.

The learning proceeds in T rounds/episodes. In each episode t, the learner starts from state x_0 and decides a stochastic policy $\pi_t : X \times A \to [0, 1]$, where $\pi_t(a|x)$ is the probability of selecting action a at state x. Simultaneously, the adversary decides a loss function $\ell_t : X \times A \to [0, 1]$, with $\ell_t(x, a)$ being the loss of selecting action a at state x. Once again, an adaptive adversary chooses ℓ_t based on all learner's actions in previous episodes, while an oblivious adversary chooses ℓ_t only knowing the learner's algorithm. Afterwards, the learner executes the policy in the MDP for J steps and generates/observes a state-action-loss sequence $(x_0, a_0, \ell_t(x_0, a_0)), \ldots, (x_{J-1}, a_{J-1}, \ell_t(x_{J-1}, a_{J-1}))$ before reaching the final state x_J. With a slight abuse of notation, we use $\ell_t(\pi) = \mathbb{E}\left[\sum_{k=1}^{J-1} \ell_t(x_k, a_k) \mid P, \pi\right]$ to denote the expected loss of executing policy π in episode t. The regret of the learner is then defined as $\text{Reg} = \sum_{t=1}^T \ell_t(\pi_t) - \min_\pi \sum_{t=1}^T \ell_t(\pi)$, where the min is over all possible policies.

Based on several prior works [38, 35], Jin et al. [25] showed the deep connection between this problem and adversarial MAB. In fact, with the help of the "occupancy measure" concept, this problem can be reformulated in a way that becomes very much akin to adversarial MAB and can be essentially solved using OMD with some importance-weighted estimators. We refer the reader to [25] and Appendix C.1 for details. The algorithm of [25] achieves $\text{Reg} = \tilde{\mathcal{O}}(J|X|\sqrt{|A|T})$ with high probability.

Since the problem has great similarity with MAB, the natural idea to improve the bound to a small-loss bound is to borrow techniques from MAB. Prior to our work, obtaining high-probability small-loss bounds for MAB can only be achieved by either the implicit exploration idea from [31] or the clipping idea from [7, 29]. Unfortunately, in Appendix C.4, we argue that neither of them works for MDPs, at least not in a direct way we can see, from perspectives of both the algorithm and the analysis.

On the other hand, our approach from Algorithm 1 immediately generalizes to MDPs without much effort. Compared to the algorithm of [25], the only essential differences are to replace their regularizer with log-barrier and to apply a similar increasing learning rate schedule. Due to space limit, we defer the algorithm to Appendix C.2 and show the main theorem below.

Theorem 4.1. *Algorithm 4 with a suitable choice of η ensures that with probability at least $1 - \delta$,* $\text{Reg} = \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L^\star \ln \frac{1}{\delta}} + |X|^5|A|^2 \ln^2 \frac{1}{\delta}\right)$, *where $L^\star = \min_\pi \sum_{t=1}^T \ell_t(\pi) \leq JT$ is the total loss of the best policy.*

We remark that our bound holds for both oblivious and adaptive adversaries, and is the first high-probability small-loss bounds for adversarial MDPs.[4] This matches the bound of [25] in the worst case (including the lower-order term $\tilde{\mathcal{O}}(|X|^5|A|^2)$ hidden in their proof), but could be much smaller as long as a good policy exists with $L^\star = o(T)$. It is still open whether this bound is optimal or not.

5   Conclusions

In this work, based on the idea of increasing learning rates we develop a new technique for obtaining high-probability regret bounds against an adaptive adversary under bandit feedback, showing that sophisticated biased estimators used in previous approaches are not necessary. We provide three examples (MAB, linear bandits, and MDPs) to show the versatility of our general approach, leading to several new algorithms and results. Although not included in this work, we point out that our approach can also be straightforwardly applied to other problems such as semi-bandits and convex bandits, based on the algorithms from [37] and [36] respectively, since they are also based on log-barrier OMD or SCRIBLE.

---

[4]Obtaining other data-dependent regret bounds as in MAB and linear bandits is challenging in this case, since there are several terms in the regret bound that are naturally only related to $L^\star$.

Broader Impact

This work is mostly theoretical, and we do not foresee any negative ethical or societal outcomes. Researchers working on theoretical aspects of online learning, bandit problems, and Markov Decision Processes may benefit from our results and find our techniques useful for other problems. In the long run, our results might lead to more stable and practical learning algorithms for applications with partial information feedback such as network routing or recommendation systems.

Acknowledgments and Disclosure of Funding

References

[1] Jacob Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. In *Conference on Learning Theory*, 2009.

[2] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

[3] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7): 4164–4175, 2012.

[4] Jacob D Abernethy, Young Hun Jung, Chansoo Lee, Audra McMillan, and Ambuj Tewari. Online learning via the differential privacy lens. In *Advances in Neural Information Processing Systems*, 2019.

[5] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 2017.

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

[7] Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in online learning in case of unbounded losses under partial monitoring. In *International conference on Algorithmic Learning Theory*, 2006.

[8] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 2009.

[9] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

[10] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2002.

[11] Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Symposium on Theory of Computing*, 2004.

[12] Peter L Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Conference On Learning Theory*, 2008.

[13] Gábor Braun and Sebastian Pokutta. An efficient high-probability algorithm for linear bandits. *arXiv preprint arXiv:1610.02072*, 2016.

[14] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

[15] Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Symposium on Theory of Computing*, 2017.

[16] Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, 2018.

[17] Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, 2019.

[18] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[19] Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, 2008.

[20] Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, 2016.

[21] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[22] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct):2369–2403, 2007.

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

[24] Elad Hazan and Zohar Karnin. Volumetric spanners: an efficient exploration basis for learning. *The Journal of Machine Learning Research*, 17(1):4062–4095, 2016.

[25] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

[26] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press (preprint), 2018.

[27] Chung-Wei Lee, Haipeng Luo, and Mengxiao Zhang. A closer look at small-loss bounds for bandits with graph feedback. In *Conference on Learning Theory*, 2020.

[28] Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, 2018.

[29] Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. In *Conference on Learning Theory*, 2018.

[30] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.

[31] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

[32] Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, 2015.

[33] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, 2013.

[34] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[35] Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, 2019.

[36] Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[37] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, 2018.

[38] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, 2013.

A Omitted details for Section 2

A.1 Proof of Theorem 2.2

First we generalize the proof of the standard Freedman's inequality in the following way. For any λ_t that is $\mathcal{F}_t$-measurable and such that $\lambda_t \leq 1/B_t$, we have with $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot|\mathcal{F}_t]$:

$$\mathbb{E}_t\left[e^{\lambda_t X_t}\right] \leq \mathbb{E}_t\left[1 + \lambda_t X_t + \lambda_t^2 X_t^2\right] = 1 + \lambda_t^2 \mathbb{E}_t\left[X_t^2\right] \leq \exp\left(\lambda_t^2 \mathbb{E}_t\left[X_t^2\right]\right). \tag{1}$$

Now for any t define random variable Z_t such that $Z_0 = 1$ and

$$Z_t \triangleq Z_{t-1} \cdot \exp(\lambda_t X_t - \lambda_t^2 \mathbb{E}_t\left[X_t^2\right]) = \exp\left(\sum_{s=1}^{t}\lambda_s X_s - \sum_{s=1}^{t}\lambda_s^2 \mathbb{E}_s\left[X_s^2\right]\right).$$

From Eq. (1), we have

$$\mathbb{E}_t\left[Z_t\right] = Z_{t-1} \cdot \exp(-\lambda_t^2 \mathbb{E}_t\left[X_t^2\right])\mathbb{E}_t\left[e^{\lambda_t X_t}\right] \leq Z_{t-1} \cdot \exp(-\lambda_t^2 \mathbb{E}_t\left[X_t^2\right])\exp(\lambda_t^2 \mathbb{E}_t\left[X_t^2\right]) \leq Z_{t-1}.$$

Therefore, taking the overall expectation we have

$$\mathbb{E}\left[Z_T\right] \leq \mathbb{E}\left[Z_{T-1}\right] \leq \cdots \leq \mathbb{E}\left[Z_0\right] = 1.$$

Using Markov's inequality, we have $\Pr\left[Z_T \geq \frac{1}{\delta'}\right] \leq \delta'$. In other words, we have with probability at least $1 - \delta'$,

$$\sum_{t=1}^{T}\lambda_t X_t \leq \ln(1/\delta') + \sum_{t=1}^{T}\lambda_t^2 \mathbb{E}_t\left[X_t^2\right]. \tag{2}$$

The proof of the standard Freedman's inequality takes all λ_t to be the same fixed value, while in our case it is important to apply Eq. (2) several times with different sets of values of λ_t. Specifically, for each $i \in [\lceil \log_2(b^2 T)\rceil]$ and $j \in [\lceil \log_2 b\rceil]$, set

$$\lambda_t = \lambda \triangleq \min\left\{2^{-j}, \sqrt{\ln(1/\delta')/2^i}\right\},$$

for $t \in \mathcal{T}_j$, where

$$\mathcal{T}_j \triangleq \left\{t : 2^{j-1} \leq \max_{s \leq t} B_s \leq 2^j\right\},$$

and $\lambda_t = 0$ otherwise. Clearly λ_t is $\mathcal{F}_t$-measurable (since $B_1, \ldots, B_t$ are $\mathcal{F}_t$-measurable). Applying Eq. (2) gives

$$\sum_{t \in \mathcal{T}_j} X_t \leq \frac{\ln(1/\delta')}{\lambda} + \sum_{t \in \mathcal{T}_j}\lambda \mathbb{E}_t\left[X_t^2\right]$$

$$\leq 2^j \ln(1/\delta') + \sqrt{2^i \ln(1/\delta')} + \lambda \sum_{t=1}^{T}\mathbb{E}_t\left[X_t^2\right] \qquad (\tfrac{1}{\lambda} \leq \max\{2^j, \sqrt{2^i/\ln(1/\delta')}\})$$

$$\leq 2\left(\max_{s \in \mathcal{T}_j} B_s\right)\ln(1/\delta') + \sqrt{2^i \ln(1/\delta')} + V\sqrt{\frac{\ln(1/\delta')}{2^i}} \qquad (2^{j-1} \leq \max_{s \in \mathcal{T}_j} B_s)$$

$$\leq 2B^\star \ln(1/\delta') + \sqrt{2^i \ln(1/\delta')} + V\sqrt{\frac{\ln(1/\delta')}{2^i}}.$$

By a union bound, the above holds with probability at least $1 - C\delta'$ for all $i \in [\lceil \log_2(b^2 T)\rceil]$ and $j \in [\lceil \log_2 b\rceil]$. In particular, since $1 \leq V \leq b^2 T$ (almost surely), there exists an $i^\star \in [\lceil \log_2(b^2 T)\rceil]$ such that $2^{i^\star - 1} \leq V \leq 2^{i^\star}$, and thus

$$\sum_{t=1}^{T} X_t = \sum_{j \in [\lceil \log_2 b\rceil]}\sum_{t \in \mathcal{T}_j} X_t \leq C \cdot \left(2B^\star \ln(1/\delta') + \sqrt{2^{i^\star}\ln(1/\delta')} + V\sqrt{\frac{\ln(1/\delta')}{2^{i^\star}}}\right)$$

$$\leq C \cdot \left(2B^\star \ln(1/\delta') + \sqrt{2V \ln(1/\delta')} + \sqrt{V \ln(1/\delta')}\right)$$

$$\leq C \cdot \left(2B^\star \ln(1/\delta') + \sqrt{8V \ln(1/\delta')}\right).$$

Finally replacing δ' with δ/C finishes the proof.

A.2 Proof of Lemma 2.1

First note that $\ell_{t,i_t} = \left\langle w_t, \widehat{\ell}_t \right\rangle$. Using standard OMD analysis (e.g., [5, Lemma 12]), we have

$$\ell_{t,i_t} - \left\langle u, \widehat{\ell}_t \right\rangle \leq D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1}) + \sum_{i=1}^{d} \eta_{t,i} w_{t,i}^2 \widehat{\ell}_{t,i}^2. \tag{3}$$

Summing the first two terms on the right hand side over t shows (here $h(y) = y - 1 - \ln y$):

$$\sum_{t=1}^{T} \left(D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1}) \right)$$

$$\leq D_{\psi_1}(u, w_1) + \sum_{t=1}^{T-1} \left(D_{\psi_{t+1}}(u, w_{t+1}) - D_{\psi_t}(u, w_{t+1}) \right) \qquad (D_{\psi_T}(u, w_{T+1}) \geq 0)$$

$$= \frac{1}{\eta} \sum_{j=1}^{d} h\left(\frac{u_j}{w_{1,j}} \right) + \sum_{j=1}^{d} \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1,j}} - \frac{1}{\eta_{t,j}} \right) h\left(\frac{u_j}{w_{t+1,j}} \right). \tag{4}$$

For the first term, since $u_j \geq \frac{1}{T}$ and $w_{1,j} = \frac{1}{d}$ for each j, we have

$$\frac{1}{\eta} \sum_{j=1}^{d} h\left(\frac{u_j}{w_{1,j}} \right) = \frac{1}{\eta} \sum_{j=1}^{d} -\ln(d u_j) \leq \frac{d \ln T}{\eta}.$$

Now we analyze the second term for each j. Note that $\eta_{T,j} = \kappa^{n_j} \eta_{1,j}$ where n_j is the number of times Algorithm 1 increases the learning rate for arm j. Let t_j be the time step such that $\eta_{T,j} = \eta_{t_j+1,j} = \kappa \eta_{t_j,j}$, that is, the last time step where the learning rate for arm j is increased. Then we have

$$\left(\frac{1}{\eta_{t_j+1,j}} - \frac{1}{\eta_{t_j,j}} \right) h\left(\frac{u_j}{w_{t_j+1,j}} \right) = \frac{1 - \kappa}{\kappa^{n_j} \eta} h\left(\frac{u_j}{w_{t_j+1,j}} \right) \leq \frac{-h\left(\frac{u_j}{w_{t_j+1,j}} \right)}{5\eta \ln T} = \frac{-h\left(\frac{u_j \rho_{T,j}}{2} \right)}{5\eta \ln T},$$

where we use the facts $1 - \kappa \leq -\frac{1}{\ln T}$ and $\kappa^{n_j} \leq 5$. The term $-h\left(\frac{u_j \rho_{T,j}}{2} \right)$ is bounded by

$$-h\left(\frac{u_j \rho_{T,j}}{2} \right) = \ln\left(\frac{u_j \rho_{T,j}}{2} \right) - \frac{u_j \rho_{T,j}}{2} + 1 \leq 1 + \ln T - \frac{u_j \rho_{T,j}}{2},$$

where the inequality is because $\frac{u_j \rho_{T,j}}{2} \leq \frac{1}{w_{t_j+1,j}} \leq T$. Plugging this result for every j back to Eq. (4), we get

$$\sum_{t=1}^{T} D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1}) \leq \frac{d \ln T}{\eta} + \sum_{j=1}^{d} \frac{2 + 2\ln T - u_j \rho_{T,j}}{10\eta \ln T} = \mathcal{O}\left(\frac{d \ln T}{\eta} \right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln T}.$$

Finally, since $\eta_{t,i} w_{t,i}^2 \widehat{\ell}_{t,i}^2 \leq \eta_{t,i_t} \ell_{t,i_t} \leq \eta_{T,i_t} \ell_{t,i_t} \leq 5\eta \ell_{t,i_t}$, summing Eq. (3) over t gives:

$$\sum_{t=1}^{T} \left(\ell_{t,i_t} - \left\langle u, \widehat{\ell}_t \right\rangle \right) \leq \sum_{t=1}^{T} \left(D_{\psi_t}(u, w_t) - D_{\psi_t}(u, w_{t+1}) \right) + \sum_{t=1}^{T} \sum_{i=1}^{d} \eta_{t,i} w_{t,i}^2 \widehat{\ell}_{t,i}^2$$

$$\leq \mathcal{O}\left(\frac{d \ln T}{\eta} \right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln T} + 5\eta \sum_{t=1}^{T} \ell_{t,i_t}$$

$$= \mathcal{O}\left(\frac{d \ln T}{\eta} + \eta \sum_{t=1}^{T} \ell_{t,i_t} \right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln T}.$$

14

A.3 Proof of Theorem 2.3

Fix any $i^\star \in [d]$ and let $u = (1 - \frac{d}{T})e^\star + \frac{1}{T}\mathbf{1}$, where $e^\star$ is the one-hot vector for $i^\star$. First note that

$$\sum_{t=1}^{T} (\ell_{t,i_t} - \ell_{t,i^\star}) = \sum_{t=1}^{T} \left(\ell_{t,i_t} - \left\langle u, \hat{\ell}_t \right\rangle\right) + \sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle + \sum_{t=1}^{T} \left\langle u - e^\star, \ell_t \right\rangle$$

$$\leq \sum_{t=1}^{T} \left(\ell_{t,i_t} - \left\langle u, \hat{\ell}_t \right\rangle\right) + \sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle + d.$$

For the first term, using Lemma 2.1, we have

$$\sum_{t=1}^{T} \left(\ell_{t,i_t} - \left\langle u, \hat{\ell}_t \right\rangle\right) \leq \mathcal{O}\left(\frac{d \ln T}{\eta} + \eta L_T\right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln T}, \tag{5}$$

where $L_T = \sum_{t=1}^{T} \ell_{t,i_t}$.

