Online Markov Decision Processes with Aggregate Bandit Feedback

Alon Cohen ALONCOHEN@GOOGLE.COM
Google Research, Tel Aviv

Haim Kaplan HAIMK@POST.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University and Google Research, Tel Aviv

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University and Google Research, Tel Aviv

Yishay Mansour MANSOUR@TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University and Google Research, Tel Aviv

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study a novel variant of online finite-horizon Markov Decision Processes with adversarially changing loss functions and initially unknown dynamics. In each episode, the learner suffers the loss accumulated along the trajectory realized by the policy chosen for the episode, and observes *aggregate bandit feedback*: the trajectory is revealed along with the cumulative loss suffered, rather than the individual losses encountered along the trajectory. Our main result is a computationally efficient algorithm with $O(\sqrt{K})$ regret for this setting, where K is the number of episodes.

We establish this result via an efficient reduction to a novel bandit learning setting we call Distorted Linear Bandits (DLB), which is a variant of bandit linear optimization where actions chosen by the learner are adversarially distorted before they are committed. We then develop a computationally-efficient online algorithm for DLB for which we prove an $O(\sqrt{T})$ regret bound, where T is the number of time steps. Our algorithm is based on online mirror descent with a self-concordant barrier regularization that employs a novel increasing learning rate schedule.

1. Introduction

Markov Decision Processes are a ubiquitous model for decision making that captures a wide array of applications including autonomous road navigation, robotics, gaming and many more. In the finite-horizon version of the model, the goal of the agent is to minimize her expected total loss over a fixed number of time steps. Classic results in finite-horizon MDPs state that the optimal policy of the agent is deterministic; namely, a mapping between each state and time step to an action for the agent to play.

In this paper, we study the problem of *Online MDPs with Aggregate Feedback* which is played for K episodes. The dynamics of the MDP are fixed but unknown to the learner. After each episode, in addition to observing her trajectory within the MDP, the agent also gets to view her total loss along this trajectory. The agent, however, does not get to observe the individual losses of specific states and actions that comprise the trajectory. This setting was recently considered in Efroni et al. (2020) where the authors derived computationally-efficient learning algorithms for the case where the losses are sampled i.i.d. from some unknown distribution. In this work, we assume that the losses are non-stochastic and may be chosen by an adversary—a significantly more challenging task. This, for example, can be seen as a relaxation of the Markov assumption, where some complex part of the environment that is hard to model influences only the learner's losses.

The adversarial setting is a variant of online MDPs (Even-Dar et al., 2009) with initially unknown model dynamics, previously considered either when full information about the losses is received (Neu et al., 2010), or with traditional bandit feedback where the agent sees the individual losses of all states and actions that were visited along each of her generated trajectories (Rosenberg and Mansour, 2019). Commonly, the main solution technique is to separate the K episodes into $O(\log K)$ epochs; in each epoch, the agent runs a no-regret algorithm using an estimate of the dynamics obtained from observations accrued up to the beginning of the epoch. To tackle bandit feedback in general, it is common practice to employ a full-information learning algorithm which is fed with an unbiased estimate of the losses in each episode. Nevertheless, in our setup we do not know the MDP dynamics, so it is hopeless for the learner to generate such an unbiased estimate since it is impossible to calculate the probability of visiting each state and action without exact knowledge of the transition distributions. This impediment was overcome in Jin et al. (2020) that followed the "optimism in the face of uncertainty" principle: they fed the learning algorithm with a certain underestimate of the loss. This drives the agent to explore under-sampled state-action pairs, helps to obtain better estimates of the dynamics, and reduces the overall bias of the loss estimators over time.

We utilize a similar approach to tackle the aggregate feedback by reducing the problem to $O(\log K)$ epochs in each of which we solve a variant of linear bandits over our current estimate of the model dynamics. We name the learning problem in each epoch *Distorted Linear Bandits* (DLB). This is a variant of the linear bandits problem in which, after choosing an action, it can be distorted (i.e., perturbed) in an adversarial manner before it is played. This distortion unavoidably introduces a non-negligible bias when trying to generate an estimate of the loss vector. The DLB problem is also interesting in its own right, capturing scenarios where there is uncertainty regarding the action that is actually taken, which might deviate significantly from the action intended to be taken—a phenomenon that occurs in applications in robotics and control, where the actions are continuous in nature.

We derive two learning algorithms for the DLB setting that yield a $O(\sqrt{K})$ regret bound, yet mitigate the estimation bias in different ways. Our first algorithm, based on EXP2 (Awerbuch and Kleinberg, 2004; McMahan and Blum, 2004), utilizes an optimistic approach by feeding the algorithm with underestimates of the loss similarly to Bartlett et al. (2008). This technique, it turns out, is not computationally efficient due to the non-convex nature of these underestimates. Our second algorithm, however, runs in polynomial-time per episode. It is a variant of Online Mirror Descent that uses a self-concordant barrier function as a regularizer (Abernethy et al., 2009) with a series of increasing learning rates. The idea of using increasing learning rates to alleviate estimation bias is used in various recent works (Bubeck et al., 2017; Agarwal et al., 2017; Lee et al., 2020). Intuitively, it gives the learner a "boost" towards playing better actions whenever the estimation bias is large.

1.1. Summary of Contributions

The main contributions of the paper are as follows:

- We introduce the setting of Online MDPs with Aggregate Bandit Feedback, where the dynamics are initially unknown and costs may be chosen by an adversary; to the best of our knowledge, such a problem has not been studied before.

- We establish an efficient reduction from Online MDPs with Aggregate Bandit Feedback to the novel Distorted Linear Bandits (DLB) problem.
- We give a computationally-efficient online learning algorithm for the DLB problem with $O(\sqrt{T})$ regret over T rounds.
- Combining the two techniques, we obtain a computationally-efficient online learning method for Online MDPs with Aggregate Bandit Feedback with $O(\sqrt{K})$ regret over K episodes.

In Section 3 we present the Online MDP with Aggregate Feedback model explicitly and give our main result. We also present the Distorted Linear Bandits (DLB) setting, the reduction between the two models, and prove our regret bound for online MDPs. In Section 4 we give our two algorithms for the DLB setting and analyze their regrets.

1.2. Additional Related Work

The study of regret minimization in reinforcement learning dates back to Jaksch et al. (2010) who considered an MDP with unknown dynamics and losses, but where the losses are sampled i.i.d. This model was further studied in Azar et al. (2017); Zanette and Brunskill (2019) that provided improved bounds.

Online MDPs were introduced in Even-Dar et al. (2009) who studied MDPs with known dynamics and adversarially changing losses. Later Neu et al. (2013) extended the online MDP to handle bandit feedback. Abbasi Yadkori et al. (2013) considered MDPs where both the dynamics and the losses change adversarially. Their algorithm, however, is not computationally-efficient as they show in a hardness result. All the above results assume access to individual losses while in this work we assume the learner observes only the aggregate loss of an episode.

Bandit linear optimization has been extensively studied under both semi-bandit and bandit feedback; for an extensive survey of this literature, see Slivkins et al. (2019); Lattimore and Szepesvári (2020). Misspecified linear bandits were introduced in Ghosh et al. (2017) where the loss of each action can be perturbed arbitrarily. They give an impossibility result for large sparse deviations and a regret bound for small deviations (see also Lattimore et al., 2020). Our model differs from misspecified linear bandit, most importantly, in that we allow for adversarial losses. In addition, we also differ both in the fact that the deviations might be large (and we can only globally bound them) and the fact that the loss is linear but with respect to a distorted action.

2. Preliminaries

Finite-Horizon MDPs. A finite-horizon Markov Decision Process is a tuple (S, A, s_1, P, ℓ, H) where S is a finite set of states; A is a finite set of actions; $s_1 \in S$ is the start state; the integer H defines the horizon. The transition function P defines a probability distribution $P(s' \mid s, a, h)$ of the next state s' given the current state s, action a, and time $h \in [H]$. The loss function is ℓ defines a loss $\ell(h, s, a, s') \in [0, 1]$ for every time $h \in [H]$ state s, action a, and next state s'.

A (randomized) policy $\pi : S \times [H] \mapsto \Delta(A)$ maps each state and time to a probability distribution over the actions. A trajectory is a sequence $(s_1, a_1, \ldots, s_H, a_H, s_{H+1})$. The probability of such trajectory with respect to a policy π and a transition function P is $\prod_{h=1}^{H} \pi(a_h \mid s_h, h) P(s_{h+1} \mid s_h, a_h, h)$. The accumulated loss of such a trajectory using a loss function ℓ is $\sum_{h=1}^{H} \ell(h, s_h, a_h, s_{h+1})$.