For the second term above, we use Theorem 2.2 with $X_t = \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle$, $B_t = \langle \rho_t, u \rangle \in [1, T]$, $b = T$, and the fact

$$\mathbb{E}_t[X_t^2] \leq \mathbb{E}_t\left[\left\langle u, \widehat{\ell}_t \right\rangle^2\right] = \mathbb{E}_t\left[\frac{u_{i_t}^2 \ell_{t,i_t}^2}{p_{t,i_t}^2}\right] \leq \sum_{i=1}^{d} u_i^2 \ell_{t,i}\rho_{T,i} \leq \langle \rho_T, u \rangle \langle u, \ell_t \rangle,$$

showing that with probability at least $1 - \delta'$,

$$\sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle \leq C\left(\sqrt{8L_u \langle \rho_T, u \rangle \ln(C/\delta')} + 2\langle \rho_T, u \rangle \ln(C/\delta')\right), \tag{6}$$

where $L_u = \left\langle u, \sum_{t=1}^{T} \ell_t \right\rangle$ and $C = \lceil \log(b) \rceil \lceil \log(b^2 T) \rceil = \lceil \log(T) \rceil \lceil 3\log(T) \rceil$. With $\eta \leq \frac{1}{40C \ln T \ln(C/\delta')}$, we then have with probability at least $1 - \delta'$,

$$\sum_{t=1}^{T} \ell_{t,i_t} - \ell_{t,i^\star}$$

$$\leq \tilde{\mathcal{O}}\left(\frac{d}{\eta} + \eta L_T\right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln T} + C\left(\sqrt{8L_u \langle \rho_T, u \rangle \ln(C/\delta')} + 2\langle \rho_T, u \rangle \ln(C/\delta')\right)$$

<div align="right">(Eq. (5) and Eq. (6))</div>

$$\leq \tilde{\mathcal{O}}\left(\frac{d}{\eta} + \eta L_T\right) + \eta 40C^2 L_u \ln(C/\delta') \ln T - \frac{\langle \rho_T, u \rangle}{20\eta \ln T} + 2C\langle \rho_T, u \rangle \ln(C/\delta')$$

<div align="right">(AM-GM inequality)</div>

$$\leq \tilde{\mathcal{O}}\left(\frac{d}{\eta} + \eta L_T \ln(1/\delta') + \eta L_u \ln(1/\delta')\right). \qquad (\eta < \tfrac{1}{40C \ln T \ln(C/\delta)})$$

Therefore, rearranging the terms, using the fact $L_u \leq L^\star + d$, and choosing $\eta = \min\left\{\sqrt{\frac{d}{L^\star} \ln(1/\delta')}, \frac{1}{40C \ln T \ln(C/\delta')}, \frac{1}{2}\right\}$, we have with probability $1 - \delta'$,

$$\sum_{t=1}^{T} \ell_{t,i_t} - \ell_{t,i^\star} = \tilde{\mathcal{O}}\left(\sqrt{dL^\star \ln(1/\delta')} + d\ln(1/\delta')\right),$$

where $L^\star = \sum_{t=1}^{T} \ell_{t,i^\star}$. This finishes the proof when the adversary is oblivious. For adaptive adversaries, taking a union bound over all possible best arms $i^\star \in [d]$ and setting $\delta' = \delta/d$, we have with probability $1 - \delta$, $\mathrm{Reg} = \widetilde{\mathcal{O}}\left(\sqrt{dL^\star \ln(d/\delta)} + d\ln(d/\delta)\right)$, finishing the proof.

Remark 1. Although the proof above requires tuning the initial learning rate η in terms of the unknown quantity $L^\star$, standard doubling trick can remove this restriction (even in the bandit setting). We refer the reader to a recent work by Lee et al. [27] for detailed exposition on how to achieve so.

Algorithm 3 d-dimensional version of Algorithm 2

Input: decision set $\Omega \subseteq \mathbb{R}^d$, a ν-self-concordant barrier $\psi(w)$ for Ω, initial learning rate η.

Define: increase factor $\kappa = e^{\frac{1}{100d \ln(\nu T)}}$, $\Psi(w, b) = 400\left(\psi\left(\frac{w}{b}\right) - 2\nu \ln b\right)$.

Initialize: $w_1 = \operatorname{argmin}_{w \in \Omega} \psi(w)$, $\boldsymbol{H}_1 = \nabla^2 \Psi(w_1, 1)$, $\eta_1 = \eta$, $\mathcal{S} = \{1\}$.

Define: shrunk decision set $\Omega' = \{w \in \Omega : \pi_{w_1}(w) \le 1 - \frac{1}{T}\}$, $J = [I_d, \mathbf{0}_d] \in \mathbb{R}^{d \times (d+1)}$.

for $t = 1, 2, \ldots, T$ **do**

1   Uniformly at random sample $\boldsymbol{s}_t$ from $\left(\boldsymbol{H}_t^{-\frac{1}{2}} e_{d+1}\right)^{\perp} \cap \mathbb{S}^{d+1}$.

2   Compute $\widetilde{w}_t = w_t + J\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{s}_t$.

3   Play $\widetilde{w}_t$, observe loss $\langle \widetilde{w}_t, \ell_t \rangle$, and construct loss estimator $\widehat{\ell}_t = d\langle \widetilde{w}_t, \ell_t \rangle J\boldsymbol{H}_t^{1/2}\boldsymbol{s}_t$.

4   Compute $w_{t+1} = \operatorname{argmin}_{w \in \Omega'}\left\{\left\langle w, \widehat{\ell}_t \right\rangle + D_{\psi_t}(w, w_t)\right\}$, where $\psi_t = \frac{1}{\eta_t}\psi$.

5   Compute $\boldsymbol{H}_{t+1} = \nabla^2 \Psi(w_{t+1}, 1)$.

6   **if** $\lambda_{\max}(\boldsymbol{H}_{t+1} - \sum_{\tau \in \mathcal{S}} \boldsymbol{H}_\tau) > 0$ **then** $\mathcal{S} \leftarrow \mathcal{S} \cup \{t+1\}$ and set $\eta_{t+1} = \eta_t \kappa$;

7   **else** set $\eta_{t+1} = \eta_t$.

B   Omitted details for Section 3

B.1   More explanation on Algorithm 2

Here, we provide a d-dimensional version of Algorithm 2 by removing the explicit lifting and performing OMD in $\mathbb{R}^d$; see Algorithm 3. It is clear that this version is exactly the same as Algorithm 2. Compared to the original SCRIBLE, one can see that besides the increasing learning rate schedule, the only difference is how the point $\widetilde{w}_t$ is computed. In particular, one can verify that $\widetilde{w}_t$ does not necessarily satisfy $\|\widetilde{w}_t - w_t\|_{\nabla^2 \psi(w_t)} = 1$, meaning that $\widetilde{w}_t$ is not necessarily on the boundary of the Dikin ellipsoid centered at w_t with respect to ψ. In other words, our algorithm provides a new sampling scheme for SCRIBLE.

B.2   Preliminary for analysis

In this section, we introduce the preliminary of self-concordant barriers and normal-barriers, including the definitions and some useful properties that will be used frequently in later analysis.

Self-concordant barriers.   Let $\psi : \operatorname{int}(\Omega) \to \mathbb{R}$ be a C^3 smooth convex function. ψ is called a self-concordant barrier on Ω if it satisfies:

- $\psi(x_i) \to \infty$ as $i \to \infty$ for any sequence $x_1, x_2, \cdots \in \operatorname{int}(\Omega) \subset \mathbb{R}^d$ converging to the boundary of Ω;

- for all $w \in \operatorname{int}(\Omega)$ and $h \in \mathbb{R}^d$, the following inequality always holds:

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 \psi(w)}{\partial w_i \partial w_j \partial w_k} h_i h_j h_k \le 2\|h\|_{\nabla^2 \psi(w)}^3.$$

We further call ψ is a ν-self-concordant barrier if it satisfies the conditions above and also

$$\langle \nabla \psi(w), h \rangle \le \sqrt{\nu}\|h\|_{\nabla^2 \psi(w)}$$

for all $w \in \operatorname{int}(\Omega)$ and $h \in \mathbb{R}^d$.

Lemma B.1 (Theorem 2.1.1 in [30]). *If ψ is a self-concordant barrier on Ω, then the Dikin ellipsoid centered at $w \in \operatorname{int}(\Omega)$, defined as $\{v : \|v - w\|_{\nabla^2 \psi(w)} \le 1\}$, is always within Ω. Moreover,*

$$\|h\|_{\nabla^2 \psi(v)} \ge \|h\|_{\nabla^2 \psi(w)}\left(1 - \|v - w\|_{\nabla^2 \psi(w)}\right)$$

holds for any $h \in \mathbb{R}^d$ and any v with $\|v - w\|_{\nabla^2 \psi(w)} \le 1$.

Lemma B.2 (Theorem 2.5.1 in [30]). *For any closed convex body $\Omega \subset \mathbb{R}^d$, there exists an $\mathcal{O}(d)$-self-concordant barrier on Ω.*

Lemma B.3 (Corollary 2.3.1 in [30]). *Let ψ be a self-concordant barrier for $\Omega \subset \mathbb{R}^d$. Then for any $w \in \text{int}(\Omega)$ and any $h \in \Omega$ such that $w + bh \in \Omega$ for all $b \geq 0$, we have*

$$\|h\|_{\nabla^2 \psi(w)} \leq -\langle \nabla \psi(w), h \rangle.$$

Normal Barriers. Let $K \subseteq \mathbb{R}^d$ be a closed and proper convex cone and let $\theta \geq 1$. A function $\psi : \text{int}(K) \to \mathbb{R}$ is called a θ-logarithmically homogeneous self-concordant barrier (or simply a θ-normal barrier) on K if it is self-concordant on $\text{int}(K)$ and is logarithmically homogeneous with parameter θ, which means

$$\psi(tw) = \psi(w) - \theta \ln t, \ \forall w \in \text{int}(K), \ t > 0.$$

The following two lemmas show the relationship between θ-normal barriers and θ-self-concordant barriers.

Lemma B.4 (Corollary 2.3.2 in [30]). *A θ-normal barrier on K is a θ-self-concordant barrier on K.*

Lemma B.5 (Proposition 5.1.4 in [30]). *Suppose f is a θ-self-concordant barrier on $K \subseteq \mathbb{R}^d$. Then the function*

$$F(w, b) = 400 \left(f\left(\frac{w}{b}\right) - 2\theta \ln b \right),$$

is a 800θ-normal barrier for $con(K) \subseteq \mathbb{R}^{d+1}$, where $con(K) = \{\mathbf{0}\} \cup \{(w, b) : \frac{w}{b} \in K, w \in \mathbb{R}^d, b > 0\}$ is the conic hull of K lifted to $\mathbb{R}^{d+1}$ (by appending 1 to the last coordinate).

Note that our regularizer Ψ defined in [Algorithm 2](#) is exactly based on this formula. We point out that, however, our entire analysis works for any $\mathcal{O}(\nu)$-normal barrier Ψ, as we will only use the following general properties of normal barriers, instead of the concrete form of Ψ. As mentioned in [Footnote 2](#), we use this concrete formula only to emphasize that, just as SCRIBLE, our algorithm requires only a self-concordant barrier of the original set Ω.

Lemma B.6 (Proposition 2.3.4 in [30]). *If ψ is a θ-normal barrier on K, then we have for all $w, u \in \text{int}(K)$,*

1. *$\|w\|^2_{\nabla^2 \psi(w)} = w^\top \nabla^2 \psi(w) w = \theta,$*

2. *$\nabla^2 \psi(w) w = -\nabla \psi(w),$*

3. *$\psi(u) \geq \psi(w) - \theta \ln \frac{-\langle \nabla \psi(w), u \rangle}{\theta}.$*

Next, we show the definition of Minkowsky functions, which is used to define the shrunk decision domain similar to the clipped simplex in multi-armed bandit setting.

Minkowsky functions. The Minkowsky function of a convex body Ω with the pole at $w \in \text{int}(\Omega)$ is a function $\pi_w : \Omega \to \mathbb{R}$ defined as

$$\pi_w(u) = \inf \left\{ t > 0 \,\middle|\, w + \frac{u - w}{t} \in \Omega \right\}.$$

The last lemma shows several useful properties using the Minkowsky function.

Lemma B.7 (Proposition 2.3.2 in [30]). *Let ψ be a ν-self-concordant barrier on $\Omega \subseteq \mathbb{R}^d$ and $u, w \in \text{int}(\Omega)$. Then for any $h \in \mathbb{R}^d$, we have*

$$\|h\|_{\nabla^2 \psi(u)} \leq \left(\frac{1 + 3\nu}{1 - \pi_w(u)} \right) \|h\|_{\nabla^2 \psi(w)},$$

$$|\langle \nabla \psi(u), h \rangle| \leq \left(\frac{\nu}{1 - \pi_w(u)} \right) \|h\|_{\nabla^2 \psi(w)},$$

$$\psi(u) - \psi(w) \leq \nu \ln \left(\frac{1}{1 - \pi_w(u)} \right).$$

B.3 Proof of Theorem 3.1

To prove the theorem, we decompose the regret against any fixed $u^\star \in \Omega$ (with $\boldsymbol{u}^\star = (u^\star, 1) \in \boldsymbol{\Omega}$) into the following three terms:

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle$$

$$= \sum_{t=1}^{T} \langle \widetilde{\boldsymbol{w}}_t - \boldsymbol{u}^\star, \boldsymbol{\ell}_t \rangle \qquad\qquad\qquad (\text{define } \boldsymbol{\ell}_t = (\ell_t, 0))$$

$$= \underbrace{\sum_{t=1}^{T} \left(\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t \rangle - \left\langle \boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t \right\rangle + \left\langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t \right\rangle \right)}_{\textsc{Deviation}} + \underbrace{\sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t \right\rangle}_{\textsc{Reg-Term}} + \sum_{t=1}^{T} \langle \boldsymbol{u} - \boldsymbol{u}^\star, \boldsymbol{\ell}_t \rangle, \quad (7)$$

where $\boldsymbol{u} = \left(1 - \frac{1}{T}\right) \cdot \boldsymbol{u}^\star + \frac{1}{T} \cdot \boldsymbol{w}_1 \in \boldsymbol{\Omega}'$. Note that the last term is trivially bounded by 2 as $\sum_{t=1}^{T} \langle \boldsymbol{u} - \boldsymbol{u}^\star, \boldsymbol{\ell}_t \rangle = \sum_{t=1}^{T} \langle u - u^\star, \ell_t \rangle = \frac{1}{T} \sum_{t=1}^{T} \langle u^\star - w_1, \ell_t \rangle \le 2$, where the last inequality is because $|\langle w, \ell_t \rangle| \le 1$ for all $w \in \Omega$. In the following sections, we show how to bound other terms. Specifically, we bound Deviation in Section B.3.1 and Reg-Term in Section B.3.2. Finally we prove Theorem 3.1 in Section B.3.3.

We will use the following notations in the remaining of this section (the first two are mentioned above already):

$$\boldsymbol{\ell}_t \triangleq (\ell_t, 0), \quad \boldsymbol{u} \triangleq (u, 1) \triangleq \left(1 - \frac{1}{T}\right) \cdot \boldsymbol{u}^\star + \frac{1}{T} \cdot \boldsymbol{w}_1 \in \boldsymbol{\Omega}', \quad \rho \triangleq \max_{t \in [T]} \|\boldsymbol{u}\|_{\boldsymbol{H}_t}, \qquad (8)$$

$$L_T \triangleq \sum_{t=1}^{T} \langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t \rangle, \quad \overline{L}_T \triangleq \sum_{t=1}^{T} |\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t \rangle|, \qquad (9)$$

$$\mathring{L}_T \triangleq \sum_{t=1}^{T} \mathbb{E}_t \left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t \rangle| \right], \quad \overline{L}_u \triangleq \sum_{t=1}^{T} |\langle u, \ell_t \rangle|. \qquad (10)$$

Before proceeding, we provide one useful lemma.

Lemma B.8. *We have* $\|\boldsymbol{u}\|_{\boldsymbol{H}_1} \le 800\nu$.

Proof. Clearly, for any $b > 0$, we have $\boldsymbol{w}_1 + b\boldsymbol{u}$ still in the conic hull of $\boldsymbol{\Omega}$. According to Lemma B.3, we thus have $\|\boldsymbol{u}\|_{\boldsymbol{H}_1} \le \langle -\nabla \Psi(\boldsymbol{w}_1), \boldsymbol{u} \rangle$. Note that Ψ is a 800ν-normal barrier by Lemma B.5. By the first order optimality condition of $\boldsymbol{w}_1$ and Lemma B.6, we then have

$$0 \le \langle \nabla \Psi(\boldsymbol{w}_1), \boldsymbol{u} - \boldsymbol{w}_1 \rangle = \langle \nabla \Psi(\boldsymbol{w}_1), \boldsymbol{u} \rangle + 800\nu.$$

Combining the above gives $\|\boldsymbol{u}\|_{\boldsymbol{H}_1} \le \langle -\nabla \Psi(\boldsymbol{w}_1), \boldsymbol{u} \rangle \le 800\nu$. $\qquad\square$

B.3.1 Bounding Deviation

We first show that $\widehat{\boldsymbol{\ell}}_t$ is an unbiased estimator of $\boldsymbol{\ell}_t$ for the first d coordinates.