The expected loss of a policy π with respect to a transition function P and loss function ℓ is

$$L^{\pi,P,\ell} = \mathbb{E}\left[\sum_{h=1}^{H} \ell(h, s_h, a_h, s_{h+1})\right] = \sum_{\substack{(h,s,a,s')\in\\ [H]\times S\times A\times S}} \ell(h,s,a,s')\ \Pr[s_h = s, a_h = a, s_{h+1} = s'].$$

Occupancy Measures. A combination of a policy π and a transition function P provide an occupancy measure $x^{\pi,P}$ such that $x^{\pi,P}(h,s,a,s')$ is the probability, according to P and π, of being at state s at time h, playing action a, and transitioning to state s'. Formally,

$$x^{\pi,P}(h,s,a,s') = \Pr_{\pi,P}[s_h = s, a_h = a, s_{h+1} = s'].$$

Any $x : [H]\times S\times A\times S \mapsto \mathbb{R}$ is an occupancy measure, if and only if

$$\begin{aligned} & x(h,s,a,s') \ge 0, && \forall (h,s,a,s') \in [H]\times S\times A\times S,\\ & \sum_{(s,a,s')\in S\times A\times S} x(h,s,a,s') = 1, && \forall h \in [H],\\ & \sum_{(a,s')\in A\times S} x(h+1,s,a,s') = \sum_{(s',a)\in S\times A} x(h,s',a,s), && \forall (s,h) \in S\times[H-1]. \end{aligned} \tag{1}$$

Indeed, any x that satisfies the conditions above corresponds to an occupancy measure for some policy π and transition function P, both can easily be extracted from x—this correspondence is therefore one-to-one. That is, given an occupancy measure x we can define the corresponding policy and dynamics as follows:

$$\pi^{(x)}(a \mid s,h) = \frac{\sum_{s'\in S} x(h,s,a,s')}{\sum_{(a,s')\in A\times S} x(h,s,a,s')}, \quad \text{and} \quad \widetilde{P}^{(x)}(s' \mid s,a,h) = \frac{x(h,s,a,s')}{\sum_{s'\in S} x(s,a,h,s')}. \tag{2}$$

For more on occupancy measures, see Rosenberg and Mansour (2019).

Self-concordant Barriers and Bregman Divergence. We next briefly review self-concordant barrier functions—a fundamental tool in interior-point methods that was also shown to be highly-useful in linear bandit optimization (Abernethy et al., 2009). Self-concordant barriers are discussed in-depth in Nemirovski (2004); we give the technical definitions in Appendix C and here focus on some useful properties of such functions that we use.

We consider a ϑ-self-concordant barrier function R over a convex set $\mathcal{S}$. In particular, for a self-concordant barrier R, the function $\|h\|_x = \sqrt{h^\top \nabla^2 R(x) h}$ is a norm, and also $\nabla R : \mathrm{int}(\mathcal{S}) \mapsto \mathbb{R}^d$ is invertible. In addition, an important property of the norm $\|\cdot\|_x$ is that for any point $y \in \mathbb{R}^d$ and $x \in \mathrm{int}(\mathcal{S})$,

$$\|y - x\|_x < 1 \implies y \in \mathrm{int}(\mathcal{S}). \tag{3}$$

We define the Bregman divergence with respect to a ϑ-self-concordant barrier R as follows:

$$B_R(y \,\|\, x) = R(y) - R(x) - \nabla R(x)\cdot(y - x).$$

The Bregman divergence is always nonnegative: $B_R(y \,\|\, x) \ge 0$ for any $x, y \in \mathrm{int}(\mathcal{S})$. Moreover, we shall need the following lower bound on the Bregman divergence (see Nemirovski, 2004):

$$B_R(y \,\|\, x) \ge \rho(\|y - x\|_x) \quad \text{for} \quad \rho(z) = z - \log(1+z). \tag{4}$$

We also require the following lemma whose proof is found in Appendix B.

Lemma 1. *Define* $\mathcal{S}_\gamma = \{(1-\gamma)x + \gamma x_1 \mid x \in \mathcal{S}\}$ *for* $x_1 = \arg\min_{x \in \mathcal{S}} R(x)$ *and some* $\gamma \in [0,1]$. *Then* $B_R(y \,\|\, x_1) \le \vartheta \log(1/\gamma)$ *for any* $y \in \mathcal{S}_\gamma$.

Online Mirror Descent with Barriers. We rely on standard properties of the Online Mirror Descent (OMD) algorithm with a self-concordant barrier function R for a domain $\mathcal{S}$ as regularization, applied to an arbitrary sequence of loss vectors $\ell_1, \ldots, \ell_T \in \mathbb{R}^d$ (Abernethy et al., 2009). Starting from an initial $x_1 \in \mathcal{S}$, OMD makes the following updates for $t = 1, \ldots, T$:[1]

$$x_{t+1} = \nabla R^{-1}\big(\nabla R(x_t) - \eta_t \ell_t\big). \tag{5}$$

This version of OMD has the following guarantee (we include a proof in Appendix B for completeness); here we use the notation $\|\ell\|_x^\star = \sqrt{\ell^\top \nabla^2 R(x)^{-1} \ell}$ for $x \in \mathcal{S}$ and $\ell \in \mathbb{R}^d$.

Lemma 2. *Let* $R : int(\mathcal{S}) \mapsto \mathbb{R}$ *be self-concordant and assume that* $\eta_t \|\ell_t\|_{x_t}^\star \le \frac{1}{2}$ *for all* t. *Then, for any* $u \in \mathcal{S}$,

$$\sum_{t=1}^T \ell_t \cdot (x_t - u) \le \frac{1}{\eta_1} B_R(u \,\|\, x_1) - \sum_{t=2}^T \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) B_R(u \,\|\, x_t) + \sum_{t=1}^T \eta_t (\|\ell_t\|_{x_t}^\star)^2.$$

Observe that when the learning rate sequence is strictly increasing, the middle term in the above bound becomes negative and can potentially serve to decrease the regret of OMD when the divergence $B_R(u \,\|\, x_t)$ is large. This observation is key to our algorithmic development in Section 4.2.

3. Setup and Overview of Results

3.1. Online MDPs with Aggregate Bandit Feedback

We consider an online version of finite-horizon MDPs in which the interaction between learner and the MDP proceeds for K episodes. Before the interaction begins, the environment assigns a sequence of loss functions $\ell_1, \ldots, \ell_k : [H] \times S \times A \times S \mapsto [0,1]$ one for each episode $k \in [K]$. The choice of loss functions is done in an arbitrary, possibly adversarial, manner.

At the start of each episode k the online algorithm defines a policy π_k. At the end of the episode the online algorithm receives the trajectory realized by π_k, i.e., $(s_1^k, a_1^k, \ldots, s_H^k, a_H^k, s_{H+1}^k)$, and the aggregate loss incurred during this trajectory with respect to ℓ_k, i.e., $\sum_{h=1}^H \ell_k(h, s_h^k, a_h^k, s_{h+1}^k)$.

We define the regret of the learner over the K episodes as

$$\mathrm{Reg}_K = \sum_{k=1}^K L^{\pi_k, P, \ell_k} - \min_\pi \sum_{k=1}^K L^{\pi, P, \ell_k},$$

where the minimum is taken over all policies π, and we let $\pi^\star$ denote a minimizer. The regret can also be written in terms of occupancy measures, by noticing that the expected loss of a policy π and transition function P with respect to a loss function ℓ is $L^{\pi,P,\ell} = x^{\pi,P} \cdot \ell$. Thus, the regret of the learner over the K episodes can be written as:

$$\mathrm{Reg}_K = \sum_{k=1}^K x^{\pi_k, P} \cdot \ell_k - \min_\pi \sum_{k=1}^K x^{\pi, P} \cdot \ell_k.$$

[1] Typically, OMD has an additional projection step when employed on a bounded domain. However, when R is a barrier, such a projection is redundant as the OMD update never steps out of the domain (as a consequence of Eq. (3)).

The main result of this paper is a computationally-efficient learning algorithm for the setting described above.

Theorem 3. *There exists an online learning algorithm for* finite-horizon MDPs with aggregated bandit feedback *that guarantees*

$$\mathbb{E}[Reg_K] = \text{poly}(H, |S|, |A|)\, O(\sqrt{K}).$$

Moreover, the per-episode runtime complexity of the algorithm is polynomial in $H, |S|, |A|$, and K.

We prove the theorem by efficiently reducing the online MDPs setting to a sequence of instances of a novel setting we term *Distorted Linear Bandits* (DLB). In what follows, we describe the DLB setting, the reduction, and prove the correctness of the reduction.

We note that the dependence on K in Theorem 3 is optimal due to a lower bound in the simpler setting of stochastic MDPs with bandit feedback (Osband and Van Roy, 2016). The dependence on $H, |S|, |A|$, however, is likely to be suboptimal. We defer the reader to Section 5 where we describe the regret bound with more explicit dependence on the problem parameters.

3.2. Distorted Linear Bandits (DLB)

In this game, the learner plays by picking vectors from a compact and convex body $\mathcal{S} \subseteq \mathbb{R}^d$. We assume that $\|y\|_1 \le H$ for all $y \in \mathcal{S}$ for some $H > 0$. Further, let $\beta > 0$ be a bias parameter. Learning in the DLB setting proceed according the following protocol: Initially, the adversary privately chooses a sequence of loss vectors $\ell_1, \ldots, \ell_T$ and a sequence of perturbation vectors $\epsilon_1, \ldots, \epsilon_T \in [0, \beta]^d$. Then, at rounds $t = 1, \ldots, T$,

(i) Learner selects $y_t \in \mathcal{S}$.

(ii) Adversary picks $z_t \in \mathbb{R}^d$, where $\|z_t\|_1 \le H$ such that $\|z_t - y_t\|_1 \le \min\{|z_t \cdot \epsilon_t|, |y_t \cdot \epsilon_t|\}$.

(iii) A random $\hat{z}_t$ is sampled such that $\mathbb{E}_t[\hat{z}_t \mid z_t] = z_t$ and $\|\hat{z}_t\|_1 \le H$, where $\mathbb{E}_t$ denotes expectation conditioned on all randomness prior to round t.

(iv) The action $\hat{z}_t$ is played; the learner suffers and observes the loss $\ell_t \cdot \hat{z}_t$; the learner additionally observes $\hat{z}_t$ and ϵ_t.

We emphasize that the z_t are arbitrary and can be chosen in an adaptive manner after the learner chooses y_t. Note, however, that we assume that the loss vectors (as well as the perturbation vectors) are chosen before the game starts; namely, the adversary is oblivious.

We define the regret in the DLB setting as follows:

$$\text{Reg}_T = \sum_{t=1}^{T} \hat{z}_t \cdot \ell_t - \min_{z \in \mathcal{S}} \sum_{t=1}^{T} z \cdot \ell_t.$$

The learner's goal is therefore to minimize the losses attained by the perturbations $\hat{z}_1, \ldots, \hat{z}_T$ of the actions $y_1, \ldots, y_T$ chosen by the learner. Clearly, the regret necessarily scales with the magnitudes of $\epsilon_1, \ldots, \epsilon_T$, and our regret bounds will ultimately depend on a parameter B that upper bounds the magnitude of the perturbations via the quantity $\sum_{t=1}^{T} (\hat{z}_t \cdot \epsilon_t)^2$. The following theorem is the main technical result of our work.

Theorem 4. *There exists an efficient (poly-time) online learning algorithm for the DLB setting whose regret is at most* $\mathrm{poly}(H, d, \beta, B)\, O(\sqrt{T})$.

We prove this theorem by showing two online learning algorithms (one is computationally-efficient; the other is not) in Section 4. We conclude this current section by describing the reduction from online MDPs with aggregate bandit feedback to DLB.

3.3. The Reduction

We now show how to reduce the MDP with aggregate feedback problem to instances of DLB described above (proofs of results of this section appear in Appendix B.) Our algorithm for learning MDPs with aggregate feedback is depicted in detail in Appendix A, and here we give a verbal description of the algorithm. The algorithm assumes the existence of a computationally-efficient online learning algorithm for DLB with $O(\sqrt{T})$ regret which exists due to Theorem 4.

The algorithm partitions the K episodes into epochs, where epoch i contains episodes k_i through $k_{i+1} - 1$ ($k_1 = 1$). Each epoch ends whenever the number of visits to some state-action pair s, a at some time step h is doubled. Thus, the total number of epochs is at most $2H|S||A| \log K$.

In epoch i, we produce an empirical estimate of the transition probabilities $\widehat{P}$ based on all observations prior to epoch i. We apply a high probability argument to bound the estimation error P of the dynamics as: $\|\widehat{P}(\cdot \mid s, a, h) - P(\cdot \mid s, a, h)\|_1 \le \epsilon_i(s, a, h)/H$ for a confidence parameter $\epsilon_i : S \times A \times [H] \mapsto [0, \beta]$ associated with epoch i. ($\epsilon_i(s, a, h)$ decreases as a function of the number of times each (s, a, h) has been visited up to epoch i.)

We fix a convex and compact $\mathcal{S}_i$ to be the set of all feasible occupancy measures based on our current estimate of the dynamics of the MDP. We claim that in each epoch, the setting admits to the distorted linear bandits problem. Indeed, we show that with high probability, $\mathcal{S}_i$ contains $x^{\pi^\star, P}$—the occupancy measure associated with the optimal policy and the true dynamics. Now, throughout epoch i, for $k = k_i, \ldots, k_{i+1} - 1$:

(i) Learner picks a policy π_k associated with some occupancy measure $y_k \in \mathcal{S}_i$.

(ii) π_k is played on the true MDP and the learner observes a trajectory $\hat{z}_k$, such that $\hat{z}_k(h, s, a, s') = 1$ iff the trajectory passed through state s at time h, played action a and transitioned to state s'. Otherwise $\hat{z}_k(h, s, a, s') = 0$. The learner suffers and observes the loss of $\ell_k \cdot \hat{z}_k$.

(iii) Let z_k be the occupancy measure of π_k and the true dynamics P; then $\mathbb{E}[\hat{z}_k \mid z_k] = z_k$. We prove that $\|z_k - y_k\|_1 \le \min\{\epsilon_i \cdot z_t, \epsilon_i \cdot y_t\}$.

Moreover, we give a bound of $\sum_{k=k_i}^{k_{i+1}-1} (\epsilon_i \cdot \hat{z}_k)^2 = \widetilde{O}(H^4 |S|^2 |A|)$ as required by the DLB setting.

We consequently apply the DLB algorithm to obtain a regret bound of $O(\sqrt{k_{i+1} - k_i}) = O(\sqrt{K})$ in each epoch, and as the number of epochs is only at most $O(\log K)$ this gives an overall regret bound of $\widetilde{O}(\sqrt{K})$ as required. The complete proof of this claim appears in Appendix B. The analysis of the running time of the algorithm is found in Section 5.

4. Algorithms for Distorted Linear Bandits

In this section we prove Theorem 4 by presenting our online algorithms for the DLB problem. The difficulty of this setting lies in the fact that the main mechanism to cope with lack of information in bandit optimization is to construct unbiased estimates of the loss vectors. In the DLB setting this is impossible to do since the actions chosen by the learner are shifted by the adversary. Nevertheless, having $\sum_{t=1}^{T}(\hat{z}_t \cdot \epsilon_t)^2$ bounded, intuitively means that the estimation bias at the actions played by the learner is bounded in an amortized sense—a useful property that we utilize in our algorithms.

4.1. Simple Approach via Optimism

Our first algorithm is based on what is arguably the most straightforward approach to the problem: construct an "optimistic" estimator to the player's loss—one whose expectation underestimates the loss of all actions at a given round, yet is sufficiently accurate in estimating the player's loss at the same round—and feed it to a standard bandit linear optimization algorithm. However, as we show in this section, such a loss estimator becomes a non-convex (in fact, concave) function of the played action, thus overall this approach leads to a computationally inefficient algorithm. We nevertheless present the algorithm in this work to illustrative why this approach fails before moving on to a more sophisticated approach in Section 4.2.

Throughout this section, we assume that the decision set $\mathcal{S}$ is finite of size $O((HT)^d)$; since for now we are not bound by computational complexity considerations, if $\mathcal{S}$ is a larger (or infinite) set we may replace $\mathcal{S}$ with a $1/(HT)$-net of $\mathcal{S}$, which has the required size. The algorithm we describe below (Algorithm 1) assumes as input an exploration distribution μ over the set $\mathcal{S}$, such that for $y \sim \mu$ it holds that $\mathbb{E}[yy^{\mathsf{T}}] \geq \lambda I$ for a constant $\lambda > 0$. Standard techniques in linear bandit optimization (e.g., Bubeck et al., 2012; Hazan and Karnin, 2016) show that under fairly general conditions on $\mathcal{S}$, one can pick an exploration distribution μ so as $\lambda = \Omega(1/\sqrt{d})$.[2]

Algorithm 1 Distorted Linear Bandits via Optimistic Biases

```
input: $\eta > 0$, $\gamma > 0$, exploration distribution $\mu$.
initialize: $w_1(y) = 1$ for all $y \in \mathcal{S}$.
for $t = 1, \ldots, T$ do
    define probability density $p_t \propto w_t$, and let $q_t = (1-\gamma)p_t + \gamma\mu$.
    sample point $y_t \sim q_t$ in the domain $\mathcal{S}$.
    adversary chooses $z_t$ such that $\|z_t - y_t\|_1 \leq \min\{|y_t \cdot \epsilon_t|, |z_t \cdot \epsilon_t|\}$, and plays $\hat{z}_t$ where
        $\mathbb{E}_t[\hat{z}_t \mid z_t] = z_t$.
    observe $\epsilon_t$ and loss $\ell_t \cdot \hat{z}_t \in [0, H]$.
    compute the second moment of $y_t$: $M_t = \mathbb{E}_t[y_t y_t^{\mathsf{T}}]$.
    compute $\hat{\ell}_t = (\ell_t \cdot \hat{z}_t) M_t^{-1} y_t$ and $\tilde{\ell}_t(y) = \hat{\ell}_t \cdot y - \sqrt{d}\,\|y\|_{M_t^{-1}} \|\epsilon_t\|_{M_t}$.
    update $w_{t+1}(y) = w_t(y) \cdot \exp(-\eta \tilde{\ell}_t(y))$,    $\forall\, y \in \mathcal{S}$.
end for
```

The algorithm relies on a standard estimator $\hat{\ell}_t = (\ell_t \cdot \hat{z}_t) M_t^{-1} y_t$ to estimate the loss vector ℓ_t. Note that if it were that $z_t = y_t$ then this would have been an unbiased estimator for the loss, i.e., $\mathbb{E}_t[\hat{\ell}_t] = \ell_t$. However, due to the adversarial perturbations z_t might be shifted away from the intended y_t. We thus modify the estimator to account for this shift and make it "optimistic," in the sense that its expectation is a lower bound on the real loss function. Given these corrected estimates, the rest of the algorithmic development follows standard lines in the linear bandit optimization literature (Dani et al., 2008; Bubeck et al., 2012).