Lemma B.9. *We have* $\mathbb{E}_t \left[\widehat{\ell}_{t,i} \right] = \ell_{t,i}$ *for* $i \in [d]$.

Proof. Let $\boldsymbol{v} = \frac{\boldsymbol{H}_t^{-1/2} e_{d+1}}{\left\| \boldsymbol{H}_t^{-1/2} e_{d+1} \right\|_2}$. First note that

$$\mathbb{E}_t[\boldsymbol{s}_t \boldsymbol{s}_t^\top] = \frac{1}{d} \left(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^\top \right) \qquad (11)$$

18

by the definition of $\boldsymbol{s}_t$. Then by the definition of $\widehat{\boldsymbol{\ell}}_t$, we have

$$
\begin{aligned}
\mathbb{E}_t\left[\widehat{\boldsymbol{\ell}}_t\right] &= \mathbb{E}_t\left[d\left\langle \boldsymbol{w}_t + \boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{s}_t, \boldsymbol{\ell}_t\right\rangle \cdot \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right] \\
&= \mathbb{E}_t\left[d\left\langle \boldsymbol{w}_t, \boldsymbol{\ell}_t\right\rangle \cdot \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t + d\cdot \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\left\langle \boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{s}_t, \boldsymbol{\ell}_t\right\rangle\right] \\
&= d\left\langle \boldsymbol{w}_t, \boldsymbol{\ell}_t\right\rangle \cdot \boldsymbol{H}_t^{\frac{1}{2}}\mathbb{E}_t\left[\boldsymbol{s}_t\right] + \mathbb{E}_t\left[d\cdot \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\boldsymbol{s}_t^\top \boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\ell}_t\right] \\
&= d\cdot \boldsymbol{H}_t^{\frac{1}{2}}\mathbb{E}_t\left[\boldsymbol{s}_t\boldsymbol{s}_t^\top\right]\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\ell}_t && (\mathbb{E}_t\left[\boldsymbol{s}_t\right] = \boldsymbol{0}\text{ by symmetry}) \\
&= \boldsymbol{H}_t^{\frac{1}{2}}\left(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^\top\right)\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\ell}_t && (\text{Eq. (11)}) \\
&= \boldsymbol{\ell}_t - \frac{e_{d+1}e_{d+1}^\top \boldsymbol{H}_t^{-1}\boldsymbol{\ell}_t}{\left\|\boldsymbol{H}_t^{-1/2}e_{d+1}\right\|_2^2}.
\end{aligned}
$$

Noticing that the first d coordinates of $e_{d+1}e_{d+1}^\top \boldsymbol{H}_t^{-1}\boldsymbol{\ell}_t$ are all zeros concludes the proof. $\qquad\square$

Now we are ready to bound DEVIATION.

Lemma B.10. *With probability at least $1-\delta$, we have*

$$
\text{DEVIATION} \le 161Cd\sqrt{(\nu + \rho^2)\mathring{L}_T\ln(C/\delta)} + C\sqrt{32\overline{L}_u\ln(C/\delta)} + 64Cd\left(\sqrt{\nu} + \rho\right)\ln(C/\delta),
$$

where $C = \Theta(\ln^2(d\nu T))$.

Proof. Define $X_t \triangleq \langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle\boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle + \left\langle\boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle$ and we have DEVIATION $= \sum_{t=1}^T X_t$. The goal is to apply our strengthened Freedman's inequality Theorem 2.2. To this end, first we show $\mathbb{E}_t[X_t] = 0$. Indeed, we have $\mathbb{E}_t[\widetilde{\boldsymbol{w}}_t] = \boldsymbol{w}_t$ and

$$
\begin{aligned}
\mathbb{E}_t\left[X_t\right] &= \langle w_t, \ell_t\rangle - \langle w_t, \ell_t\rangle + \langle u, \ell_t - \ell_t\rangle - \mathbb{E}_t\left[(\boldsymbol{w}_{t,d+1} - \boldsymbol{u}_{t,d+1})\widehat{\boldsymbol{\ell}}_{t,d+1}\right] && (\text{Lemma B.9}) \\
&= 0. && (\boldsymbol{w}_{t,d+1} = \boldsymbol{u}_{t,d+1} = 1)
\end{aligned}
$$

Next, we bound X_t by a $\mathcal{F}_t$-measurable random variable $B_t \triangleq 32d\sqrt{\nu} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_t}$. This can be shown using the properties of a normal barrier:

$$
\begin{aligned}
X_t &= \langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle\boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle + \left\langle\boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle \\
&= \langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle\boldsymbol{w}_t, d\cdot\langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle\cdot\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right\rangle + \left\langle\boldsymbol{u}, d\cdot\langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t - \boldsymbol{\ell}_t\right\rangle \\
&= \langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle\left(1 - d\boldsymbol{w}_t^\top\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right) + d\langle\widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle\boldsymbol{u}^\top\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t - \langle\boldsymbol{u}, \boldsymbol{\ell}_t\rangle \\
&\le 2 + d\left|\boldsymbol{w}_t^\top\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right| + d\left|\boldsymbol{u}^\top\boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right| && (|\langle w, \ell_t\rangle| \le 1\text{ for any } w\in\Omega) \\
&\le 2 + d\|\boldsymbol{w}_t\|_{\boldsymbol{H}_t} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_t} && (\text{by Cauchy-Schwarz inequality and }\boldsymbol{s}_t^\top\boldsymbol{s}_t = 1) \\
&\le 2 + 20d\sqrt{2\nu} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_t} && (\text{Lemma B.5 and Lemma B.6}) \\
&\le 32d\sqrt{\nu} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_t}. && (\nu \ge 1)
\end{aligned}
$$

Then, we show that B_t is bounded by a constant $b \triangleq 2\times 10^6 d\nu^2 T$ for all t:

$$
\begin{aligned}
B_t &\le 32d\sqrt{\nu} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_1}\cdot\left(\frac{1 + 2400\nu}{1 - \pi_{\boldsymbol{w}_1}(\boldsymbol{w}_t)}\right) && (\text{Lemma B.7}) \\
&\le 32d\sqrt{\nu} + d\|\boldsymbol{u}\|_{\boldsymbol{H}_1}(1 + 2400\nu)T && (\boldsymbol{w}_t\in\boldsymbol{\Omega}') \\
&\le 32d\sqrt{\nu} + 800d\nu(1 + 2400\nu)T && (\text{Lemma B.8}) \\
&\le 2\times 10^6 d\nu^2 T. && (\nu\ge 1)
\end{aligned}
$$

The last step before applying Theorem 2.2 is to calculate $\mathbb{E}_t[X_t^2]$. We first write

$$\mathbb{E}_t[X_t^2] = \mathbb{E}_t\left[\left(\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle \boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle + \left\langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle\right)^2\right]$$

$$\leq 2\mathbb{E}_t\left[\left(\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle \boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle\right)^2\right] + 2\mathbb{E}_t\left[\left\langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle^2\right]. \tag{12}$$

The first term is bounded by:

$$\mathbb{E}_t\left[\left(\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle \boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle\right)^2\right]$$

$$= \mathbb{E}_t\left[\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle^2 \left(1 - \left\langle \boldsymbol{w}_t, d\cdot \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right\rangle\right)^2\right]$$

$$\leq \mathbb{E}_t\left[\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle^2 \left(2d^2\left(\boldsymbol{w}_t^\top \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right)^2 + 2\right)\right]$$

$$\leq \mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle| \left(2d^2\left(\boldsymbol{w}_t^\top \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right)^2 + 2\right)\right] \qquad (\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle \leq 1)$$

$$\leq \mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle| \left(2d^2\|\boldsymbol{w}_t\|_{\boldsymbol{H}_t}^2\|\boldsymbol{s}_t\|_2^2 + 2\right)\right] \qquad \text{(Cauchy-Schwarz inequality)}$$

$$\leq \mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle| \left(1600d^2\nu + 2\right)\right] \qquad (\|\boldsymbol{s}_t\|_2^2 = 1 \text{ and Lemma B.6})$$

$$\leq 1602d^2\nu\mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\right].$$

Similarly, the second term is bounded by:

$$\mathbb{E}_t\left[\left\langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle^2\right] \leq \mathbb{E}_t\left[\left(-\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle + d\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle\boldsymbol{u}^\top \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right)^2\right]$$

$$\leq \mathbb{E}_t\left[2|\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle| + 2d^2|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\cdot\left(\boldsymbol{u}^\top \boldsymbol{H}_t^{\frac{1}{2}}\boldsymbol{s}_t\right)^2\right] \qquad (\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle \leq 1)$$

$$\leq \mathbb{E}_t\left[2|\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle| + 2d^2|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\cdot\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\right].$$

Plugging these bounds to Eq. (12), we have

$$\mathbb{E}_t[X_t^2] \leq 3204d^2\nu\mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\right] + 4|\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle| + 4d^2\mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\right]\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2.$$

Summing over t gives

$$\sum_{t=1}^T \mathbb{E}_t[X_t^2] \leq 3204d^2\sum_{t=1}^T\left(\nu + \|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\right)\mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\right] + 4\sum_{t=1}^T|\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle|$$

$$\leq 3204d^2\left(\nu + \max_{t\in[T]}\|\boldsymbol{u}\|_{\boldsymbol{H}_t}^2\right)\sum_{t=1}^T\mathbb{E}_t\left[|\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle|\right] + 4\sum_{t=1}^T|\langle \boldsymbol{u}, \boldsymbol{\ell}_t\rangle|$$

$$= 3204d^2\left(\nu + \rho^2\right)\mathring{L}_T + 4\overline{L}_u.$$

Therefore, choosing $B^\star = 32d(\sqrt{\nu}+\rho)$, $b = 2\times10^6 d\nu^2 T$, $C = \lceil\log_2 b\rceil\lceil\log_2 b^2 T\rceil = \Theta(\ln^2(d\nu T))$ and using Theorem 2.2, we obtain with probability $1-\delta$,

$$\sum_{t=1}^T X_t = \sum_{t=1}^T\left(\langle \widetilde{\boldsymbol{w}}_t, \boldsymbol{\ell}_t\rangle - \left\langle \boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle + \left\langle \boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right\rangle\right)$$

$$\leq C\sqrt{25632d^2\left(\nu + \rho^2\right)\mathring{L}_T\ln(C/\delta) + 32\overline{L}_u\ln(C/\delta)} + 64Cd\left(\sqrt{\nu} + \rho\right)\ln(C/\delta).$$

Finally, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, the first term above is bounded by

$$161C\sqrt{d^2\left(\nu + \rho^2\right)\mathring{L}_T\ln(C/\delta)} + C\sqrt{32\overline{L}_u\ln(C/\delta)},$$

which finishes the proof. $\qquad\qquad\square$

B.3.2 Bounding REG-TERM

The goal of this section is to prove the following bound on REG-TERM.

Lemma B.11. *Let $\mathcal{S}$ be its final value after running Algorithm 2 for T rounds and $\mathcal{S}' = \mathcal{S} \setminus \{1, T+1\}$. Then as long as $\eta \leq \frac{1}{80d}$, we have*

$$\text{REG-TERM} \leq \tilde{\mathcal{O}}\left(\frac{\nu}{\eta}\right) - \frac{\sum_{s \in \mathcal{S}'} \|\boldsymbol{u}\|_{\boldsymbol{H}_s}}{5\eta a d \ln(\nu T)} + 40\eta d^2 \overline{L}_T.$$

for $a = 100$.

To prove this lemma, we first prove three useful lemmas. The first one shows that the number of times Algorithm 2 increases the learning rate is upper bounded by $\mathcal{O}(d \log_2(d\nu T))$.

Lemma B.12. *Assume that $T \geq 8$. Let n be the number of times Algorithm 2 increases the learning rate. Then $n \leq a d \log_2(\nu T)$ for $a = 100$. Consequently, we have $\eta_t \leq 5\eta$ for all $t \in [T]$.*

Proof. Let $\mathcal{S} = \{t_1, \ldots, t_{n+1}\}$ be its final value after running Algorithm 2 for T rounds, which means n is the number of times the algorithm has increased the learning rate, $t_1 = 1$, and for $i = 2, \ldots, n+1$, $\eta_{t_i} = \eta_{t_{i-1}}\kappa$ holds. Let $\boldsymbol{A}_i = \sum_{j=1}^{i} \boldsymbol{H}_{t_j}$. Then for any $i > 1$, according to the update rule, there exists a vector $p \in \mathbb{R}^{d+1}$ such that $p^\top \boldsymbol{H}_{t_i} p \geq p^\top \boldsymbol{A}_{i-1} p$ and thus $p^\top \boldsymbol{A}_i p \geq 2p^\top \boldsymbol{A}_{i-1} p$. Since a self-concordant function is strictly convex, $\boldsymbol{A}_i$ is positive definite for all $i \in [n]$. Therefore, let $q = \boldsymbol{A}_{i-1}^{\frac{1}{2}} p$ and we have $q^\top \boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{A}_i \boldsymbol{A}_{i-1}^{-\frac{1}{2}} q \geq 2\|q\|_2^2$. This implies that the largest eigenvalue of $\boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{A}_i \boldsymbol{A}_{i-1}^{-\frac{1}{2}}$ is at least 2. Furthermore, the smallest eigenvalue of $\boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{A}_i \boldsymbol{A}_{i-1}^{-\frac{1}{2}}$ is at least 1 since

$$\boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{A}_i \boldsymbol{A}_{i-1}^{-\frac{1}{2}} = \boldsymbol{A}_{i-1}^{-\frac{1}{2}} \left(\boldsymbol{A}_{i-1} + \boldsymbol{H}_{t_i}\right) \boldsymbol{A}_{i-1}^{-\frac{1}{2}} = I + \boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{H}_{t_i} \boldsymbol{A}_{i-1}^{-\frac{1}{2}} \succeq I.$$

Therefore, we have

$$2 \leq \det(\boldsymbol{A}_{i-1}^{-\frac{1}{2}} \boldsymbol{A}_i \boldsymbol{A}_{i-1}^{-\frac{1}{2}}) = \frac{\det(\boldsymbol{A}_i)}{\det(\boldsymbol{A}_{i-1})},$$

which implies that $\det(\boldsymbol{A}_{n+1}) \geq 2^n \det(\boldsymbol{A}_1)$.

Next we show an upper bound for $\frac{\det(\boldsymbol{A}_{n+1})}{\det(\boldsymbol{A}_1)}$. Consider any $(d+1)$-dimensional unit vector $\boldsymbol{r}$. For each $i \in [n+1]$, applying Lemma B.7 with $h = \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{r}$, $u = \boldsymbol{w}_{t_i}$ and $w = \boldsymbol{w}_1$, we have,

$$\|h\|_{\boldsymbol{H}_{t_i}}^2 = \boldsymbol{r}^\top \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{H}_{t_i} \boldsymbol{H}_1^{-\frac{1}{2}} \boldsymbol{r} \leq \left(\frac{1 + 2400\nu}{1 - \pi_{\boldsymbol{w}_1}(\boldsymbol{w}_{t_i})}\right)^2 \|h\|_{\boldsymbol{H}_1}^2 \leq (1 + 2400\nu)^2 T^2.$$

Taking a summation over all $i \in [n+1]$, we obtain

$$\boldsymbol{r}^\top \boldsymbol{A}_1^{-\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{A}_1^{-\frac{1}{2}} \boldsymbol{r} \leq (n+1)(1 + 2400\nu)^2 T^2,$$

which means that

$$\lambda_{\max}\left(\boldsymbol{A}_1^{-\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{A}_1^{-\frac{1}{2}}\right) \leq (n+1)(1 + 2400\nu)^2 T^2,$$

and thus

$$\frac{\det(\boldsymbol{A}_{n+1})}{\det(\boldsymbol{A}_1)} = \det\left(\boldsymbol{A}_1^{-\frac{1}{2}} \boldsymbol{A}_{n+1} \boldsymbol{A}_1^{-\frac{1}{2}}\right) \leq \left((n+1)(1 + 2400\nu)^2 T^2\right)^{d+1}.$$

Combining with $\frac{\det(\boldsymbol{A}_{n+1})}{\det(\boldsymbol{A}_1)} \geq 2^n$, we have

$$n \leq (d+1)\log_2(n+1) + 2(d+1)\log_2\left((1 + 2400\nu)T\right) \leq a d \log_2(\nu T),$$

for $a = 100$. To show that $\eta_t \leq 5\eta$ for t, notice that $\exp(\log_2(\nu T)/\ln(\nu T)) \leq 5$. Therefore,

$$\eta_t \leq \kappa^n \eta = \exp\left(\frac{n}{a d \ln(\nu T)}\right) \eta \leq 5\eta,$$

finishing the proof. $\qquad\square$

The second lemma gives a lower bound of the Bregman divergence between $\boldsymbol{u}$ and $\boldsymbol{w}_t$, which contains an important term to cancel DEVIATION in later analysis.