[2] Some of these techniques rely on solving intractable optimization problems, but recall that in the context of this section we are not concerned by the computational complexity of the resulting algorithm.

Concretely, we define the following *bias-corrected* loss functions:

$$\tilde{\ell}_t(y) = \hat{\ell}_t \cdot y - \sqrt{d}\,\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}, \qquad \forall\, t \in [T], y \in \mathcal{S}\,.$$

Then, the algorithm essentially performs multiplicative-weights updates on the modified loss functions $\tilde{\ell}_t(y)$, which can be seen to be a *concave* function of y. In general, it is a hard problem to sample from the resulting distributions q_t given that these losses are concave. (If, on the other hand, they were convex, then the resulting distributions would have been log-concave for which efficient sampling algorithms are well-known.) Therefore, the algorithm is computationally inefficient.

We prove that Algorithm 1 provides the following regret guarantee.

Theorem 5. *Set $\eta = (2H\beta d)^{-1}\sqrt{\log|\mathcal{S}|/T}$, $\gamma = 2H^2(H+\beta\sqrt{d})\eta/\lambda$. Then, given that $B \ge \sum_{t=1}^T(\hat{z}_t \cdot \epsilon_t)^2$ (almost surely), for any $y^\star \in \mathcal{S}$, Algorithm 1 satisfies that*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t \cdot (\hat{z}_t - y^\star)\right] = \widetilde{O}\left(H\beta d + \beta d\sqrt{B} + \frac{H^3}{\beta\lambda d} + \frac{H^2}{\lambda\sqrt{d}}\right)\sqrt{T},$$

provided that $\beta \ge 1$ and $T \ge (4H^2(H+\beta\sqrt{d})^2\log|\mathcal{S}|)/(\lambda^2\beta^2 d^2)$.

We only sketch the proof here, deferring details and precise bounds to Appendix B.

Proof (sketch). We begin by showing that $\tilde{\ell}_t$ is indeed an underestimate of the true loss (see Lemma 6 below): $\mathbb{E}_t[\tilde{\ell}_t(y)] \le \ell_t \cdot y$ for any $y \in \mathcal{S}$. For the converse direction, we show that in expectation over the learner's decision, $\tilde{\ell}_t$ is close to ℓ_t in the following sense:

$$\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} p_t(y)\tilde{\ell}_t(y)\right] \ge \mathbb{E}_t[\ell_t \cdot \hat{z}_t] - \gamma H - 5d\|\epsilon_t\|_{M_t}.$$

With these two results at hand, we argue that the regret of Algorithm 1 is bounded by the regret of the Multiplicative Weights updates, plus an additive error term that scales with the perturbations ϵ_t:

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t \cdot (\hat{z}_t - y^\star)\right] \le \mathbb{E}\left[\sum_{t=1}^T\sum_{y\in\mathcal{S}} p_t(y)\big(\tilde{\ell}_t(y) - \tilde{\ell}_t(y^\star)\big)\right] + \gamma HT + 5d\,\mathbb{E}\left[\sum_{t=1}^T \|\epsilon_t\|_{M_t}\right]. \tag{6}$$

Next, we apply a standard second-order regret bound of Multiplicative Weights to obtain the following:

$$\sum_{t=1}^T\sum_{y\in\mathcal{S}} p_t(y)\big(\tilde{\ell}_t(y) - \tilde{\ell}_t(y^\star)\big) \le \frac{\log|\mathcal{S}|}{\eta} + \eta\sum_{y\in\mathcal{S}} p_t(y)\big(\tilde{\ell}_t(y)\big)^2,$$

and we bound the term $\mathbb{E}_t[\sum_{y\in\mathcal{S}} p_t(y)(\tilde{\ell}_t(y))^2] \le 8(H\beta d)^2$ using simple algebra.

The theorem is now given by combining the second-order regret bound above together with Eq. (6), and by bounding the bias terms using the DLB setting assumptions, as $\mathbb{E}[\sum_{t=1}^T\|\epsilon_t\|_{M_t}] \le 2\beta\sqrt{BT}$. ■

We now prove that as mentioned, the expectation of $\tilde{\ell}_t$ is an underestimate of the true loss.

Lemma 6. $\mathbb{E}_t[\tilde{\ell}_t(y)] \le \ell_t \cdot y$ *for any $y \in \mathcal{S}$.*

Proof. Observe that

$$\begin{aligned}
\mathbb{E}_t[\hat{\ell}_t] &= \mathbb{E}_t[(\ell_t \cdot z_t) M_t^{-1} y_t] \\
&= \mathbb{E}_t[M_t^{-1} y_t y_t^\top \ell_t] + \mathbb{E}_t[M_t^{-1} y_t (z_t - y_t) \cdot \ell_t] \\
&= \ell_t + M_t^{-1} \mathbb{E}_t[y_t (z_t - y_t) \cdot \ell_t].
\end{aligned}$$

Our assumptions imply that $|(z_t - y_t) \cdot \ell_t| \le \|z_t - y_t\|_1 \|\ell_t\|_\infty \le |y_t \cdot \epsilon_t|$. Thus, by two applications of Cauchy-Schwartz, for any $y \in \mathcal{S}$ we obtain

$$\begin{aligned}
\mathbb{E}_t[|(\hat{\ell}_t - \ell_t) \cdot y|] &= \mathbb{E}_t\left[|y^\top M_t^{-1} y_t| \cdot |(z_t - y_t) \cdot \ell_t|\right] \\
&\le \|y\|_{M_t^{-1}} \mathbb{E}_t\left[\|y_t\|_{M_t^{-1}} |y_t \cdot \epsilon_t|\right] \\
&\le \|y\|_{M_t^{-1}} \sqrt{\mathbb{E}_t[\|y_t\|_{M_t^{-1}}^2]\, \mathbb{E}_t[(y_t \cdot \epsilon_t)^2]} \qquad (7) \\
&= \|y\|_{M_t^{-1}} \sqrt{\mathbb{E}_t[y_t^\top M_t^{-1} y_t]} \sqrt{\epsilon_t^\top \mathbb{E}_t[y_t y_t^\top] \epsilon_t} \\
&= \sqrt{d} \|y\|_{M_t^{-1}} \|\epsilon_t\|_{M_t}.
\end{aligned}$$

This means that

$$\mathbb{E}_t[\tilde{\ell}_t(y) - \ell_t \cdot y] = \mathbb{E}_t[(\hat{\ell}_t - \ell_t) \cdot y] - \sqrt{d} \|y\|_{M_t^{-1}} \|\epsilon_t\|_{M_t} \le 0$$

which proves that $\mathbb{E}[\tilde{\ell}_t(y)] \le \ell_t \cdot y$ for any $y \in \mathcal{S}$. ■

4.2. Efficient Approach via OMD with Increasing Learning Rates

Our previous algorithm enjoys an $\widetilde{O}(\sqrt{T})$ regret bound, but it is inherently computationally inefficient. In this section we take a different approach that leads to an algorithm with $\widetilde{O}(\sqrt{T})$ regret, but one that can also be implemented efficiently.

Algorithm 2 Distorted Linear Bandits via Increasing Learning Rates

```
input: $\eta_0 > 0$, $\vartheta$-self-concordant barrier $R : \mathrm{int}(\mathcal{S}) \mapsto \mathbb{R}$.
init: $x_1 = \arg\min_{x \in \mathcal{S}} R(x)$.
for $t = 1, \ldots, T$ do
    sample $u_t$ uniformly at random from the unit sphere of $\mathbb{R}^d$.
    predict $y_t = x_t + \nabla^2 R(x_t)^{-1/2} u_t$.
    adversary chooses $z_t$ such that $\|z_t - y_t\|_1 \le \min\{|y_t \cdot \epsilon_t|, |z_t \cdot \epsilon_t|\}$, and plays $\hat{z}_t$ where
        $\mathbb{E}_t[\hat{z}_t \mid z_t] = z_t$.
    observe $\hat{z}_t, \epsilon_t$, and loss $\ell_t \cdot \hat{z}_t \in [0, H]$.
    construct $\tilde{\ell}_t = d(\ell_t \cdot \hat{z}_t) \nabla^2 R(x_t)^{1/2} u_t$.
    update $\eta_t^{-1} = \eta_{t-1}^{-1} - 2d\,|\hat{z}_t \cdot \epsilon_t|$.
    set $x_{t+1} = \nabla R^{-1}(\nabla R(x_t) - \eta_t \tilde{\ell}_t)$.
end for
```

Algorithm 2 is based on Online Mirror Descent with a self-concordant barrier R as a regularizer (Abernethy et al., 2009). The algorithm maintains a sequence of points $x_1, \ldots, x_T \in \mathcal{S}$. In line 5, the algorithm makes a prediction y_t by sampling uniformly at random from the ellipsoid $\{y : \|y - x_t\|_{x_t} \le 1\}$, known as the *Dikin Ellipsoid* associated with R at x_t, that is always contained in $\mathcal{S}$ (this follows from Eq. (3)). Then, according to the DLB protocol, the algorithm receives ϵ_t, $\hat{z}_t$ and loss $\ell_t \cdot \hat{z}_t$ such that $z_t = \mathbb{E}_t[\hat{z}_t \mid z_t]$ where z_t is a perturbation of y_t.

The algorithm proceeds to construct an estimator $\tilde{\ell}_t$ of the loss vector ℓ_t in line 8. Note that if we replace $\hat{z}_t$ with y_t in line 8, then $\tilde{\ell}_t$ would be an unbiased estimator. However, as this is not the case, the algorithm must mitigate the bias in the $\tilde{\ell}_t$, and does that by increasing its learning rate according to the perturbation magnitude $|\hat{z}_t \cdot \epsilon_t|$ (line 9). Finally, in line 10, the algorithm performs the mirror descent update.

Algorithm 2 can be implemented efficiently as long as $\mathcal{S}$ is not degenerate (namely, $\mathcal{S}$ is compact and has volume in $\mathbb{R}^d$, and thus admits a proper self-concordant barrier R) and as long as gradients and Hessians of R can be computed efficiently. We defer a more detailed discussion of implementation issues to Section 5.

Our main result regarding the algorithm is as follows.

Theorem 7. *Algorithm 2 with $\eta_0 = \widetilde{\Theta}(\vartheta/(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T}))$ provides the following regret guarantee, for any $y^\star \in \mathcal{S}$:*

$$\mathbb{E}\left[\sum_{t=1}^{T}(\hat{z}_t - y^\star)\cdot \ell_t\right] = \widetilde{O}\big(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T}\big),$$

provided that $B \ge \max\{\sum_{t=1}^{T}(\hat{z}_t \cdot \epsilon_t)^2, H\}$ (almost surely).

Here we sketch the proof of Theorem 7 highlighting the key ideas; the complete proof and precise bounds can be found in Appendix B.

Proof (sketch). The first part of the proof is straightforward. We split the regret into three terms:

$$\mathbb{E}\left[\sum_{t=1}^{T}(\hat{z}_t - y^\star)\cdot \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}(z_t - x_t)\cdot \ell_t\right] + \mathbb{E}\left[\sum_{t=1}^{T}(x_t - y^\star_\gamma)\cdot \ell_t\right] + \mathbb{E}\left[\sum_{t=1}^{T}(y^\star_\gamma - y^\star)\cdot \ell_t\right], \quad (8)$$

where $y^\star_\gamma = (1-\gamma)y^\star + \gamma x_1 \in \mathcal{S}_\gamma$ for sufficiently small γ. (Following a standard technique, we introduce $y^\star_\gamma$ as otherwise we would eventually have to bound $B_R(y^\star \,\|\, x_1)$ which might be arbitrarily large; by introducing $y^\star_\gamma$, we instead would have to bound $B_R(y^\star_\gamma \,\|\, x_t)$, which is bounded by Lemma 1.) The first summand in Eq. (8) pertains to the bias generated by the perturbation of y_t to z_t, and is bounded by $\sqrt{BT}$; the third summand bounds the loss difference between that of $y^\star$ and of $y^\star_\gamma$, and is bounded by $2\gamma HT$. All of these quantities are $\widetilde{O}\big(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T}\big)$.

The heart of the proof focuses on bounding the second summand. To this end, we apply Lemma 8 (see below) to bound the instantaneous regret of the algorithm at each time step t, by the instantaneous regret using the loss estimator $\tilde{\ell}_t$ plus an additional bias term that scales with $\|x_t - y^\star_\gamma\|_{x_t}$. This results with

$$(x_t - y^\star_\gamma)\cdot \ell_t \le \mathbb{E}_t[(x_t - y^\star_\gamma)\cdot \tilde{\ell}_t] + d\|x_t - y^\star_\gamma\|_{x_t}\mathbb{E}_t[|\hat{z}_t \cdot \epsilon_t|], \quad (9)$$

and we proceed in bounding $\mathbb{E}[\sum_{t=1}^{T}\mathbb{E}_t[(x_t - y^\star_\gamma)\cdot \tilde{\ell}_t]] = \mathbb{E}[\sum_{t=1}^{T}(x_t - y^\star_\gamma)\cdot \tilde{\ell}_t]$. Since the algorithm is taking OMD steps with loss vectors $\tilde{\ell}_t$, we can apply Lemma 2 to get

$$\mathbb{E}\left[\sum_{t=1}^{T}(x_t - y^\star_\gamma)\cdot \tilde{\ell}_t\right] \le \mathbb{E}\left[\frac{1}{\eta_1}B_R(y^\star_\gamma \,\|\, x_1) - \sum_{t=2}^{T}\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)B_R(y^\star_\gamma \,\|\, x_t) + \sum_{t=1}^{T}\eta_t(\|\tilde{\ell}_t\|^\star_{x_t})^2\right].$$

Handling the first and third terms is standard (following Abernethy et al., 2009), and they are shown to be bounded by $O((\vartheta/\eta_0)\log(1/\gamma))$ and $O(\eta_0 d^2 H^2 T)$ respectively, both are $\widetilde{O}(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T})$ for our choice of parameters. The middle term in the bound above is what enables the algorithm to compensate for the bias in the loss estimation by employing an increasing learning rate schedule. Indeed, together with Eq. (9) we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(x_t - y_\gamma^\star)\cdot \ell_t\right] \le d\,\mathbb{E}\left[\sum_{t=1}^{T}\|x_t - y_\gamma^\star\|_{x_t}|\hat{z}_t \cdot \epsilon_t|\right] - \mathbb{E}\left[\sum_{t=2}^{T}\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)B_R(y_\gamma^\star \,\|\, x_t)\right] \\ + \widetilde{O}\big(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T}\big).\end{aligned} \tag{10}$$

The key observation is that the divergence $B_R(y_\gamma^\star \,\|\, x_t)$ here is directly related to the bias term $\|y_\gamma^\star - x_t\|_{x_t}$ via Lemma 9 (found below), as $B_R(y_\gamma^\star \,\|\, x_t) \ge \frac{1}{2}\|y_\gamma^\star - x_t\|_{x_t} - 1$. Now, with our particular setting of learning rates (line 9) the second term in Eq. (10) is upper bounded by $-d\sum_{t=1}^{T}\|y_\gamma^\star - x_t\|_{x_t}\mathbb{E}_t[|\hat{z}_t \cdot \epsilon_t|] + O((\vartheta/\eta_0)\log(1/\gamma))$ (in expectation), which precisely cancels out the first summation over the bias terms and gives the $\widetilde{O}\big(d\vartheta\sqrt{BT} + dH\sqrt{\vartheta T}\big)$ regret bound. ■

The following lemma bounds the instantaneous regrets suffered by the algorithm, by the algorithm's estimates of the instantaneous regret plus an additive bias term that scales as $\|x_t - x\|_{x_t}$.

Lemma 8. *Let $x \in \mathcal{S}$. Then, $(x_t - x)\cdot \ell_t \le \mathbb{E}_t[(x_t - x)\cdot \tilde{\ell}_t] + d\|x_t - x\|_{x_t}\mathbb{E}_t[|\hat{z}_t \cdot \epsilon_t|]$.*

Proof. Recall that x_t is determined given the randomness up to time t. We have that

$$\begin{aligned}\mathbb{E}_t\big[(x_t - x)\cdot \tilde{\ell}_t\big] &= \mathbb{E}_t\big[(x_t - x)\cdot d(\ell_t \cdot \hat{z}_t)\nabla^2 R(x_t)^{1/2}u_t\big] \\ &= \mathbb{E}_t\big[(x_t - x)\cdot d\big(\ell_t \cdot \mathbb{E}_t[\hat{z}_t \mid z_t]\big)\nabla^2 R(x_t)^{1/2}u_t\big] \\ &= \mathbb{E}_t\big[(x_t - x)\cdot d(\ell_t \cdot z_t)\nabla^2 R(x_t)^{1/2}u_t\big] \\ &= \underbrace{\mathbb{E}_t\big[(x_t - x)\cdot d(\ell_t \cdot x_t)\nabla^2 R(x_t)^{1/2}u_t\big]}_{(1)} + \underbrace{\mathbb{E}_t\big[(x_t - x)\cdot d\big(\ell_t \cdot (y_t - x_t)\big)\nabla^2 R(x_t)^{1/2}u_t\big]}_{(2)} \\ &\qquad - \underbrace{\mathbb{E}_t\big[(x_t - x)\cdot d\big(\ell_t \cdot (y_t - z_t)\big)\nabla^2 R(x_t)^{1/2}u_t\big]}_{(3)}.\end{aligned}$$

Next, we analyze each of the three summands above. As the only randomness given the history up to time t is in u_t, we have $(1) = 0$, and as $y_t - x_t = \nabla^2 R(x_t)^{-1/2}u_t$, we have

$$\begin{aligned}(2) &= \mathbb{E}_t\big[(x_t - x)\cdot d\nabla^2 R(x_t)^{1/2}u_t\,(y_t - x_t)\cdot \ell_t\big] \\ &= \mathbb{E}_t\big[(x_t - x)\cdot d\,\nabla^2 R(x_t)^{1/2}u_t u_t^{\mathsf{T}}\nabla^2 R(x_t)^{-1/2}\ell_t\big] \\ &= (x_t - x)\cdot d\,\nabla^2 R(x_t)^{1/2}\mathbb{E}_t\big[u_t u_t^{\mathsf{T}}\big]\nabla^2 R(x_t)^{-1/2}\ell_t \\ &= (x_t - x)\cdot d\,\nabla^2 R(x_t)^{1/2}\cdot \tfrac{1}{d}I \cdot \nabla^2 R(x_t)^{-1/2}\ell_t \\ &= (x_t - x)\cdot \ell_t.\end{aligned}$$

For term (3), two applications of Hölder's inequality yield

$$(3) \le d\,\mathbb{E}_t\big[\|x_t - x\|_{x_t}\|\nabla^2 R(x_t)^{1/2}u_t\|_{x_t}^\star\|\ell_t\|_\infty\|y_t - z_t\|_1\big].$$

Now, to obtain the lemma, we use our assumption that $\|\ell_t\|_\infty \leq 1$, that

$$\mathbb{E}_t[\|y_t - z_t\|_1] \leq \mathbb{E}_t[|z_t \cdot \epsilon_t|] = \mathbb{E}_t\left[|\mathbb{E}_t[\hat{z}_t \mid z_t] \cdot \epsilon_t|\right] \leq \mathbb{E}_t[|\hat{z}_t \cdot \epsilon_t|],$$

by Jensen's inequality, and finally $\|R(x_t)^{1/2}u_t\|^\star_{x_t} = 1$ due to Lemma 19 (see Appendix B). ■

The next lemma lower bounds the Bregman divergence of any point $x \in \mathcal{S}$ from x_t by an order of their distance in local norm; i.e., $\|x - x_t\|_{x_t}$.