Lemma B.13. *For all $t \in [T]$, $D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \geq -800\nu \ln(800\nu T) - 800\nu + \|\boldsymbol{u}\|_{\boldsymbol{H}_t}$.*

Proof. Note again that Ψ is a 800ν-normal barrier of Ω by Lemma B.5. By the definition of Bregman divergence, we have

$$
\begin{aligned}
D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) &= \Psi(\boldsymbol{u}) - \Psi(\boldsymbol{w}_t) - \langle \nabla\Psi(\boldsymbol{w}_t), \boldsymbol{u} - \boldsymbol{w}_t \rangle \\
&\geq -800\nu \ln \frac{-\boldsymbol{u}^\top \nabla\Psi(\boldsymbol{w}_t)}{800\nu} - \langle \nabla\Psi(\boldsymbol{w}_t), \boldsymbol{u} \rangle - 800\nu. \quad \text{(Lemma B.6 and Lemma B.8)}
\end{aligned}
$$

According to Lemma B.7 and Lemma B.8, we know that

$$
\left| \boldsymbol{u}^\top \nabla\Psi(\boldsymbol{w}_t) \right| \leq \left(\frac{800\nu}{1 - \pi_{\boldsymbol{w}_1}(\boldsymbol{w}_t)} \right) \|\boldsymbol{u}\|_{\boldsymbol{H}_1} \leq 800\nu T \|\boldsymbol{u}\|_{\boldsymbol{H}_1} \leq 640000\nu^2 T.
$$

On the other hand, according to Lemma B.3, we have

$$
-\nabla\Psi(\boldsymbol{w}_t)^\top \boldsymbol{u} \geq \|\boldsymbol{u}\|_{\boldsymbol{H}_t}.
$$

Combining everything, we have

$$
D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \geq -800\nu \left(\ln(800\nu T) + 1\right) + \|\boldsymbol{u}\|_{\boldsymbol{H}_t},
$$

finishing the proof. $\qquad\square$

The third lemma gives a bound for the so-called stability term.

Lemma B.14. *If $\eta \leq \frac{1}{80d}$, then Algorithm 2 guarantees $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \leq 40\eta \|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}}$ for all $t \in [T]$.*

Proof. Let $F_t(\boldsymbol{w}) = \left\langle \boldsymbol{w}, \widehat{\boldsymbol{\ell}}_t \right\rangle + \frac{1}{\eta_t} D_\Psi(\boldsymbol{w}, \boldsymbol{w}_t)$. We have

$$
\begin{aligned}
F_t(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_{t+1}) &= (\boldsymbol{w}_t - \boldsymbol{w}_{t+1})^\top \widehat{\boldsymbol{\ell}}_t - \frac{1}{\eta_t} D_\Psi(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \\
&\leq (\boldsymbol{w}_t - \boldsymbol{w}_{t+1})^\top \widehat{\boldsymbol{\ell}}_t \leq \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \cdot \|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}},
\end{aligned} \tag{13}
$$

where the last line uses the nonnegativity of Bregman divergence and also Hölder's inequality. On the other hand, by Taylor's theorem, there exists a point $\boldsymbol{\xi}$ on the segment connecting $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$ such that

$$
\begin{aligned}
&F_t(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_{t+1}) \\
&= \nabla F_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}_{t+1})\nabla^2 F_t(\boldsymbol{\xi})(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) \\
&\geq \frac{1}{2}(\boldsymbol{w}_t - \boldsymbol{w}_{t+1})\nabla^2 F_t(\boldsymbol{\xi})(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) \\
&\qquad\qquad\qquad \text{(by first order optimality of } \boldsymbol{w}_{t+1} = \operatorname{argmin}_{\boldsymbol{w} \in \Omega'} F_t(\boldsymbol{w})) \\
&= \frac{1}{2\eta_t} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 \Psi(\boldsymbol{\xi})}^2.
\end{aligned} \tag{14}
$$

Next we will prove $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2 \Psi(\boldsymbol{\xi})} \geq \frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t}$. To do so, we first show $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \leq \frac{1}{2}$. It is in turn sufficient to show

$$
F_t(\boldsymbol{w}') \geq F_t(\boldsymbol{w}_t), \quad \text{for all } \boldsymbol{w}' \text{ such that } \|\boldsymbol{w}' - \boldsymbol{w}_t\|_{\boldsymbol{H}_t} = \frac{1}{2},
$$

since $\boldsymbol{w}_{t+1}$ is the minimizer of the convex function F_t. Indeed, using Taylor's theorem again and denoting $\boldsymbol{w}' - \boldsymbol{w}_t$ by $\boldsymbol{h}$, we have a point $\boldsymbol{\xi}'$ on the segment between $\boldsymbol{w}'$ and $\boldsymbol{w}_t$ such that

$$
\begin{aligned}
F_t\left(\boldsymbol{w}'\right) &= F_t\left(\boldsymbol{w}_t\right) + \nabla F_t\left(\boldsymbol{w}_t\right)^{\top}\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^{\top}\nabla^2 F_t(\boldsymbol{\xi}')\boldsymbol{h} \\
&= F_t\left(\boldsymbol{w}_t\right) + \widehat{\boldsymbol{\ell}}_t^{\top}\boldsymbol{h} + \frac{1}{2\eta_t}\|\boldsymbol{h}\|^2_{\nabla^2\Psi(\boldsymbol{\xi}')} \\
&\geq F_t\left(\boldsymbol{w}_t\right) + \widehat{\boldsymbol{\ell}}_t^{\top}\boldsymbol{h} + \frac{1}{2\eta_t}\|\boldsymbol{h}\|^2_{\boldsymbol{H}_t}\left(1 - \|\boldsymbol{w}_t - \boldsymbol{\xi}'\|_{\boldsymbol{H}_t}\right)^2 && \text{(Lemma B.1)} \\
&\geq F_t\left(\boldsymbol{w}_t\right) + \widehat{\boldsymbol{\ell}}_t^{\top}\boldsymbol{h} + \frac{1}{160\eta} && (\|\boldsymbol{h}\|_{\boldsymbol{H}_t} = \tfrac{1}{2}, \|\boldsymbol{w}_t - \boldsymbol{\xi}'\|_{\boldsymbol{H}_t} \leq \tfrac{1}{2}, \text{ and Lemma B.12}) \\
&\geq F_t\left(\boldsymbol{w}_t\right) - \|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}}\|\boldsymbol{h}\|_{\boldsymbol{H}_t} + \frac{1}{160\eta} && \text{(Hölder's inequality)} \\
&\geq F_t\left(\boldsymbol{w}_t\right) - \frac{d}{2} + \frac{1}{160\eta}. && (\|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}} \leq d\,|\langle\widetilde{w}_t, \ell_t\rangle| \leq d)
\end{aligned}
$$

Under the condition $\eta \leq \frac{1}{80d}$ we have thus shown $F_t(\boldsymbol{w}') \geq F_t(\boldsymbol{w}_t)$ and consequently $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \leq \frac{1}{2}$ and $\|\boldsymbol{w}_t - \boldsymbol{\xi}\|_{\boldsymbol{H}_t} \leq \frac{1}{2}$. Now according to Lemma B.1 again, we have

$$
\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\nabla^2\Psi(\boldsymbol{\xi})} \geq \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t}(1 - \|\boldsymbol{w}_t - \boldsymbol{\xi}\|_{\boldsymbol{H}_t}) \geq \frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t}.
$$

Plugging it into Eq. (14) and combining Eq. (13) give

$$
\|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}} \geq \frac{1}{8\eta_t}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \geq \frac{1}{40\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t},
$$

where the last inequality uses Lemma B.12. Rearranging finishes the proof. $\qquad\square$

Now we are ready to prove the bound for REG-TERM stated in Lemma B.11.

Proof of Lemma B.11. We first verify that $\boldsymbol{u}$ is in Ω'. Indeed, according to the definition of $\boldsymbol{u}$, we have

$$
\boldsymbol{w}_1 + \frac{1}{1 - \frac{1}{T}} \cdot (\boldsymbol{u} - \boldsymbol{w}_1) = \boldsymbol{u}^\star \in \Omega,
$$

which by the definition of Minkowsky function shows that $\pi_{\boldsymbol{w}_1}(\boldsymbol{u}) \leq 1 - 1/T$ and thus $\boldsymbol{u} \in \Omega'$. According to the standard analysis of Online Mirror Descent, for example, Lemma 6 of [37], we then have

$$
\left\langle\boldsymbol{w}_t, \widehat{\boldsymbol{\ell}}_t\right\rangle - \left\langle\boldsymbol{u}, \widehat{\boldsymbol{\ell}}_t\right\rangle \leq D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) + \left\langle\boldsymbol{w}_t - \boldsymbol{w}_{t+1}, \widehat{\boldsymbol{\ell}}_t\right\rangle. \tag{15}
$$

We first focus on the term $D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1})$. Taking a summation over $t = 1, 2, \dots, T$, we have

$$
\begin{aligned}
\sum_{t=1}^{T} D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) &\leq D_{\Psi_1}(\boldsymbol{u}, \boldsymbol{w}_1) + \sum_{t=1}^{T-1}\left(D_{\Psi_{t+1}}(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1})\right) \\
&\leq D_{\Psi_1}(\boldsymbol{u}, \boldsymbol{w}_1) + \sum_{i=2}^{n}\left(\frac{1}{\eta_{t_i}} - \frac{1}{\eta_{t_i - 1}}\right)D_{\Psi}(\boldsymbol{u}, \boldsymbol{w}_{t_i}),
\end{aligned}
$$

where we recall the definition of $t_1, \dots, t_n$ defined in the beginning of the proof of Lemma B.12. The first term can be bounded by

$$
\begin{aligned}
D_{\Psi_1}(\boldsymbol{u}, \boldsymbol{w}_1) = \frac{1}{\eta}D_{\Psi}(\boldsymbol{u}, \boldsymbol{w}_1) &= \frac{\Psi(\boldsymbol{u}) - \Psi(\boldsymbol{w}_1)}{\eta} - \frac{1}{\eta}\cdot\langle\nabla\Psi(\boldsymbol{w}_1), \boldsymbol{u} - \boldsymbol{w}_1\rangle \\
&\leq \frac{\Psi(\boldsymbol{u}) - \Psi(\boldsymbol{w}_1)}{\eta} && \text{(by first order optimality of } \boldsymbol{w}_1) \\
&\leq \frac{800\nu\ln T}{\eta}. && \text{(Lemma B.7)}
\end{aligned}
$$

For the second term, using $1 - \kappa \leq -\frac{1}{ad\ln(\nu T)}$ for $a = 100$ and Lemma B.12, we have

$$\frac{1}{\eta_{t_i}} - \frac{1}{\eta_{t_i-1}} \leq \frac{1-\kappa}{\eta_{t_i}} \leq -\frac{1}{5\eta ad\ln(\nu T)}.$$

Therefore,

$$\sum_{t=1}^{T} D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_t) - D_{\Psi_t}(\boldsymbol{u}, \boldsymbol{w}_{t+1})$$

$$\leq \frac{800\nu\ln T}{\eta} - \sum_{i=2}^{n} \frac{1}{5\eta ad\ln(\nu T)} \cdot D_{\Psi}(\boldsymbol{u}, \boldsymbol{w}_{t_i})$$

$$\leq \tilde{\mathcal{O}}\left(\frac{\nu}{\eta}\right) - \frac{1}{5\eta ad\ln(\nu T)} \cdot \sum_{i=2}^{n} \left(\|\boldsymbol{u}\|_{\boldsymbol{H}_{t_i}} - 800\nu - 800\nu\ln(800\nu T)\right) \qquad \text{(Lemma B.13)}$$

$$= \tilde{\mathcal{O}}\left(\frac{\nu}{\eta}\right) - \frac{1}{5\eta ad\ln(\nu T)} \sum_{i=2}^{n} \|\boldsymbol{u}\|_{\boldsymbol{H}_{t_i}}.$$

For the second term in Eq. (15), that is, $\left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, \widehat{\boldsymbol{\ell}}_t \right\rangle$, taking summation over $t \in [T]$ we have

$$\sum_{t=1}^{T} \left\langle \boldsymbol{w}_t - \boldsymbol{w}_{t+1}, \widehat{\boldsymbol{\ell}}_t \right\rangle \leq \sum_{t=1}^{T} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_{\boldsymbol{H}_t} \|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}} \qquad \text{(Hölder's inequality)}$$

$$\leq 40\eta \sum_{t=1}^{T} \|\widehat{\boldsymbol{\ell}}_t\|_{\boldsymbol{H}_t^{-1}}^2 \qquad \text{(Lemma B.14)}$$

$$= 40\eta \sum_{t=1}^{T} d^2 \langle \widetilde{w}_t, \ell_t \rangle^2 \, \boldsymbol{s}_t^\top \boldsymbol{H}_t^{1/2} \boldsymbol{H}_t^{-1} \boldsymbol{H}_t^{1/2} \boldsymbol{s}_t$$

$$\leq 40\eta \sum_{t=1}^{T} d^2 \left| \langle \widetilde{w}_t, \ell_t \rangle \right| = 40\eta d^2 \overline{L}_T.$$

Combining everything finishes the proof. $\qquad\square$

B.3.3 Proof of Theorem 3.1

To prove Theorem 3.1, we first prove the following main lemma.

Lemma B.15. *Algorithm 2 with $\eta \leq \frac{1}{640aCd^2\ln(\nu T)\ln(C/\delta)}$ guarantees that with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle$$

$$\leq \tilde{\mathcal{O}}\left(\frac{\nu}{\eta} + \eta d^2 \overline{L}_T + \sqrt{\overline{L}_u \ln(1/\delta)}\right) + (\sqrt{\nu} + \rho)\left(161Cd\sqrt{\ln(C/\delta)\mathring{L}_T} - \frac{1}{10\eta ad\ln(\nu T)}\right),$$

where $a = 100$, $C = \Theta(\ln^2(d\nu T))$ is defined in Lemma B.9, and we recall all other notations defined in Equations (8)-(10).

24

Proof. Recall the decomposition of regret shown in [Eq. (7)]. Combining the result of [Lemma B.10] and [Lemma B.11], we have when $\eta \leq \frac{1}{80d}$,

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}}\left(\frac{\nu}{\eta}\right) - \frac{\sum_{s \in \mathcal{S}'} \|u\|_{H_s}}{5\eta a d \ln(\nu T)} + 40\eta d^2 \overline{L}_T + 64 C d \left(\sqrt{\nu} + \rho\right) \ln(C/\delta)$$

$$+ 161 C d \sqrt{(\nu + \rho^2)\, \mathring{L}_T \ln(C/\delta)} + C\sqrt{32 \overline{L}_u \ln(C/\delta)}.$$

$$= \tilde{\mathcal{O}}\left(\frac{\nu}{\eta} + \eta d^2 \overline{L}_T + \sqrt{\overline{L}_u \ln(C/\delta)}\right) - \frac{\sum_{s \in \mathcal{S}'} \|u\|_{H_s}}{5\eta a d \ln(\nu T)}$$

$$+ 64 C d \left(\sqrt{\nu} + \rho\right) \ln(C/\delta) + 161 C \sqrt{d^2 (\nu + \rho^2)\, \mathring{L}_T \ln(C/\delta)}. \qquad (16)$$

Now consider the value of $\rho = \|u\|_{H_{t^\star}}$ where $t^\star \in \mathrm{argmax}_{t \in [T]} \|u\|_{H_t}$, compared to the negative term above. Suppose $t^\star \in \mathcal{S}$, then we have

$$\rho \leq \max\left\{\|u\|_{H_1}, \sum_{s \in \mathcal{S}'} \|u\|_{H_s}\right\} \leq 800\nu + \sum_{s \in \mathcal{S}'} \|u\|_{H_s},$$

where we use [Lemma B.8] again to bound $\|u\|_{H_1}$. On the other hand, if $t^\star \notin \mathcal{S}$, then according to the update rule of $\mathcal{S}$ in [Algorithm 2], we have $H_{t^\star} \preceq H_1 + \sum_{s \in \mathcal{S}'} H_s$, which means

$$\rho = \sqrt{\|u\|_{H_{t^\star}}^2} \leq \sqrt{\|u\|_{H_1}^2 + \sum_{s \in \mathcal{S}'} \|u\|_{H_s}^2} \leq 800\nu + \sum_{s \in \mathcal{S}'} \|u\|_{H_s}.$$

Therefore, we continue to bound the last three terms in [Eq. (16)] as

$$\frac{800\nu - \rho}{5\eta a d \ln(\nu T)} + 64 C d \left(\sqrt{\nu} + \rho\right) \ln(C/\delta) + 161 C d \sqrt{(\nu + \rho^2)\, \mathring{L}_T \ln(C/\delta)}$$

$$\leq \mathcal{O}\left(\frac{\nu}{\eta}\right) - \frac{\sqrt{\nu} + \rho}{5\eta a d \ln(\nu T)} + 64 C d \left(\sqrt{\nu} + \rho\right) \ln(C/\delta) + 161 C d \sqrt{(\nu + \rho^2)\, \mathring{L}_T \ln(C/\delta)}$$

$$\leq \mathcal{O}\left(\frac{\nu}{\eta}\right) - \frac{\sqrt{\nu} + \rho}{10\eta a d \ln(\nu T)} + 161 C d \sqrt{(\nu + \rho^2)\, \mathring{L}_T \ln(C/\delta)} \qquad (\eta \leq \tfrac{1}{640 a C d^2 \ln(\nu T) \ln(C/\delta)})$$

$$\leq \mathcal{O}\left(\frac{\nu}{\eta}\right) + (\sqrt{\nu} + \rho)\left(161 C d \sqrt{\ln(C/\delta)\mathring{L}_T} - \frac{1}{10\eta a d \ln(\nu T)}\right).$$

Plugging this back into [Eq. (16)] finishes the proof. $\qquad \square$

Now we are ready to prove the main theorem. For convenience, we restate the theorem below.