Lemma 9. *Let $x \in \mathcal{S}$. Then, $B_R(x \,\|\, x_t) \geq \frac{1}{2}\|x - x_t\|_{x_t} - 1$.*

Proof. Recall that $B_R(x \,\|\, x_t) \geq \rho(\|x - x_t\|_{x_t})$ by Eq. (4) where $\rho(z) = z - \log(1+z)$. Since $\rho(z)$ is convex, we can lower bound

$$\rho(z) \geq \rho(1) + \rho'(1) \cdot (z-1) = \frac{1}{2} - \log(2) + \frac{1}{2}z \geq \frac{1}{2}z - 1,$$

which yields the lemma's statement for $z = \|x - x_t\|_{x_t}$. ■

5. Efficient Implementation of the Reduction

In this section we complete the proof of Theorem 3 by showing a computationally-efficient reduction between Finite-Horizon MDPs with Aggregate Feedback and that of distorted linear bandits.

Recall the reduction in Section 3.3 in which we showed how to solve a Finite-Horizon MDPs with Aggregated Feedback by constructing a sequence of $O(\log K)$ instances (epochs) of the distorted linear bandits problem and running a no-regret algorithm in each such instance (which exists due to Theorem 4). In subsequent sections we reviewed Algorithms 1 and 2, both of which guarantee no-regret for DLB. In this section we make the choice of the algorithm for the reduction explicit by fixing it to be Algorithm 2. Note that the reduction itself, as well as Algorithm 2, can be implemented in polynomial-time as long as in each epoch i, the barrier R chosen for $\mathcal{S}_i$ can be computed efficiently. However, Algorithm 2 is made for the case in which $\mathcal{S}$ has volume in $\mathbb{R}^d$ which is not the case of our body $\mathcal{S}_i$. Thus, in what follows we give two options on how to alleviate this problem and build an efficiently-computable barrier function for each option. In option 1, we show how to alter Algorithm 2 to accommodate the case for $\mathcal{S}_i$ not being fully-dimensional. In option 2, we keep Algorithm 2 as it is, but change the reduction so that $\mathcal{S}_i$ has a small volume in $\mathbb{R}^d$.

Option 1. We follow a technique used in Lee et al. (2020). The set $\mathcal{S}_i$ consists of an intersection between linear equations (Eqs. (17) to (19)) of the form $c_i \cdot x = d_i$ for $i = 1, \ldots, p$ and linear inequalities (Eqs. (16) and (20)) of the form $a_i \cdot x \leq b_i$ for $i = 1, \ldots, m$ where $m = O(|S|^2 H|A|)$. Our approach is to set the log barrier $R(x) = -\sum_{i=1}^m \log(b_i - a_i \cdot x)$ over the inequalities (note that its barrier parameter ϑ is m; see Nemirovski, 2004). However, we still have to handle the linear equations in order to make sure that Algorithm 2 will not generate predictions that are not in $\mathcal{S}_i$.

Recall that Algorithm 2 is essentially a variant of OMD, which commonly has a projection step that does not appear in Algorithm 2. First, we add a projection step in Algorithm 2 after line 10 onto the affine subspace defined by the linear equations of $\mathcal{S}_i$: we replace line 10 with $x'_{t+1} = \nabla R^{-1}(\nabla R(x_t) - \eta_t \tilde{\ell}_t)$, and then add after line 10: $x_{t+1} = \arg\min_{x: Cx=d} B_R(x \,\|\, x'_{t+1})$, where C is a matrix whose columns are $c_1, \ldots, c_p$. This validates that the iterates $x_1, x_2, \ldots$ are in $\mathcal{S}_i$.

Second, recall that originally y_k, is sampled uniformly at random from the Dikin ellipsoid centered at x_t: $\{y : \|y - x_t\|_{x_t} \leq 1\}$. Concretely, $y_k = x_k + \nabla^2 R(x_k)^{-1/2} u_k$ (line 5) for u_k sampled uniformly at random from the unit sphere of $\mathbb{R}^d$. We also like to make sure that y_k is in the aforementioned affine subspace, by instead sampling y_k uniformly at random from the intersection of the Dikin ellipsoid with the affine subspace. To achieve this, we let W be an orthogonal matrix whose range spans the null space of C. We now sample u_k uniformly from the unit sphere in $\mathbb{R}^p$. We replace line 5 in Algorithm 2 by choosing $y_k = x_k + WW^\intercal \nabla^2 R(x_k)^{-1/2} W u_k$, so now $y_k - x_k$ is in the null space spanned by $c_1, \ldots, c_p$. Moreover, we have $y_k \in \mathcal{S}_i$ as we show that $\|y_k - x_k\|_{x_k} = 1$ and by Eq. (3) (proof in Appendix B).

The estimators $\{\tilde{\ell}_t\}_{t=1}^T$ have to be changed accordingly. We change line 8 by redefining $\tilde{\ell}_t = p(\ell_t \cdot \hat{z}_t) WW^\intercal \nabla^2 R(x_t)^{1/2} W u_t$. The proof of Lemma 8 is altered to accommodate for these changes, using the fact that $x_k - x$ is in the span of W. In addition, we replace the technical results in Lemma 19, by these of Lemma 21 (both found in Appendix B). The rest of the proof of the analysis of Algorithm 2 remains without any further changes.

Option 2. In this option, instead of altering Algorithm 2, we alter $\mathcal{S}_i$ to give it a small volume in $\mathbb{R}^d$. We replace the p linear equations of the form $c_i \cdot x = d_i$ with linear inequalities of the form $|c_i \cdot x - d_i| \leq 1/\mathrm{poly}(K)$. We then set the barrier on the new body to be the log barrier of the new set of linear inequalities:

$$R(x) = -\sum_{i=1}^{m} \log(b_i - a_i \cdot x) - \sum_{i=1}^{p} \log(\mathrm{poly}(K)^{-1} - |d_i - c_i \cdot x|),$$

which also has a barrier parameter of $\vartheta = O(|S|^2|A|H)$ (number of linear inequalities defining the new body; see Nemirovski, 2004).

The issue here is that, when running Algorithm 2 on the new body, we might choose y_k that is on the exterior of $\mathcal{S}_i$. However, we could then replace y_k by its projection, which we denote by y'_k, onto $\mathcal{S}_i$ and play that projection instead. Note that $\|y_k - y'_k\|_1 \leq O(1/\mathrm{poly}(K))$. This ensures that $\|y_k - z_k\|_1 \leq |\epsilon_i \cdot z_k| + O(1/\mathrm{poly}(K))$ and fulfills the assumptions of the distorted linear bandits setting (Section 3.2) thus ensuring that Algorithm 2 will maintain its $\widetilde{O}(\sqrt{K})$ regret bound.

Finally, we note that taking either option 1 or option 2 yields the following, more explicit, regret guarantee (proof in Appendix B).

Corollary 10. *There exists a learning algorithm for finite-horizon MDPs with aggregate bandit feedback that attains a regret bound of*

$$\mathbb{E}[Reg_K] = \widetilde{O}\big(|S|^6|A|^{5/2}H^5\sqrt{K}\big),$$

and runs in time per episode that is polynomial in $H, |S|, |A|, K$.

Acknowledgments

HK is supported by Israel Science Foundation grant no. 1595-19, and the Blavatnik Family Foundation. TK is supported in part by the Israeli Science Foundation (ISF) grant no. 2549/19, by the Len Blavatnik and the Blavatnik Family foundation, and by the Yandex Initiative in Machine Learning. YM received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17) and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

Yasin Abbasi Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. *Advances in neural information processing systems*, 26:2508–2516, 2013.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. *Conference on Learning Theory*, 2009.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38, 2017.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342. Omnipress, 2008.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1, 2012.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. *arXiv preprint arXiv:2008.06036*, 2020.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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*, pages 4860–4869, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33, 2020.