Theorem B.16. *[Algorithm 2] with an appropriate choice of η ensures that with probability at least $1 - \delta$:*

$$\mathrm{Reg} = \begin{cases} \tilde{\mathcal{O}}(d^2 \nu \sqrt{T \ln \frac{1}{\delta}} + d^2 \nu \ln \frac{1}{\delta}), & \text{against an oblivious adversary;} \\ \tilde{\mathcal{O}}(d^2 \nu \sqrt{dT \ln \frac{1}{\delta}} + d^3 \nu \ln \frac{1}{\delta}), & \text{against an adaptive adversary.} \end{cases}$$

Moreover, if $\langle w, \ell_t \rangle \geq 0$ for all $w \in \Omega$ and all t, then T in the bounds above can be replaced by $L^\star = \min_{u \in \Omega} \sum_{t=1}^{T} \langle u, \ell_t \rangle$, that is, the total loss of the best action.

Proof. Using [Lemma B.15] and the fact that $|\langle \widetilde{w}_t, \ell_t \rangle| \leq 1$ and $|\langle u, \ell_t \rangle| \leq 1$ for all $t \in [T]$, we have

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}}\left(\frac{\nu}{\eta} + \eta d^2 T + \sqrt{T \ln \frac{1}{\delta}}\right) + (\sqrt{\nu} + \rho)\left(161 C d \sqrt{T \ln(C/\delta)} - \frac{1}{10\eta a d \ln(\nu T)}\right).$$

With

$$\eta = \min\left\{\frac{1}{640 a C d^2 \ln(\nu T) \ln(C/\delta)}, \frac{1}{1610 a C d^2 \ln(\nu T)\sqrt{T \ln(C/\delta)}}\right\},$$

25

the last term becomes nonpositive, and we arrive at

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}} \left(d^2 \nu \sqrt{T \ln \frac{1}{\delta}} + d^2 \nu \ln \frac{1}{\delta} \right), \tag{17}$$

for any fixed $u^\star \in \Omega$, which completes the proof for the oblivious case. To obtain a bound for an adaptive adversary, we discrete the feasible set Ω and then take a union bound. Specifically, define B_Ω as follows:

$$B_\Omega \triangleq \lceil \alpha \rceil \lceil \beta \rceil, \quad \alpha \triangleq \max_{w, w' \in \Omega} \|w - w'\|_\infty, \quad \beta \triangleq \max_{\ell \in \Omega^\circ} \|\ell\|_\infty,$$

where $\Omega^\circ \triangleq \{\ell : |\langle w, \ell \rangle| \leq 1, \ \forall w \in \Omega\}$ is the set of feasible loss vectors. Then we discretize Ω into a finite set $\overline{\Omega}$ of $(B_\Omega T)^d$ points, such that for any $u^\star \in \Omega$, there exists $\overline{u} \in \overline{\Omega}$, such that $\|\overline{u} - u^\star\|_\infty \leq \frac{1}{\lceil \beta \rceil T}$. This means that

$$\left| \sum_{t=1}^{T} \langle \overline{u} - u^\star, \ell_t \rangle \right| \leq \sum_{t=1}^{T} \frac{d}{\lceil \beta \rceil T} \cdot \max_i \ell_{t,i} \leq d.$$

Therefore, it suffices to only consider regret against the points in $\overline{\Omega}$. Taking a union bound and replacing δ with $\frac{\delta}{(B_\Omega T)^d}$ in Eq. (17) finish the proof for the worst-case bound for adaptive adversaries.

In the remaining of the proof, we show that if $\langle w, \ell_t \rangle \in [0, 1]$ for all $w \in \Omega$ and $t \in [T]$, T can be replaced by $L^\star$ in both bounds. As $\langle w, \ell_t \rangle$ is always positive, we have $\mathbb{E}_t[|\langle \widetilde{w}_t, \ell_t \rangle|] = \mathbb{E}_t[\langle \widetilde{w}_t, \ell_t \rangle] = \langle w_t, \ell_t \rangle$, $\overline{L}_u = \sum_{t=1}^{T} \langle u, \ell_t \rangle \leq L^\star + 2$, and $\overline{L}_T = L_T = \sum_{t=1}^{T} \langle \widetilde{w}_t, \ell_t \rangle$. Using standard Freedman's inequality, we have with probability at least $1 - \delta$,

$$\mathring{L}_T - \overline{L}_T \leq \frac{\mathring{L}_T}{2} + 3 \ln(1/\delta).$$

Rearranging gives

$$\mathring{L}_T \leq 2 L_T + 6 \ln(1/\delta).$$

Using Lemma B.15 again, we have

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}} \left(\frac{\nu}{\eta} + \eta d^2 L_T + \sqrt{L^\star \ln \frac{1}{\delta}} \right)$$

$$+ (\sqrt{\nu} + \rho) \left(161 C d \sqrt{\ln(C/\delta) \left(2 L_T + 6 \ln \frac{1}{\delta} \right)} - \frac{1}{10 \eta a d \ln(\nu T)} \right).$$

With $\eta = \min \left\{ \frac{1}{640 a C d^2 \ln(\nu T) \ln(C/\delta)}, \frac{1}{1610 a C d^2 \ln(\nu T) \sqrt{(2 L_T + 6 \ln(1/\delta)) \ln(C/\delta)}} \right\}$, the last term becomes nonpositive, and we arrive at

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}} \left(d^2 \nu \sqrt{L_T \ln \frac{1}{\delta}} + \sqrt{L^\star \ln \frac{1}{\delta}} + d^2 \nu \ln \frac{1}{\delta} \right)$$

Solving the quadratic inequality in terms of $\sqrt{L_T}$ gives the following high probability regret bound

$$\sum_{t=1}^{T} \langle \widetilde{w}_t - u^\star, \ell_t \rangle \leq \tilde{\mathcal{O}} \left(d^2 \nu \sqrt{L^\star \ln \frac{1}{\delta}} + d^2 \nu \ln \frac{1}{\delta} \right).$$

This finishes the proof for the case with oblivious adversaries, and the case with adaptive adversaries is again by taking a union bound as done earlier. □

Remark 2. The tuning of η in the proof above depends on the unknown quantity L_T. In fact, the issue seems even more severe than that pointed out in Remark 1 because L_T depends on the algorithm's behavior, which in turns depends on η itself. We point out that, however, this can again be addressed using a doubling trick, making the algorithm completely parameter-free. We omit the details but refer the reader to Lee et al. [27, Algorithm 4] for very similar ideas.

C Omitted details for Section 4

C.1 Preliminary

In this section, we introduce the concept of *occupancy measure* (used in previous works already; see [25]), which helps reformulate adversarial MDP problems in a way very similar to adversarial MAB problems. For a state x, let $k(x)$ denote the index of the layer to which state x belongs. Given a policy π and a transition function P, we define occupancy measure $w^{P,\pi} \in \mathbb{R}^{X \times A \times X}$ as follows:

$$w^{P,\pi}(x, a, x') = \mathbb{P}[x_k = x, a_k = a, x_{k+1} = x'|P, \pi],$$

where $k = k(x)$. In other words, $w^{P,\pi}(x, a, x')$ is the probability of visiting the triple (x, a, x') if we execute policy π in an MDP with transition function P.

According to this definition, we have the following two properties for any occupancy measure w. First, based on the layered structure, we know that each layer is visited exactly once in each episode, which means for each $k = 0, 1, \ldots, J$, we have

$$\sum_{x \in X_k, a \in A, x' \in X_{k+1}} w(x, a, x') = 1. \tag{18}$$

Second, the probability of entering one state when coming from the previous layer equals to the probability of leaving the state to the next layer. Therefore, for each $k = 1, 2, \ldots, J - 1$, we have

$$\sum_{x' \in X_{k-1}, a \in A} w(x', a, x) = \sum_{x' \in X_{k+1}, a \in A} w(x, a, x'), \tag{19}$$

for all $x \in X_k$.

Moreover, the following lemma shows that if w satisfies the above two properties, then w is an occupancy measure with respect to some transition function P^w and policy π^w.

Lemma C.1 (Lemma 3.1 in [34]). *For any $w \in [0, 1]^{|X| \times |A| \times |X|}$, it satisfies Eq. (18) and Eq. (19) if and only if it is a valid occupancy measure associated with the following induced transition function P^w and policy π^w:*

$$P^w(x'|x, a) = \frac{w(x, a, x')}{\sum_{y \in X_{k(x)+1}} w(x, a, y)}, \ \pi^w(a|x) = \frac{\sum_{x' \in X_{k(x)+1}} w(x, a, x')}{\sum_{a' \in A} \sum_{x' \in X_{k(x)+1}} w(x, a', x')}.$$

Following [25], we denote by Δ the set of all valid occupancy measures. For a fixed transition function, we denote by $\Delta(P) \subseteq \Delta$ the set of occupancy measures whose induced transition function P^w is exactly P. In addition, we denote by $\Delta(\mathcal{P}) \subseteq \Delta$ the set of occupancy measures whose induced transition function P^w belongs to a set of transition functions $\mathcal{P}$. With a slightly abuse of notation, we define $w(x, a) = \sum_{x' \in X_{k(x)+1}} w(x, a, x')$ for all $x \neq x_J$ and $a \in A$. Using the notations introduced above, we know that the expected loss of using policy π at round t is exactly $\langle w^{P,\pi}, \ell_t \rangle \triangleq \sum_{x,a} w^{P,\pi}(x, a)\ell_t(x, a)$. Let π_t be the policy chosen at round t. Then the total expected loss (with respect to randomness of the transition function) is $\sum_{t=1}^T \langle w^{P,\pi_t}, \ell_t \rangle$ and the total regret can be written as:

$$\text{Reg} = \sum_{t=1}^T \ell_t(\pi_t) - \min_\pi \sum_{t=1}^T \ell_t(\pi) = \sum_{t=1}^T \langle w_t - u^\star, \ell_t \rangle = L_T - L^\star, \tag{20}$$

where $u^\star = w^{P,\pi^\star}$ is the occupancy measure induced by the optimal policy $\pi^\star = \arg\min_\pi \sum_{t=1}^T \ell_t(\pi)$, $w_t = w^{P,\pi_t}$, $L_T \triangleq \sum_{t=1}^T \langle w_t, \ell_t \rangle$, and $L^\star \triangleq \sum_{t=1}^T \langle u^\star, \ell_t \rangle$. When the regret is written in this way, it is clear that the problem is very similar to MAB or linear bandits with $\Delta(P)$ being the decision set and ℓ_t parametrizing the linear loss function at time t.

C.2 Algorithm for MDPs

In this section, we introduce our algorithm that achieves high-probability small-loss regret bound for the MDP setting. The full pseudocode of the algorithm is shown in Algorithm 4. The algorithm is very similar to UOB-REPS introduced in [25], except for the following two modifications.

Algorithm 4 Upper Occupancy Bound Log Barrier Policy Search

Input: state space X, action space A, learning rate η, and confidence parameter δ.

Define: $\kappa = e^{\frac{1}{T \ln T}}$, COMP-UOB is Algorithm 3 of [25], and

$$\Omega = \left\{ \hat{w} : \hat{w}(x, a, x') \geq \frac{1}{T^3 |X|^2 |A|}, \forall k \in \{0, 1, \ldots, J-1\}, x \in X_k, a \in A, x' \in X_{k+1} \right\}.$$

Initialization: Set epoch index $i = 1$ and confidence set $\mathcal{P}_1$ as the set of all transition functions. For all $k = 0, \ldots, J-1, (x, a, x') \in X_k \times A \times X_{k+1}$, set

$$\widehat{w}_1(x, a, x') = \frac{1}{|X_k||A||X_{k+1}|}, \quad \pi_1 = \pi^{\widehat{w}_1}, \quad \eta_1(x, a) = \eta, \quad \rho_1(x, a) = 2|X_k||A|,$$

$$\phi_1(x, a) = \text{COMP-UOB}(\pi_1, x, a, \mathcal{P}_1), \quad N_0(x, a) = N_1(x, a) = G_0(x'|x, a) = G_1(x'|x, a) = 0.$$

1 **for** $t = 1, 2, \ldots, T$ **do**
2   Execute policy π_t for J steps and obtain trajectory $x_k, a_k, \ell_t(x_k, a_k)$ for $k = 0, \ldots, J-1$.
3   Construct loss estimators for all $(x, a) \in X \times A$:

$$\widehat{\ell}_t(x, a) = \frac{\ell_t(x, a)}{\phi_t(x, a)} \mathbb{1}_t(x, a), \quad \text{where} \quad \mathbb{1}_t(x, a) = \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}. \qquad (21)$$

4   Update counters: for each $k = 0, 1, \ldots, J-1$,

$$N_i(x_k, a_k) \leftarrow N_i(x_k, a_k) + 1, \quad G_i(x_{k+1}|x_k, a_k) \leftarrow G_i(x_{k+1}|x_k, a_k) + 1.$$

5   **if** $\exists k, N_i(x_k, a_k) \geq \max\{1, 2N_{i-1}(x_k, a_k)\}$ **then**
6   Increase epoch index $i \leftarrow i + 1$.
7   Initialize new counters: $N_i = N_{i-1}, G_i = G_{i-1}$ (copy all entries).
8   Compute confidence set

$$\mathcal{P}_i = \big\{ \hat{P} : \big| \hat{P}(x'|x, a) - \bar{P}_i(x'|x, a) \big| \leq \epsilon_i(x'|x, a),$$

$$\forall (x, a, x') \in X_k \times A \times X_{k+1}, k = 0, 1, \ldots, J-1 \big\},$$

where $\bar{P}_i(x'|x, a) = \frac{G_i(x'|x, a)}{\max\{1, N_i(x, a)\}}$ and

$$\epsilon_i(x'|x, a) \triangleq 4\sqrt{\frac{\bar{P}_i(x'|x, a) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x, a) - 1\}}} + \frac{28 \ln\left(\frac{T|X||A|}{\delta}\right)}{3 \max\{1, N_i(x, a) - 1\}}.$$

9   Compute $\widehat{w}_{t+1} = \text{argmin}_{w \in \Delta(\mathcal{P}_i) \cap \Omega} \left\{ \langle w, \widehat{\ell}_t \rangle + D_{\psi_t}(w, \widehat{w}_t) \right\}$, where

$$\psi_t(w) = \sum_{k=0}^{J-1} \sum_{(x, a, x') \in X_k \times A \times X_{k+1}} \frac{1}{\eta_t(x, a)} \ln\left(\frac{1}{w(x, a, x')}\right). \qquad (22)$$

10   Update policy $\pi_{t+1} = \pi^{\widehat{w}_{t+1}}$.
11   **for** *each* $(x, a) \in X \times A$ **do**
12   Update upper occupancy bound:

$$\phi_{t+1}(x, a) = \max_{\hat{P} \in \mathcal{P}_i} w^{\hat{P}, \pi_{t+1}}(x, a) = \text{COMP-UOB}(\pi_{t+1}, x, a, \mathcal{P}_i). \qquad (23)$$

13   **if** $\frac{1}{\phi_{t+1}(x, a)} \geq \rho_t(x, a)$ **then** $\rho_{t+1}(x, a) = \frac{2}{\phi_{t+1}(x, a)}, \eta_{t+1}(x, a) = \eta_t(x, a) \cdot \kappa$.
14   **else** $\rho_{t+1}(x, a) = \rho_t(x, a), \eta_{t+1}(x, a) = \eta_t(x, a)$.