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Conference on Learning Theory*, pages 109–123, 2004.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *Conference on Learning Theory*, pages 231–243, 2010.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59(3):676–691, 2013.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2209–2218, 2019.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Appendix A. Reduction Algorithm

Algorithm 3 Reduction from online MDPs with aggregate feedback to DLB

```
**Init**: $N_1(s,a,h) = 0,\ N(s,a,h,s') = 0, \quad \forall (s,a,h,s') \in S \times A \times [H] \times S,\ k = 1.$
**for** epoch $i = 1, 2, \dots$ **do**
    **construct** empirical transition function:
        $$\widehat{P}_i(s' \mid s,a,h) = \frac{N_i(s,a,h,s')}{\max\{N_i(s,a,h), 1\}}, \quad \forall (s,a,h,s') \in S \times A \times [H] \times S. \qquad (11)$$
    **set** confidence bounds:
        $$\epsilon_i(s,a,h) = 5H\sqrt{\frac{|S| + \log(H|S||A|K/\delta)}{\max\{N_i(s,a,h), 1\}}}, \quad \forall (s,a,h) \in S \times A \times [H]. \qquad (12)$$
    **construct** polytope of feasible occupancy measure $\mathcal{S}_i$ (Eqs. (17) to (20)).
    **init**: $n_i(s,a,h) = 0,\ n_i(s,a,h,s') = 0$ **for all** $(s,a,h,s') \in S \times A \times [H] \times S$.
    **while** $n_i(s,a,h) < \max\{N_i(s,a,h), 1\}$ **for all** $(s,a,h) \in S \times A \times [H]$ **do**
        **predict** occupancy measure $y_k \in \mathcal{S}_i$ using algorithm from Theorem 4.
        **play** $\pi_k$ such that $\pi_k = \pi^{(y_k)}$ (recall Eq. (2)).
        **observe** trajectory $\hat{z}_k$, and aggregate loss $\ell_k \cdot \hat{z}_k$.
        **feed** DLB algorithm with $\hat{z}_k$, $\ell_k \cdot \hat{z}_k$, and $\epsilon_i$.
        **increment**: $k = k + 1$, $n_i(s,a,h,s') = n_i(s,a,h,s') + \hat{z}_k(s,a,h,s')$, and $n_i(s,a,h) = \sum_{s' \in S} n_i(s,a,h,s')$.
    **end while**
    **update**: $N_{i+1}(s,a,h) = N_i(s,a,h) + n_i(s,a,h)$, $N_{i+1}(s,a,h,s') = N_i(s,a,h,s') + n_i(s,a,h,s')$.
**end for**
```

Appendix B. Deferred Proofs

B.1. Proof of Lemma 1

Proof. Note that $B_R(y \,\|\, x_1) \le R(y) - R(x_1)$ since $\nabla R(x_1) \cdot (y - x_1) \ge 0$ by the first-order optimality criterion of x_1. Since $y = (1-\gamma)x + \gamma x_1$ for some $x \in \mathcal{S}$,

$$\pi_{x_1}(y) = \inf\left\{t > 0 : x_1 + t^{-1}(1-\gamma)(y - x_1) \in \mathcal{S}\right\} \le 1 - \gamma.$$

We now bound $R(y) - R(x_1)$ using Eq. (22). ■

B.2. Proof of Lemma 2

For the proof we shall need the following fact about Bregman divergences. For any $x, y, z \in \operatorname{int}(\mathcal{S})$, it satisfies the following equation (easily shown):

$$B_R(y \,\|\, x) = B_R(y \,\|\, z) + B_R(z \,\|\, x) - (\nabla R(x) - \nabla R(z)) \cdot (y - z). \qquad (13)$$

Proof. First let us show that $B_R(u \,\|\, x'_{t+1}) \ge B_R(u \,\|\, x_{t+1})$. Note that B_R is convex in its first argument, and x_{t+1} minimizes $B_R(\,\cdot\, \|\, x'_{t+1})$, entails $(\nabla R(x_{t+1}) - \nabla R(x'_{t+1})) \cdot (x_{t+1} - u) \le 0$, due to the first-order optimality of convex functions. Therefore, by Eq. (13),

$$\begin{aligned} B_R(u \,\|\, x'_{t+1}) &= B_R(u \,\|\, x_{t+1}) + B_R(x_{t+1} \,\|\, x'_{t+1}) - (\nabla R(x_{t+1}) - \nabla R(x'_{t+1})) \cdot (x_{t+1} - u) \\ &\ge B_R(u \,\|\, x_{t+1}), \end{aligned} \tag{14}$$

by the first-order optimality criterion of the projection step and the non-negativity of the Bregman divergence.

Next, we follow the standard mirror-descent analysis, reusing Eq. (13), to obtain

$$\eta_t \ell_t \cdot (x_t - u) = (\nabla R(x_t) - \nabla R(x'_{t+1})) \cdot (x_t - u) = B_R(u \,\|\, x_t) - B_R(u \,\|\, x'_{t+1}) + B_R(x_t \,\|\, x'_{t+1}).$$

Combining with Eq. (14) and summing over $t = 1, \ldots, T$:

$$\sum_{t=1}^{T} \ell_t \cdot (x_t - u) \le \sum_{t=1}^{T} \frac{1}{\eta_t} (B_R(u \,\|\, x_t) - B_R(u \,\|\, x_{t+1})) + \sum_{t=1}^{T} \frac{1}{\eta_t} B_R(x_t \,\|\, x'_{t+1}),$$

where, using $B_R(u \,\|\, x_{T+1}) \ge 0$,

$$\sum_{t=1}^{T} \frac{1}{\eta_t} (B_R(u \,\|\, x_t) - B_R(u \,\|\, x_{t+1})) \le \frac{1}{\eta_1} B_R(u \,\|\, x_1) - \sum_{t=2}^{T} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) B_R(u \,\|\, x_t).$$

Now denote $z = x_t - x'_{t+1}$. For the term $B_R(x_t \,\|\, x'_{t+1})$, Eq. (4) entails that

$$\begin{aligned} B_R(x_t \,\|\, x'_{t+1}) &= R(x_t) - R(x'_{t+1}) - \nabla R(x'_{t+1}) \cdot z \\ &\le \big(\nabla R(x_t) - \nabla R(x'_{t+1})\big) \cdot z - \rho\big(\|z\|_{x_t}\big) \\ &= \eta_t \ell_t \cdot z - \rho\big(\|z\|_{x_t}\big) \\ &\le \eta_t \|\ell_t\|_{x_t}^{\star} \cdot \|z\|_{x_t} - \rho\big(\|z\|_{x_t}\big) && \text{(Hölder inequality)} \\ &\le \sup_{\alpha \in \mathbb{R}} \{\eta_t \|\ell_t\|_{x_t}^{\star} \cdot \alpha - \rho(\alpha)\} \\ &= \rho^{\star}\big(\eta_t \|\ell_t\|_{x_t}^{\star}\big), \end{aligned}$$

where $\rho^{\star}$ is the Fenchel conjugate of ρ: $\rho^{\star}(x) = -x - \log(1 - x)$ defined for any $x < 1$. The final statement is then given using $\rho^{\star}(x) \le x^2$ for any $x \in [0, 1/2]$. ■

B.3. Proof of Theorem 3

Theorem 3 (restated). *There exists an online algorithm for Finite-Horizon MDPs with Aggregated Feedback of expected regret,*

$$\mathbb{E}[Reg_K] = \text{poly}(|S|, |A|, H)\ \widetilde{O}(\sqrt{K}),$$

in K episodes.

In the remainder of this section we prove that the assumptions of the DLB setting hold in each epoch with high probability, and bound the constants β, H, B (defined in Section 3.2). The following lemma quantifies how concentrated are our empirical estimates of the dynamics (Eq. (11)) around the true values.

Lemma 11. *With probability at least $1 - \delta$, the following holds for all epochs $i = 1, 2, \dots$ simultaneously:*

$$\|P(\cdot \mid s, a, h) - \widehat{P}_i(\cdot \mid s, a, h)\|_1 \leq 5\sqrt{\frac{|S| + \log\big(H|S||A|K/\delta\big)}{\max\{N_i(s, a, h), 1\}}}, \qquad \forall (s, a, h) \in S \times A \times [H]. \quad (15)$$

To prove the lemma, we need the following simple technical result.