First, in [25], they propose a loss estimator akin to the importance-weighted estimator using the so-called *upper occupancy bound*, denoted by $\phi_t(x, a)$ in our notation. Indeed, the actual probability

$w_t(x, a)$ of visiting state-action pair (x, a) is unknown (due to the unknown transition), and thus standard unbiased importance-weighted estimators do not apply directly. Instead, since the algorithm maintains a confidence set $\mathcal{P}_i$ (for epoch i) of all the plausible transition functions based on observations, one can calculate the largest probability of visiting state-action pair (x, a) under policy π_t, among all the plausible transition functions, which is exactly the definition of $\phi_t(x, a)$ and can be computed efficiently via the sub-routine COMP-UOB as shown in [25]. In addition, Jin et al. [25] also apply the idea of implicit exploration from [31] and introduce an extra bias with a parameter $\gamma > 0$, leading to the following loss estimator:

$$\widehat{\ell}_t(x, a) = \frac{\ell_t(x, a)}{\phi_t(x, a) + \gamma} \mathbb{1}_t(x, a),$$

which is crucial for them to derive a high-probability bound. As one can see in Eq. (21), the first difference of our algorithm is that we remove this implicit exploration (that is, $\gamma = 0$), similarly to our MAB algorithm in Section 2. As we later explain in Appendix C.4, removing this implicit exploration is important for obtaining a small-loss bound.

Second, while UOB-REPS uses the entropy regularizer with a fixed learning rate, we use the log-barrier regularizer with time-varying and individual learning rates for each state-action pair, defined in Eq. (22), which is a direct generalization of Algorithm 1 for MAB. The way we increase the learning rate is also essentially identical to the MAB case; see the last part of Algorithm 4. We also point out that the analogue of the clipped simplex used in Algorithm 1 is now $\Delta(\mathcal{P}_i) \cap \Omega$ where $\Delta(\mathcal{P}_i)$ is the set of occupancy measures with induced transition functions in the confidence set $\mathcal{P}_i$, and Ω (defined at the beginning of Algorithm 4) contains all $\widehat{w}$ with each entry not smaller than $1/(T^3|X|^2|A|)$, which ensures that the learning rates cannot be increased by too many times.

C.3 Proof of Theorem 4.1

In this section, we analyze Algorithm 4 and prove Theorem 4.1. We start with decomposing the regret into five terms (recall the definitions of w_t and $u^\star$ in Eq. (20) and $\widehat{w}_t$ and $\widehat{\ell}_t$ in Algorithm 4):

$$\sum_{t=1}^{T} \langle w_t - u^\star, \ell_t \rangle = \underbrace{\sum_{t=1}^{T} \langle w_t - \widehat{w}_t, \ell_t \rangle}_{\text{ERROR}} + \underbrace{\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t \right\rangle}_{\text{BIAS-1}} + \underbrace{\sum_{t=1}^{T} \left\langle \widehat{w}_t - u, \widehat{\ell}_t \right\rangle}_{\text{REG-TERM}}$$

$$+ \underbrace{\sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle}_{\text{BIAS-2}} + \underbrace{\sum_{t=1}^{T} \langle u - u^\star, \ell_t \rangle}_{\text{BIAS-3}}.$$

Here u is defined as

$$u = \left(1 - \frac{1}{T}\right) u^\star + \frac{1}{T|A|} \sum_{a \in A} w^{P_0, \pi_a}, \tag{24}$$

where π_a is the policy that chooses action a at every state, and the definition of the transition function P_0 is deferred to Lemma C.4. Note that u is random in the case with adaptive adversaries.

In the remaining of this subsection, we first provide a few useful lemmas in Appendix C.3.1, and then bound ERROR in Appendix C.3.2, BIAS-1 in Appendix C.3.3, BIAS-2 in Appendix C.3.4, and REG-TERM in Appendix C.3.5. Note that BIAS-3 can be trivially bounded by J as

$$\text{BIAS-3} = \sum_{t=1}^{T} \langle u - u^\star, \ell_t \rangle \le \frac{1}{T|A|} \sum_{a \in A} \sum_{t=1}^{T} \langle w^{\pi_a, P_0}, \ell_t \rangle \le J. \tag{25}$$

We finally put everything together and prove Theorem 4.1 in Appendix C.3.6.

C.3.1 Useful lemmas

The first two lemmas are from [25].

Lemma C.2 (Lemma 2 in [25]). *With probability at least $1 - 4\delta$, we have for all $k = 0, 1, \ldots, J-1$ and $(x, a, x') \in X_k \times A \times X_{k+1}$,*

$$\left|P(x'|x, a) - \bar{P}_i(x'|x, a)\right| \leq \frac{\epsilon_i(x'|x, a)}{2}. \tag{26}$$

Consequently, we have $P \in \mathcal{P}_i$ for all i.

Lemma C.3 (Lemma 10 in [25]). *With probability at least $1 - \delta$, we have for all $k = 0, \ldots, J-1$,*

$$\sum_{t=1}^{T} \sum_{x \in X_k, a \in A} \frac{w_t(x, a)}{\max\{1, N_{i_t}(x, a)\}} = \tilde{\mathcal{O}}\left(|X_k| \cdot |A| + \ln \frac{1}{\delta}\right)$$

$$\sum_{t=1}^{T} \sum_{x \in X_k, a \in A} \frac{w_t(x, a) - \mathbb{1}_t(x, a)}{\sqrt{\max\{1, N_{i_t}(x, a)\}}} \leq \sum_{t=1}^{T} \sum_{x \in X_k, a \in A} \frac{w_t(x, a)}{\max\{1, N_{i_t}(x, a)\}} + \tilde{\mathcal{O}}\left(\ln \frac{1}{\delta}\right)$$

$$\leq \tilde{\mathcal{O}}\left(|X_k| \cdot |A| + \ln \frac{1}{\delta}\right),$$

where i_t is the index of the epoch to which episode t belongs.

Next, we prove a lemma showing that there exists a transition function P_0 that always lies in the confidence set $\mathcal{P}_i$ of the algorithm, such that for any action $a \in A$ and any two states x, x' in consecutive layers, the probability of reaching x' by taking action a at state x is at least $\frac{1}{T|X|}$.

Lemma C.4. *With probability at least $1 - 4\delta$, there exists $P_0 \in \cap_i \mathcal{P}_i$ such that for all $k < J, x \in X_k, a \in A$, and $x' \in X_{k+1}$, we have $P_0(x'|x, a) \geq \frac{1}{T|X|}$.*

Proof. The construction of P_0 is as follows. First we start with $P_0 = P$. Then for each fixed (x, a), we focus on the distribution $P_0(\cdot|x, a)$. In particular, for all $x' \in X_{k(x)+1}$ such that $P_0(x'|x, a) < \frac{1}{T|X|}$, we move the weight from the largest entry of $P_0(\cdot|x, a)$ to this entry so that $P_0(x'|x, a) = \frac{1}{T|X|}$ and $P_0(\cdot|x, a)$ remains a valid distribution. Repeat the same for all (x, a) pairs finishes the construction of P_0.

Clearly, P_0 satisfies $P_0(x'|x, a) \geq \frac{1}{T|X|}$, and it remains to show $P_0 \in \mathcal{P}_i$ for all i. To this end, we first note that

$$|P_0(x'|x, a) - P(x'|x, a)| \leq \frac{|X_{k(x')}|}{T|X|} \leq \frac{1}{T} \leq \frac{\epsilon_i(x'|x, a)}{2}$$

holds for all $k = 0, 1, \ldots, J-1$ and $(x, a, x') \in X_k \times A \times X_{k+1}$. Combining this with Eq. (26) then shows that $\left|P_0(x'|x, a) - \bar{P}_i(x'|x, a)\right| \leq \epsilon_i(x'|x, a)$, indicating that P_0 is indeed in $\mathcal{P}_i$ by the definition of $\mathcal{P}_i$. $\qquad \square$

The next lemma shows that the upper occupancy bound for each state-action pair is lower bounded.

Lemma C.5. *We have $\phi_t(x, a) \geq \frac{1}{T^3|X|^2|A|}$ for all $x \in X$ and $a \in A$.*

Proof. This is simply by the definition of ϕ_t in Eq. (23) and the definition of Ω: $\phi_t(x, a) \geq \widehat{w}_t(x, a) \geq \frac{1}{T^3|X|^2|A|}$. $\qquad \square$

The last lemma is an improvement of [25, Lemma 4] and is important for bounding ERROR and BIAS-1 in terms of $\sqrt{L_T}$, as opposed to $\sqrt{T}$ (which is the case in [25]).

Lemma C.6. *With probability at least $1 - 6\delta$, for any t and any collection of transition functions $\{P_t^x\}_{x \in X}$ such that $P_t^x \in \mathcal{P}_{i_t}$ for all x (where i_t is the index of the epoch to which episode t belongs), we have*

$$\sum_{t=1}^{T} \sum_{x \in X, a \in A} \left|w^{P_t^x, \pi_t}(x, a) - w_t(x, a)\right| \ell_t(x, a)$$

$$= \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T \ln \frac{1}{\delta}} + |X|^5|A|^2 \ln \frac{1}{\delta} + |X|^4|A| \ln^2 \frac{1}{\delta}\right).$$

Proof. The proof is technical but mostly follows the same ideas of that for [25, Lemma 4]. We first assume that the events of Lemma C.2 and Lemma C.3 hold, which happens with probability at least $1 - 5\delta$. According to the proof of [25, Lemma 4] (specifically their Eq. (15)), we have for any pair (x, a),

$$\left| w^{P_t^x, \pi_t}(x, a) - w_t(x, a) \right| \cdot \ell_t(x, a)$$

$$\leq \sum_{m=0}^{k(x)-1} \sum_{x_m, a_m, x_{m+1}} \epsilon_{i_t}^{\star}(x_{m+1}|x_m, a_m) w_t(x_m, a_m) w^{P_t^x, \pi_t}(x, a|x_{m+1}) \cdot \ell_t(x, a), \qquad (27)$$

where $\epsilon_{i_t}^{\star}(x'|x, a) = \mathcal{O}\left(\sqrt{\frac{P(x'|x,a)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x,a)\}}} + \frac{\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x,a)\}} \right)$, and for an occupancy measure w, $w(x, a|x')$ denotes the probability of encountering the pair (x, a) given that x' was visited earlier, under policy π^w and P^w. By their Eq. (16), we also have

$$|w^{P_t^x, \pi_t}(x, a|x_{m+1}) - w_t(x, a|x_{m+1})|$$

$$\leq \pi_t(a|x) \sum_{h=m+1}^{k(x)-1} \sum_{x'_h, a'_h, x'_{h+1}} \epsilon_{i_t}^{\star}(x'_{h+1}|x'_h, a'_h) w_t(x'_h, a'_h|x_{m+1}), \qquad (28)$$

Combining Eq. (27) and Eq. (28), summing over all t and (x, a) and using shorthands $z_m \triangleq (x_m, a_m, x_{m+1})$ and $z'_h \triangleq (x'_h, a'_h, x'_{h+1})$, we have

$$\sum_{t=1}^{T} \sum_{x \in X, a \in A} |w^{P_t^x, \pi_t}(x, a) - w_t(x, a)| \cdot \ell_t(x, a)$$

$$\leq \sum_{t,x,a} \sum_{m=0}^{k(x)-1} \sum_{z_m} \epsilon_{i_t}^{\star}(x_{m+1}|x_m, a_m) w_t(x_m, a_m) w_t(x, a|x_{m+1}) \ell_t(x, a)$$

$$+ \sum_{t,x,a} \sum_{m=0}^{k(x)-1} \sum_{z_m} \epsilon_{i_t}^{\star}(x_{m+1}|x) w_t(x_m, a_m) \cdot \left(\pi_t(a|x) \sum_{h=m+1}^{k(x)-1} \sum_{z'_h} \epsilon_{i_t}^{\star}(x'_{h+1}|x'_h, a'_h) w_t(x'_h, a'_h|x_{m+1}) \right)$$

$$(\ell_t(x, a) \leq 1)$$

$$= \sum_{t} \sum_{k<J} \sum_{m=0}^{k-1} \sum_{z_m} \epsilon_{i_t}^{\star}(x_{m+1}|x_m, a_m) w_t(x_m, a_m) \sum_{x \in X_k, a \in A} w_t(x, a|x_{m+1}) \ell_t(x, a)$$

$$+ \sum_{t} \sum_{0 \leq m < h < k < J} \sum_{z_m, z'_h} \epsilon_{i_t}^{\star}(x_{m+1}|x) w_t(x_m, a_m) \epsilon_{i_t}^{\star}(x'_{h+1}|x'_h, a'_h) w_t(x'_h, a'_h|x_{m+1}) \cdot \left(\sum_{x \in X_k, a \in A} \pi_t(a|x) \right)$$

$$= \sum_{0 \leq m < k < J} \sum_{t, z_m} \epsilon_{i_t}^{\star}(x_{m+1}|x_m, a_m) w_t(x_m, a_m) \sum_{x \in X_k, a \in A} w_t(x, a|x_{m+1}) \ell_t(x, a)$$

$$+ \sum_{0 \leq m < h < k < J} |X_k| \sum_{t, z_m, z'_h} \epsilon_{i_t}^{\star}(x_{m+1}|x) w_t(x_m, a_m) \epsilon_{i_t}^{\star}(x'_{h+1}|x'_h, a'_h) w_t(x'_h, a'_h|x_{m+1})$$

$$\leq \sum_{0 \leq m < k < J} \sum_{t, z_m} \epsilon_{i_t}^{\star}(x_{m+1}|x_m, a_m) w_t(x_m, a_m) \sum_{x \in X_k, a \in A} w_t(x, a|x_{m+1}) \ell_t(x, a)$$

$$+ |X| \sum_{0 \leq m < h < J} \sum_{t, z_m, z'_h} \epsilon_{i_t}^{\star}(x_{m+1}|x) w_t(x_m, a_m) \epsilon_{i_t}^{\star}(x'_{h+1}|x'_h, a'_h) w_t(x'_h, a'_h|x_{m+1})$$

$$\triangleq B_1 + |X| B_2.$$

It remains to bound B_1 and B_2. First, B_2 is exactly the same as in the proof of [25, Lemma 4]. Below, we outline the proof with the dependence on all parameters explicit (indeed, this is hidden in their

proof). First, according to their analysis, B_2 is bounded by

$$\tilde{\mathcal{O}}\left(\sum_{0\leq m<h<J}\sum_{t,z_m,z'_h}\sqrt{\frac{P(x_{m+1}|x_m,a_m)\ln\frac{1}{\delta}}{\max\{1,N_{i_t}(x_m,a_m)\}}}w_t(x_m,a_m)\sqrt{\frac{P(x'_{h+1}|x'_h,a'_h)\ln\frac{1}{\delta}}{\max\{1,N_{i_t}(x'_h,a'_h)\}}}w_t(x'_h,a'_h|x_{m+1})\right)$$

$$+\tilde{\mathcal{O}}\left(\sum_{0\leq m<h<J}\sum_{t,z_m,z'_h}\frac{w_t(x_m,a_m)\ln\frac{1}{\delta}}{\max\{1,N_{i_t}(x_m,a_m)\}}\right)+\tilde{\mathcal{O}}\left(\sum_{0\leq m<h<J}\sum_{t,z_m,z'_h}\frac{w_t(x'_h,a'_h)\ln\frac{1}{\delta}}{\max\{1,N_{i_t}(x'_h,a'_h)\}}\right).$$

They show that the first term is bounded by $\tilde{\mathcal{O}}(|X|^2|A|\ln^2(1/\delta))$. For the second term, we have

$$\sum_{0\leq m<h<J}\sum_{t,z_m,z'_h}\frac{w_t(x_m,a_m)\ln\frac{1}{\delta}}{\max\{1,N_{i_t}(x_m,a_m)\}}$$

$$\leq\left(\sum_{h=0}^{J-1}|X_h|\cdot|A|\cdot|X_{h+1}|\ln\frac{1}{\delta}\right)\sum_{m=0}^{J-1}|X_{m+1}|\cdot\sum_{t,x\in X_m,a\in A}\frac{w_t(x_m,a_m)}{\max\{1,N_{i_t}(x_m,a_m)\}}$$

$$\leq\mathcal{O}\left(|X|^2|A|\ln\frac{1}{\delta}\right)\cdot\tilde{\mathcal{O}}\left(|X|^2|A|+|X|\ln\frac{1}{\delta}\right)\qquad\text{(Lemma C.3)}$$

$$\leq\tilde{\mathcal{O}}\left(|X|^4|A|^2\ln\frac{1}{\delta}+|X|^3|A|\ln\frac{1}{\delta}\right).$$

The third term can be bounded in the exact same way. Therefore, we arrive at

$$|X|B_2\leq\tilde{\mathcal{O}}\left(|X|^5|A|^2\ln(1/\delta)+|X|^4|A|\ln^2(1/\delta)\right).\qquad(29)$$

Next we show that B_1 is bounded by $\tilde{\mathcal{O}}(|X|\sqrt{J|A|L_T\ln(1/\delta)}+|X|^3|A|\ln(1/\delta))$. According to the definition of $\epsilon_{i_t}^\star$, we have

$$B_1=\mathcal{O}\left(\sum_{0\leq m<k<J}\sum_{t,z_m}w_t(x_m,a_m)\left(\sum_{x\in X_k,a\in A}w_t(x,a|x_{m+1})\ell_t(x,a)\right)\right.$$

$$\left.\cdot\sqrt{\frac{P(x_{m+1}|x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}}\right)+\mathcal{O}\left(\sum_{0\leq m<k<J}\sum_{t,z_m}\frac{w_t(x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}\right).$$

$$(30)$$

According to Lemma C.3, the second term is bounded as

$$\mathcal{O}\left(\sum_{0\leq m<k<J}\sum_{t,z_m}\frac{w_t(x_m,a_m)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1,N_{i_t}(x_m,a_m)\}}\right)\leq\tilde{\mathcal{O}}\left(J|X|^2|A|+J|X|\ln\frac{1}{\delta}\right).\qquad(31)$$

In the following, we define $\ell_t(k|x,a)\triangleq\sum_{x_k\in X_k,a_k\in A}\ell_t(x_k,a_k)w_t(x_k,a_k|x,a)$ where $w_t(x',a'|x,a)$ is the probability of encountering pair (x',a') given that pair (x,a) was encoun-

tered earlier, under policy π_t and transition P. For the first term of Eq. (30), we then have

$$\sum_{0 \le m < k < J} \sum_{t, z_m} w_t(x_m, a_m) \left(\sum_{x \in X_k, a \in A} w_t(x, a | x_{m+1}) \ell_t(x, a) \right) \cdot \sqrt{\frac{P(x_{m+1} | x_m, a_m) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$$

$$\le \sum_{0 \le m < k < J} \sum_{t, z_m} w_t(x_m, a_m) \sqrt{\left(\sum_{x \in X_k, a \in A} w_t(x, a | x_{m+1}) \ell_t(x, a) \right) \cdot \frac{P(x_{m+1} | x_m, a_m) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$$

$$\le \sum_{0 \le m < k < J} \sum_{t, x_m, a_m} w_t(x_m, a_m) \sqrt{|X_{m+1}| \cdot \frac{\ell_t(k | x_m, a_m) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$$

(Cauchy-Schwarz inequality)

$$\le \sum_{0 \le m < k < J} \sqrt{|X_{m+1}| \ln\left(\frac{T|X||A|}{\delta}\right)}$$

$$\cdot \sum_{t, x_m, a_m} \left(\mathbb{1}_t(x_m, a_m) \sqrt{\frac{\ell_t(k | x_m, a_m)}{\max\{1, N_{i_t}(x_m, a_m)\}}} + \frac{w_t(x_m, a_m) - \mathbb{1}_t(x_m, a_m)}{\sqrt{\max\{1, N_{i_t}(x_m, a_m)\}}} \right). \qquad (32)$$

According to Lemma C.3 again, we have for all $m = 0, 1, \ldots, J-1$,

$$\sum_{t=1}^{T} \sum_{x_m, a_m} \frac{w_t(x_m, a_m) - \mathbb{1}_t(x_m, a_m)}{\sqrt{\max\{1, N_{i_t}(x_m, a_m)\}}} \le \tilde{\mathcal{O}}\left(|X_m||A| + \ln(1/\delta)\right).$$

For the term $\sum_{t, x_m, a_m} \mathbb{1}_t(x_m, a_m) \sqrt{\frac{\ell_t(k | x_m, a_m)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$, using Cauchy-Schwarz inequality, we have

$$\sum_{t, x_m, a_m} \mathbb{1}_t(x_m, a_m) \sqrt{\frac{\ell_t(k | x_m, a_m)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$$

$$\le \sum_{x_m, a_m} \sqrt{\sum_{t=1}^{T} \frac{\mathbb{1}_t(x_m, a_m)}{\max\{1, N_{i_t}(x_m, a_m)\}}} \cdot \sqrt{\sum_{t=1}^{T} \mathbb{1}_t(x_m, a_m) \ell_t(k | x_m, a_m)}$$

(Cauchy-Schwarz inequality)

$$\le \mathcal{O}\left(\sqrt{|X_m||A| \left(\sum_{t=1}^{T} \sum_{x \in X_m, a \in A} \mathbb{1}_t(x_m, a_m) \ell_t(k | x, a) \right) \cdot \ln T} \right), \qquad (33)$$

where the last step uses Cauchy-Schwarz inequality again and the fact $\sum_{t=1}^{T} \frac{\mathbb{1}_t(x_m, a_m)}{\max\{1, N_{i_t}(x_m, a_m)\}} \le \mathcal{O}(\ln T)$. Combining Eq. (32) and Eq. (33), we have

$$\sum_{0 \le m < k < J} \sum_{t, z_m} w_t(x_m, a_m) \left(\sum_{x \in X_k, a \in A} w_t(x, a | x_{m+1}) \ell_t(x, a) \right) \cdot \sqrt{\frac{P(x_{m+1} | x_m, a_m) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_{i_t}(x_m, a_m)\}}}$$

$$\le \tilde{\mathcal{O}}\left(\sum_{0 \le m < k < J} \sqrt{|X_m||A||X_{m+1}| \ln \frac{1}{\delta}} \sqrt{\sum_{t=1}^{T} \sum_{x \in X_m, a \in A} \mathbb{1}_t(x_m, a_m) \ell_t(k | x, a)} \right)$$

$$\le \tilde{\mathcal{O}}\left(\sum_{m=0}^{J-1} \sqrt{J|X_m||A||X_{m+1}| \sum_{t=1}^{T} \sum_{k > m} \sum_{x \in X_m, a \in A} \mathbb{1}_t(x_m, a_m) \ell_t(k | x, a) \ln \frac{1}{\delta}} \right).$$

(Cauchy-Schwarz inequality)

Further note that

$$\mathbb{E}_t\left[\sum_{k>m}\sum_{x\in X_m,a\in A}\mathbb{1}_t(x_m,a_m)\ell_t(k|x,a)\right]$$

$$=\sum_{k>m}\sum_{x\in X_m,a\in A}w_t(x,a)\ell_t(k|x,a)$$

$$=\sum_{x\in X_m,a\in A}w_t(x,a)\sum_{k>m}\sum_{x'\in X_k,a'\in A}w_t(x',a'|x,a)\ell_t(x',a')$$

$$=\sum_{k>m}\sum_{x'\in X_k,a'\in A}w_t(x',a')\ell_t(x',a')$$

$$\leq\langle w_t,\ell_t\rangle$$

and

$$\mathbb{E}_t\left[\left(\sum_{k>m}\sum_{x\in X_m,a\in A}\mathbb{1}_t(x_m,a_m)\ell_t(k|x,a)\right)^2\right]\leq J\langle w_t,\ell_t\rangle.$$

Using Freedman inequality J times with parameter δ/J for $m=0,1,\dots,J-1$ and taking a union bound, we have with probability $1-\delta$, for all $m=0,1,\dots,J-1$,

$$\sum_{t=1}^T\sum_{k>m}\sum_{x\in X_m,a\in A}\mathbb{1}_t(x_m,a_m)\ell_t(k|x,a)-\sum_{t=1}^T\langle w_t,\ell_t\rangle$$

$$\leq\tilde{\mathcal{O}}\left(\sqrt{J\sum_{t=1}^T\langle w_t,\ell_t\rangle\ln\frac{1}{\delta}}+J\ln\frac{1}{\delta}\right)=\tilde{\mathcal{O}}\left(\sqrt{JL_T\ln\frac{1}{\delta}}+J\ln\frac{1}{\delta}\right).$$

Therefore, using AM-GM inequality, we have

$$\sum_{t=1}^T\sum_{k>m}\sum_{x\in X_m,a\in A}\mathbb{1}_t(x_m,a_m)\ell_t(k|x,a)\leq\tilde{\mathcal{O}}\left(L_T+J\ln\frac{1}{\delta}\right).$$

Combining the results above and Eq. (31), we know that with probability at least $1-\delta$,

$$B_1\leq\tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T\ln\frac{1}{\delta}}+J|X|\sqrt{|A|}\ln\frac{1}{\delta}\right)+\tilde{\mathcal{O}}\left(J|X|^2|A|+J|X|\ln\frac{1}{\delta}\right)$$

$$\leq\tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T\ln\frac{1}{\delta}}+|X|^3|A|\ln\frac{1}{\delta}\right). \tag{34}$$

Finally, combining Eq. (29) and Eq. (34) and considering the probability of the events of Lemma C.2 and Lemma C.3, we have with probability $1-6\delta$,

$$B_1+|X|B_2\leq\tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T\ln\frac{1}{\delta}}+|X|^5|A|^2\ln\frac{1}{\delta}+|X|^4|A|\ln^2\frac{1}{\delta}\right),$$

finishing the proof. □

C.3.2 Bounding ERROR

Lemma C.7. *With probability at least $1-6\delta$, we have*

$$\text{ERROR}=\sum_{t=1}^T\langle w_t-\widehat{w}_t,\ell_t\rangle=\tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T\ln\frac{1}{\delta}}+|X|^5|A|^2\ln\frac{1}{\delta}+|X|^4|A|\ln^2\frac{1}{\delta}\right),$$

Proof. Note that according to the definition of $\widehat{w}_t$, the transition function $P^{\widehat{w}_t}$ induced by $\widehat{w}_t$ is in $\mathcal{P}_{i_t}$. Therefore, applying Lemma C.6, we know that with probability at least $1 - 6\delta$,

$$
\begin{aligned}
\text{ERROR} &= \sum_{t=1}^{T} \langle \widehat{w}_t - w_t, \ell_t \rangle \\
&\leq \sum_{t=1}^{T} \sum_{x \in X, a \in A} |\widehat{w}_t(x, a) - w_t(x, a)| \, \ell_t(x, a) \\
&\leq \tilde{\mathcal{O}} \left(|X| \sqrt{J |A| L_T \ln \frac{1}{\delta}} + |X|^5 |A|^2 \ln \frac{1}{\delta} + |X|^4 |A| \ln^2 \frac{1}{\delta} \right),
\end{aligned}
$$

completing the proof. □

C.3.3 Bounding BIAS-1

Lemma C.8. *With probability at least $1 - 7\delta$, we have*

$$
\text{BIAS-1} = \sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t \right\rangle \leq \tilde{\mathcal{O}} \left(|X| \sqrt{J |A| L_T \ln \frac{1}{\delta}} + |X|^5 |A|^2 \ln \frac{1}{\delta} + |X|^4 |A| \ln^2 \frac{1}{\delta} \right).
$$

Proof. First we write

$$
\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t \right\rangle = \sum_{t=1}^{T} \left\langle \widehat{w}_t, \mathbb{E}_t \left[\widehat{\ell}_t \right] - \widehat{\ell}_t \right\rangle + \sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \mathbb{E}_t \left[\widehat{\ell}_t \right] \right\rangle.
$$

Since $\widehat{w}_t(x, a) \leq \phi_t(x, a)$ by the definition of ϕ_t, we have

$$
\left\langle \widehat{w}_t, \widehat{\ell}_t \right\rangle \leq \sum_{k=1}^{J} \sum_{x \in X_k, a \in A} \frac{\widehat{w}_t(x, a)}{\phi_t(x, a)} \cdot \mathbb{1}_t(x, a) \leq J,
$$

$$
\mathbb{E}_t \left[\left\langle \widehat{w}_t, \widehat{\ell}_t \right\rangle^2 \right] \leq \mathbb{E}_t \left[J \cdot \left\langle \widehat{w}_t, \widehat{\ell}_t \right\rangle \right] = J \sum_{x,a} \widehat{w}_t(x, a) \cdot \frac{\ell_t(x, a)}{\phi_t(x, a)} \cdot w_t(x, a) \leq J \cdot \langle w_t, \ell_t \rangle,
$$

and thus according to Freedman inequality, we have with probability at least $1 - \delta$,

$$
\sum_{t=1}^{T} \left\langle \widehat{w}_t, \mathbb{E}_t \left[\widehat{\ell}_t \right] - \widehat{\ell}_t \right\rangle \leq \mathcal{O} \left(\sqrt{J \sum_{t=1}^{T} \langle w_t, \ell_t \rangle \ln \frac{1}{\delta}} + J \cdot \ln \frac{1}{\delta} \right) = \mathcal{O} \left(\sqrt{J L_T \ln \frac{1}{\delta}} + |X| \ln \frac{1}{\delta} \right).
\tag{35}
$$

For the second term, we have

$$
\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \mathbb{E}_t \left[\widehat{\ell}_t \right] \right\rangle = \sum_{t,x,a} \widehat{w}_t(x, a) \ell_t(x, a) \cdot \left(1 - \frac{w_t(x, a)}{\phi_t(x, a)} \right) \leq \sum_{t,x,a} |\phi_t(x, a) - w_t(x, a)| \cdot \ell_t(x, a).
$$

By the definition of ϕ_t, one has $\phi_t = w^{P_t^x, \pi_t}$ for $P_t^x = \operatorname{argmax}_{\hat{P} \in \mathcal{P}_{i_t}} \sum_a w^{\hat{P}, \pi_t}(x, a)$. Therefore, according to Lemma C.7, we have with probability at least $1 - 6\delta$,

$$
\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \mathbb{E}_t \left[\widehat{\ell}_t \right] \right\rangle \leq \tilde{\mathcal{O}} \left(|X| \sqrt{J |A| L_T \ln \frac{1}{\delta}} + |X|^5 |A|^2 \ln \frac{1}{\delta} + |X|^4 |A| \ln^2 \frac{1}{\delta} \right).
\tag{36}
$$

Combining Eq. (35) and Eq. (36) proves the lemma.

□

C.3.4 Bounding BIAS-2

Lemma C.9. *With probability at least* $1 - 5\delta$, *we have*

$$\text{BIAS-2} = \sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle \leq C \sum_{x \in X, a \in A} u(x,a) \sqrt{8\rho_T(x,a) \sum_{t=1}^{T} \ell_t(x,a) \ln \frac{C|X||A|}{\delta}}$$
$$+ 2C \left\langle u, \rho_T \right\rangle \ln \frac{C|X||A|}{\delta},$$

for some constant $C = \tilde{\mathcal{O}}(1)$.

Proof. First we write

$$\sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \ell_t \right\rangle = \sum_{t=1}^{T} \left\langle u, \mathbb{E}_t \left[\widehat{\ell}_t \right] - \ell_t \right\rangle + \sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \mathbb{E}_t \left[\widehat{\ell}_t \right] \right\rangle.$$

The first term is nonpositive under the event of [Lemma C.2](#) as for any $(x,a) \in X \times A$, $w_t(x,a) \leq \phi_t(x,a)$ by the definition of ϕ_t and thus

$$\mathbb{E}_t \left[\widehat{\ell}_t(x,a) \right] - \ell_t(x,a) = w_t(x,a) \cdot \frac{\ell_t(x,a)}{\phi_t(x,a)} - \ell_t(x,a) \leq 0. \tag{37}$$

For the second term, note that for each $(x,a) \in X \times A$, we have

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{\phi_t(x,a)} \cdot \mathbb{1}_t(x,a) \leq T^3 |X|^2 |A|, \qquad \text{([Lemma C.5](#))}$$

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{\phi_t(x,a)} \cdot \mathbb{1}_t(x,a) \leq \rho_t(x,a),$$

and

$$\sum_{t=1}^{T} \mathbb{E}_t \left[\widehat{\ell}_t(x,a)^2 \right] \leq \sum_{t=1}^{T} \mathbb{E}_t \left[\frac{\ell_t(x,a)}{\phi_t(x,a)^2} \cdot \mathbb{1}_t(x,a) \right] \leq \rho_T(x,a) \sum_{t=1}^{T} \ell_t(x,a).$$

Therefore, using [Theorem 2.2](#) with $X_t = \widehat{\ell}_t(x,a) - \mathbb{E}_t \left[\widehat{\ell}_t(x,a) \right]$, $B_t = \rho_t(x,a)$, $B^\star = \rho_T(x,a)$, $b = T^3 |X|^2 |A|$, $C = \lceil \log_2 b \rceil \lceil \log_2 b^2 T \rceil = \tilde{\mathcal{O}}(1)$, we have with probability at least $1 - \frac{\delta}{|X||A|}$,

$$\sum_{t=1}^{T} \widehat{\ell}_t(x,a) - \mathbb{E}_t \left[\widehat{\ell}_t(x,a) \right] \leq C \left(\sqrt{8\rho_T(x,a) \sum_{t=1}^{T} \ell_t(x,a) \ln \frac{C|X||A|}{\delta}} + 2\rho_T(x,a) \ln \frac{C|X||A|}{\delta} \right).$$

Taking a union bound over all $(x,a) \in X \times A$, multiplying both sides by $u(x,a)$, and summing up all these inequalities, we have with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \left\langle u, \widehat{\ell}_t - \mathbb{E}_t \left[\widehat{\ell}_t \right] \right\rangle$$

$$\leq C \sum_{x \in X, a \in A} u(x,a) \left(\sqrt{8\rho_T(x,a) \sum_{t=1}^{T} \ell_t(x,a) \ln \frac{C|X||A|}{\delta}} + 2\rho_T(x,a) \ln \frac{C|X||A|}{\delta} \right). \tag{38}$$

Combining [Eq. (37)](#) and [Eq. (38)](#) finishes the proof.

$\square$

C.3.5 Bounding REG-TERM

Lemma C.10. *With probability at least $1 - 4\delta$, we have*

$$\text{REG-TERM} = \sum_{t=1}^{T} \left\langle \widehat{w}_t - u, \widehat{\ell}_t \right\rangle \leq \tilde{\mathcal{O}}\left(\frac{|X|^2|A|}{\eta}\right) + 5\eta\overline{L}_T - \frac{\langle u, \rho_T \rangle}{70\eta \ln T},$$

where $\overline{L}_T = \sum_{t=1}^{T} \sum_{x \in X, a \in A} \mathbb{1}_t(x,a)\ell_t(x,a)$.

Proof. We condition on the event of Lemma C.2. First, we prove that $u \in \Delta(\mathcal{P}_i) \cap \Omega$ for all i (recall its definition in Eq. (24)). Indeed, for any fixed $(x, a, x') \in X_k \times A \times X_{k+1}$, $k = 0, 1, \ldots, J-1$, we have (with $w^{P_0, \pi_a}(x)$ being the probability of visiting x under P_0 and π_a)

$$u(x, a, x') \geq \frac{1}{T|A|} w^{P_0, \pi_a}(x, a, x')$$

$$= \frac{1}{T|A|} w^{P_0, \pi_a}(x) P_0(x'|x, a)$$

$$\geq \frac{1}{T|A|} \left(\sum_{x'' \in X_{k(x)-1}} w^{P_0, \pi_a}(x'') \cdot P_0(x|x'', a) \right) \cdot \frac{1}{T|X|} \qquad \text{(Lemma C.4)}$$

$$\geq \frac{1}{T^3|X|^2|A|} \left(\sum_{x'' \in X_{k(x)-1}} w^{P_0, \pi_a}(x'') \right), \qquad \text{(Lemma C.4 again)}$$

$$= \frac{1}{T^3|X|^2|A|},$$

which shows $u \in \Omega$. On the other hand, since $P \in \mathcal{P}_i$ under Lemma C.2 and $P_0 \in \mathcal{P}_i$ as well by Lemma C.4, we have $u^\star \in \Delta(\mathcal{P}_i)$ and $w^{P_0, \pi_a} \in \Delta(\mathcal{P}_i)$, which indicates that, as a convex combination of $u^\star$ and w^{P_0, π_a} for all a, u has to be in $\Delta(\mathcal{P}_i)$ as well.

Therefore, by standard OMD analysis (e.g., [5, Lemma 12]), we have

$$\left\langle \widehat{w}_t - u, \widehat{\ell}_t \right\rangle$$

$$\leq D_{\psi_t}(u, \widehat{w}_t) - D_{\psi_t}(u, \widehat{w}_{t+1}) + \sum_{k=0}^{J-1} \sum_{(x,a,x') \in X_k \times A \times X_{k+1}} \eta_t(x,a) \widehat{w}_t^2(x,a,x') \widehat{\ell}_t^2(x,a)$$

$$\leq D_{\psi_t}(u, \widehat{w}_t) - D_{\psi_t}(u, \widehat{w}_{t+1}) + \sum_{k=0}^{J-1} \sum_{(x,a) \in X_k \times A} \eta_t(x,a) \widehat{w}_t^2(x,a) \widehat{\ell}_t^2(x,a)$$

$$\left(\sum_{x' \in X_{k+1}} \widehat{w}_t(x,a,x')^2 \leq \widehat{w}_t(x,a)^2 \right)$$

$$\leq D_{\psi_t}(u, \widehat{w}_t) - D_{\psi_t}(u, \widehat{w}_{t+1}) + \sum_{x \in X, a \in A} \eta_t(x,a) \mathbb{1}_t(x,a) \ell_t(x,a). \qquad (\widehat{w}_t(x,a) \leq \phi_t(x,a))$$

Summing over t gives

$$\sum_{t=1}^{T} \left\langle \widehat{w}_t - u, \widehat{\ell}_t \right\rangle \leq D_{\psi_1}(u, \widehat{w}_1) + \sum_{t=1}^{T-1} D_{\psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\psi_t}(u, \widehat{w}_{t+1})$$

$$+ \sum_{t=1}^{T} \sum_{x \in X, a \in A} \eta_t(x,a) \mathbb{1}_t(x,a) \ell_t(x,a). \qquad (39)$$

Next, for a fixed (x, a) pair, let $n(x, a)$ be the total number of times the learning rate for (x, a) has increased, such that $\eta_T(x, a) = \eta\kappa^{n(x,a)}$, and let $t_1, \ldots, t_{n(x,a)}$ be the rounds where $\eta_t(x, a)$ is increased, such that $\eta_{t_i+1}(x, a) = \eta_{t_i}(x, a)\kappa$. Then since $\frac{1}{\phi_{t_{n(x,a)}+1}(x,a)} > \rho_{t_{n(x,a)}}(x, a) >$

$2\rho_{t_{n(x,a)}-1}(x,a) > \cdots > 2^{n(x,a)-1}\rho_1(x,a) > 2^{n(x,a)}|A|$ and $\frac{1}{\phi_{t_{n(x,a)}+1}(x,a)} \leq T^3|X|^2|A|$
(Lemma C.5), we have $n \leq \log_2\left(T^3|X|^2\right) \leq 7\log_2 T$.

Therefore, we have $\eta_t(x,a) \leq \eta e^{\frac{7\log_2 T}{7\ln T}} \leq 5\eta$ for any t, $x \in X$, and $a \in A$, and the last term in Eq. (39) is thus bounded by $5\eta\overline{L}_T$. For the second term, with $h(y) = y - 1 - \ln y$, we have

$$\sum_{t=1}^{T-1} D_{\psi_{t+1}}(u, \widehat{w}_{t+1}) - D_{\psi_t}(u, \widehat{w}_{t+1})$$

$$\leq \sum_{t=1}^{T-1}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} \left(\frac{1}{\eta_{t+1}(x,a)} - \frac{1}{\eta_t(x,a)}\right) h\left(\frac{u(x,a,x')}{\widehat{w}_{t+1}(x,a,x')}\right)$$

$$\leq \sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} \frac{1-\kappa}{\eta\cdot\kappa^{n(x,a)}} \cdot h\left(\frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')}\right)$$

$$\leq -\frac{1}{35\eta\ln T}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} h\left(\frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')}\right)$$

$$\qquad\qquad\qquad\qquad\qquad (1-\kappa \leq -\tfrac{1}{7\ln T} \text{ and } \kappa^{n(x,a)} \leq e^{\frac{7\log_2 T}{7\ln T}} \leq 5)$$

$$= -\frac{1}{35\eta\ln T}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} \left(\frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')} - 1 - \ln\frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')}\right)$$

$$\leq \frac{|X|^2|A|(1+6\ln T)}{35\eta\ln T} - \frac{1}{35\eta\ln T}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} \frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')}$$

$$\qquad\qquad\qquad\qquad\qquad (\ln\frac{u(x,a,x')}{\widehat{w}_{t_{n(x,a)}+1}(x,a,x')} \leq 6\ln T)$$

$$\leq \frac{|X|^2|A|}{5\eta} - \frac{1}{35\eta\ln T}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A, x'\in X_{k+1}} \frac{u(x,a,x')}{\phi_{t_{n(x,a)}+1}(x,a)}$$

$$= \frac{|X|^2|A|}{5\eta} - \frac{1}{35\eta\ln T}\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A} \frac{u(x,a)}{\phi_{t_{n(x,a)}+1}(x,a)}$$

$$= \frac{|X|^2|A|}{5\eta} - \frac{\langle u, \rho_T\rangle}{70\eta\ln T}. \qquad\qquad (\rho_T(x,a) = \tfrac{2}{\phi_{t_{n(x,a)}+1}(x,a)})$$

Finally, we bound the first term in Eq. (39):

$$D_{\psi_1}(u, \widehat{w}_1) = \frac{1}{\eta}\left(\sum_{k=0}^{J-1}\sum_{(x,a,x')\in X_k\times A\times X_{k+1}} h\left(\frac{u(x,a,x')}{\widehat{w}_1(x,a,x')}\right)\right)$$

$$= \frac{1}{\eta}\left(\sum_{k=0}^{J-1}\sum_{(x,a,x')\in X_k\times A\times X_{k+1}} h\left(|X_k|\cdot|A|\cdot|X_{k+1}|\cdot u(x,a,x')\right)\right)$$

$$= \frac{1}{\eta}\left(\sum_{k=0}^{J-1}\sum_{(x,a,x')\in X_k\times A\times X_{k+1}} \ln\left(\frac{1}{|X_k|\cdot|A|\cdot|X_{k+1}|\cdot u(x,a,x')}\right)\right)$$

$$\leq \tilde{\mathcal{O}}\left(\frac{|X|^2|A|}{\eta}\right).$$

Combining all the bounds finishes the proof. □

C.3.6   Putting everything together

Now we are ready to prove Theorem 4.1. For completeness, we restate the theorem below.

Theorem C.11. *Algorithm 4 with a suitable choice of η ensures that with probability at least $1 - \delta$,*
$\text{Reg} = \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L^\star \ln \frac{1}{\delta}} + |X|^5|A|^2 \ln^2 \frac{1}{\delta}\right).$

Proof. First, note that

$$\mathbb{E}_t\left[\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A} \mathbb{1}_t(x,a)\cdot \ell_t(x,a)\right] = \langle w_t, \ell_t\rangle \leq J,$$

$$\mathbb{E}_t\left[\left(\sum_{k=0}^{J-1}\sum_{x\in X_k, a\in A} \mathbb{1}_t(x,a)\cdot \ell_t(x,a)\right)^2\right] \leq J\cdot \langle w_t, \ell_t\rangle.$$

Therefore, using Freedman's inequality, we have with probability at least $1 - \delta$

$$\overline{L}_T - L_T \leq 2\sqrt{JL_T \ln\frac{1}{\delta}} + J\ln\frac{1}{\delta},$$

where $\overline{L}_T$ is defined in Lemma C.10. Furthermore, using AM-GM inequality, we have with probability at least $1 - \delta$,

$$\overline{L}_T \leq 2L_T + 2J\ln\frac{1}{\delta}. \tag{40}$$

Choosing $\eta \leq \frac{1}{280C\ln(C|X||A|/\delta)\ln T}$, combining Lemma C.7, Lemma C.8, Lemma C.9 and Lemma C.10 and letting $L_u \triangleq \sum_{t=1}^T \langle u, \ell_t\rangle$, we have with probability at least $1 - 22\delta$:

$L_T - L^\star$

$$\leq \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T \ln\frac{1}{\delta}} + \frac{|X|^2|A|}{\eta}\right) + 5\eta\overline{L}_T + \underbrace{\left(2C\langle u, \rho_T\rangle \ln\frac{C|X||A|}{\delta} - \frac{\langle u, \rho_T\rangle}{140\eta\ln T}\right)}_{\text{TERM}1}$$

$$+ \sum_{x\in X, a\in A} u(x,a)\underbrace{\left(C\sqrt{8\rho_T(x,a)\sum_{t=1}^T \ell_t(x,a)\ln\frac{C|X||A|}{\delta}} - \frac{\rho_T(x,a)}{140\eta\ln T}\right)}_{\text{TERM}2}$$

$$+ \tilde{\mathcal{O}}\left(|X|^5|A|^2\ln\frac{1}{\delta} + |X|^4|A|\ln^2\frac{1}{\delta}\right)$$

$$\leq \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T \ln\frac{1}{\delta}} + \frac{|X|^2|A|}{\eta} + \eta L_u\ln\frac{1}{\delta} + |X|^5|A|^2\ln\frac{1}{\delta} + |X|^4|A|\ln^2\frac{1}{\delta}\right) + 10\eta L_T$$

(TERM1 is nonpositive, AM-GM inequality for TERM2, and Eq. (40))

$$\leq \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T \ln\frac{1}{\delta}} + \frac{|X|^2|A|}{\eta} + \eta L^\star\ln\frac{1}{\delta} + |X|^5|A|^2\ln\frac{1}{\delta} + |X|^4|A|\ln^2\frac{1}{\delta}\right) + 10\eta L_T.$$

(Eq. (25))

As $\eta \leq \frac{1}{280C\ln(C|X||A|/\delta)\ln T} \leq \frac{1}{20}$, rearranging the terms gives

$$L_T - L^\star \leq \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L_T \ln\frac{1}{\delta}} + \frac{|X|^2|A|}{\eta} + \eta L^\star\ln\frac{1}{\delta} + |X|^5|A|^2\ln\frac{1}{\delta} + |X|^4|A|\ln^2\frac{1}{\delta}\right).$$

Finally, choosing $\eta = \min\left\{\sqrt{\frac{|X|^2|A|}{L^\star\ln\frac{1}{\delta}}}, \frac{1}{280C\ln(C|X||A|/\delta)\ln T}\right\}$, $\delta = \delta'/22$, and solving the quadratic inequality, we have with probability at least $1 - \delta'$,

$$L_T - L^\star \leq \tilde{\mathcal{O}}\left(|X|\sqrt{J|A|L^\star\ln\frac{1}{\delta'}} + |X|^5|A|^2\ln\frac{1}{\delta'} + |X|^4|A|\ln^2\frac{1}{\delta'}\right),$$

finishing the proof. $\qquad\square$

Remark 3. Similarly to the MAB case, the proof above requires tuning the initial learning rate η in terms of the unknown quantity $L^\star$, and again, using standard doubling trick can remove this restriction, as pointed out in Remark 1.

C.4 Issues of other potential approaches

In this section, we discuss why the idea of clipping [7] or implicit exploration [31] may not be directly applicable to achieve near-optimal high-probability small-loss bounds.

Implicit exploration. First, we consider the idea of implicit exploration. As mentioned in Appendix C.2, this means using the following loss estimator: $\widehat{\ell}_t = \frac{\ell_t(x,a)}{\phi_t(x,a)+\gamma} \cdot \mathbb{1}_t\{x,a\}$ for all $x \in X$ and $a \in A$ and some parameter $\gamma > 0$, and without using our increasing learning schedule. The concentration results of [25, Lemma 12] show that the deviation contains a term of order $1/\gamma$, meaning that γ cannot be too small.

Repeating the same analysis, one can see that the main difficulty of obtaining high-probability small-loss bounds in this case is to bound BIAS-2 by the loss of the algorithm $L_T = \sum_{t=1}^{T} \langle w_t, \ell_t \rangle$ or $L^\star$, instead of the number of episodes T. Indeed, consider the term $\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \mathbb{E}_t\left[\widehat{\ell}_t\right] \right\rangle$:

$$
\sum_{t=1}^{T} \left\langle \widehat{w}_t, \ell_t - \mathbb{E}_t\left[\widehat{\ell}_t\right] \right\rangle = \sum_{t=1}^{T} \sum_{x \in X, a \in A} \widehat{w}_t(x,a)\ell_t(x,a) \cdot \left(1 - \frac{w_t(x,a)}{\phi_t(x,a)+\gamma}\right)
$$

$$
\leq \sum_{t=1}^{T} \sum_{x \in X, a \in A} |\phi_t(x,a) - w_t(x,a)|\ell_t(x,a) + \frac{\gamma}{\gamma + \phi_t(x,a)} \cdot \widehat{w}_t(x,a)\ell_t(x,a).
$$

The first term can still be bounded by $\mathcal{O}\left(|X|\sqrt{J|A|L_T} + |X|^5|A|^2 \ln\frac{1}{\delta} + |X|^4|A| \ln^2\frac{1}{\delta}\right)$ according to Lemma C.8. For the second term, while it is at most $\gamma \sum_{t=1}^{T} \sum_{x,a} \ell_t(x,a) \leq \gamma|X||A|T$, it is not clear at all how to bound it in terms of L_T or $L^\star$. For MAB (where there is only one state x_0), it is possible to show that $\sum_{t=1}^{T} \ell_t(x_0,a) \leq \sum_{t=1}^{T} \ell_t(x_0,a^\star) + \tilde{\mathcal{O}}(\frac{1}{\eta} + \frac{1}{\gamma})$ for all $a \neq a^\star$ where $a^\star$ is the best action, making it possible to connect the second term with $L^\star$. However, we do not see a way of doing similar analysis for general MDPs.

Clipping. On the other hand, the idea of clipping for MAB is to clip all small probabilities so that actions with probability smaller than γ are never selected. Even from an algorithmic perspective, it is not clear how to generalize this idea to MDPs, because it is possible that for a state x, $\widehat{w}_t(x,a)$ is smaller than γ for *all* a. In this case, the clipping idea suggests not to "pick" (x,a) at all for any a, but there is no way to ensure that if the transition function is such that x can always be visited with some positive probability regardless of the policy we execute.

Moreover, even if there is a way to fix this, the analysis of clipping for MAB is also similar to the idea of implicit exploration in terms of obtaining small-loss bounds of order $\tilde{\mathcal{O}}(\sqrt{L^\star})$, and as we argued already, even for implicit exploration there are difficulties in generalizing the analysis to MDPs.