Lemma 12 (Weissman et al., 2003). *Let $p(\cdot)$ be a distribution over m elements, and let $\bar{p}_t(\cdot)$ be the empirical distribution defined by t i.i.d. samples from $p(\cdot)$. Then, with probability at least $1 - \delta$,*

$$\|\bar{p}_t(\cdot) - p(\cdot)\|_1 \leq 2\sqrt{\frac{m + \log(\delta^{-1})}{t}}.$$

Proof of Lemma 11. Note that any state-action pair can be sampled at time h during the episode at most K times over the entire K episodes. Then, the lemma from Lemma 12 and a union bound over all $(s, a, h) \in S \times A \times [H]$ and over all possible number of times in which (s, a, h) can be sampled in total. ■

Now, let i be any epoch. Before defining the set of feasible occupancy measures for epoch i, $\mathcal{S}_i$, let us first simplify our notation. We write for any occupancy measure x,

$$x(h, s, a) = \sum_{s' \in S} x(h, s, a, s'); \quad x(h, s) = \sum_{a \in A} x(h, s, a); \quad \text{and} \quad x(h) = \sum_{s \in S} x(h, s).$$

We define $\mathcal{S}_i$ as follows:

$$\begin{aligned}
\mathcal{S}_i = \Big\{ x \in \mathbb{R}^{[H] \times S \times A \times S} \; : & \\
& x(h, s, a, s') \geq 0, && \forall (h, s, a, s') \in [H] \times S \times A \times S && (16)\\
& x(h) = 1, && \forall h \in [H], && (17)\\
& x(1, s) = \mathbb{I}\{s = s_1\}, && \forall s \in S. && (18)\\
& x(h+1, s) = \sum_{(s', a) \in S \times A} x(h, s', a, s), && \forall (h, s) \in [H-1] \times S. && (19)\\
& \|\widetilde{P}^{(x)}(\cdot \mid s, a, h) - \widehat{P}_i(\cdot \mid s, a, h)\|_1 \leq \frac{\epsilon_i(s, a, h)}{H}, && \forall (h, s, a) \in [H] \times S \times A \Big\}. && (20)
\end{aligned}$$

Eqs. (16) to (19) simply define an occupancy measure, while Eq. (20) requires that the next-state distribution associated with the occupancy measure, $\widetilde{P}^{(x)}$ (Eq. (2)), are close to the empirical next-state distribution (Eq. (11)). The following lemma states that $\mathcal{S}_i$ contains all occupancy measures associated with the true model dynamics P.

Lemma 13. *Suppose that Eq. (15) holds, and let x^π, P be an occupancy measure corresponding to some policy π and the true model dynamics. Then $x \in \mathcal{S}_i$.*

Proof. By definition of an occupancy measure, we have that Eqs. (16) to (19) hold, and that

$$\widetilde{P}^{(x)}(s' \mid s, a, h) = P(s' \mid s, a, h), \qquad \forall (h, s, a, s') \in [H] \times S \times A \times S,$$

where P is the true dynamics. Thus Eq. (20) holds by Lemma 11 and our claim follows. ■

The next lemma bounds the difference in norm between any two occupancy measures in $\mathcal{S}_i$ that correspond to the same policy (proof is lone and deferred to Appendix B.4 below).

Lemma 14. *Suppose that Eq. (15) holds, and let $x \in \mathcal{S}_i$. Let x' be the occupancy measure of $\pi^{(x)}$ under the true model dynamics P. Then, $\|x - x'\|_1 \le \min\{\epsilon_i \cdot x, \epsilon_i \cdot x'\}$.*

Lastly, note that according to the DLB setting, one has to know an a-priori upper bound on $\sum_{t=1}^{T} (\hat{z}_t \cdot \epsilon_t)^2$. The bound is given by the following lemma.

Lemma 15. *Let $k_1, k_2, \ldots$ be such that k_i is the initial episode for epoch i. Then, for every epoch i,*

$$\sum_{k=k_i}^{k_{i+1}-1} (\epsilon_i \cdot \hat{z}_k)^2 \le 25H^4 |S||A| \left(|S| + \log \frac{H|S||A|K}{\delta} \right).$$

Proof. We have that $\hat{z}_k(s,a,h)$ is the empirical trajectory of episode $k \in [k_i, k_{i+1} - 1]$. Therefore, $n_i(s,a,h) = \sum_{k=k_i}^{k_{i+1}-1} \hat{z}_k(s,a,h)$. Since during epoch i we have $n_i(s,a,h) \le \max\{N_i(s,a,h), 1\}$, at the end of epoch i we have $n_i(s,a,h) \le \max\{N_i(s,a,h), 1\} + 1$, since the last trajectory might add 1. Also note that $\hat{z}_t$ is a vector whose elements are zero or one with exactly H non-zeros. Therefore,

$$\begin{aligned}
\sum_{k=k_i}^{k_{i+1}-1} (\epsilon_i \cdot \hat{z}_k)^2 &\le \sum_{k=k_i}^{k_{i+1}-1} H \sum_{\substack{(s,a,h) \\ \in S \times A \times [H]}} \hat{z}_k(s,a,h) \cdot \epsilon_i(s,a,h)^2 \\
&= \sum_{\substack{(s,a,h) \\ \in S \times A \times [H]}} n_i(s,a,h) \cdot 25H^3 \cdot \frac{|S| + \log(H|S||A|K/\delta)}{\max\{N_i(s,a,h), 1\}} \\
&\le \sum_{\substack{(s,a,h) \\ \in S \times A \times [H]}} \max\{N_i(s,a,h), 1\} \cdot 25H^3 \cdot \frac{|S| + \log(H|S||A|K/\delta)}{\max\{N_i(s,a,h), 1\}} \\
&\le \sum_{\substack{(s,a,h) \\ \in S \times A \times [H]}} 25H^3 \left(|S| + \log \frac{H|S||A|K}{\delta} \right) \\
&\le 25H^4 |S||A| \left(|S| + \log \frac{H|S||A|K}{\delta} \right),
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz, the second is replacing the sum over $\hat{z}_k(s,a,h)$ by $n_i(s,a,h)$, and the third uses the inequality $n_i(s,a,h) \le \max\{N_i(s,a,h), 1\}$ from definition of Algorithm 3. ■

We now prove the main theorem.

Proof of Theorem 3. We run the algorithm of Theorem 4 on $\mathcal{S}_i$ in epoch i, for every i, resetting the algorithm between epochs. Theorem 4 bounds the expected regret in each epoch, whereas Lemmas 11, 13 and 14 imply that the DLB setting holds in each epoch with high probability.

To avoid having to deal with probabilistic dependencies, we only bound the expected regret. To do so, we can assume that there are exactly $2H|S||A| \log K$ epochs (by adding epochs with zero episodes), and that each epoch is run for exactly K episodes (by padding with zero losses and the remaining episodes).

The analysis proceeds as follows. We set $\delta = 1/(HK)$, $\beta = 5H\sqrt{|S| + \log(H|S||A|K/\delta)}$, $B = \beta^2|S||A|H^2$, and $d = |S|^2|A|H$. Recall that Eq. (15) holds with probability at least $1-\delta$, and consider some epoch i. When Eq. (15) holds, we have $x \in \mathcal{S}_i$ by Lemma 13 as well as that $\|y_k - z_k\|_1 \le \min\{y_k \cdot \epsilon_i, z_k \cdot \epsilon_i\}$ for all episodes k during the epoch by Lemma 14. Moreover, we have that $\|\epsilon_i\|_\infty \le \beta$ and that $\sum_{k=k_i}^{k_{i+1}-1}(\epsilon_i \cdot \hat{z}_k)^2 \le B$ (Lemma 15). Thus, conditioned on that Eq. (15) holds up to epoch i (which depends only on randomness prior to epoch i), the algorithm of Theorem 4 obtains an expected regret bound in epoch i of

$$\mathrm{poly}(d, \beta, H, B)\, O(\sqrt{K}) = \mathrm{poly}(H, |S|, |A|)\, \widetilde{O}(\sqrt{K}).$$

If, on the other hand, Eq. (15) does not hold, the regret in epoch i is at most HK which happens with probability at most δ. Therefore, by the choice of δ, we obtain that the expected regret in epoch i is at most $\mathrm{poly}(|S|, |A|, H)\, \widetilde{O}(\sqrt{K})$, where now the expectation is taken with respect to any randomness prior to the start of the epoch as well as during the epoch.

We conclude the proof by summing over all epochs, which yields the final regret bound. ■

B.4. Proof of Lemma 14

Proof. To simplify notation, we write

$$x(h,s,a) = \sum_{s' \in S} x(h,s,a,s'), \quad \text{and} \quad x(h,s) = \sum_{a \in A} x(h,s,a).$$

Define $\widetilde{P}(s' \mid s,a,h) = \frac{x(h,s,a,s')}{x(h,s,a)}$ and recall that $\pi_h(a \mid s) = \frac{x(h,s,a)}{x(h,s)}$. For $h = 1$, we have

$$\begin{aligned}
\sum_{\substack{(s,a,s') \\ \in S\times A\times S}} \left|x(1,s,a,s') - x'(1,s,a,s')\right| &= \sum_{\substack{(s,a,s') \\ \in S\times A\times S}} \left|x(1,s)\widetilde{P}(s' \mid s,a,1) - x'(1,s)P(s' \mid s,a,1)\right|\pi_1(a \mid s) \\
&= \sum_{\substack{(a,s') \\ \in A\times S}} \left|\widetilde{P}(s' \mid s_1,a,1) - P(s' \mid s_1,a,1)\right|\pi_1(a \mid s_1) \quad \text{(Eq. (18))} \\
&\le \sum_{a \in A} \frac{\epsilon_i(s_1,a,1)}{H}\pi_1(a \mid s_1) \quad \text{(Eq. (20))} \\
&\le \sum_{\substack{(s,a) \\ \in S\times A}} \frac{\epsilon_i(s,a,1)}{H} x'(1,s,a).
\end{aligned}$$

Next, for $h > 1$,

$$
\begin{aligned}
&\sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| x(h,s,a,s') - x'(h,s,a,s') \right| \\
&= \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| x(h,s)\widetilde{P}(s' \mid s,a,h) - x'(h,s)P(s' \mid s,a,h) \right| \cdot \pi_h(a \mid s) \\
&= \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| \sum_{\substack{(a'',s'')\\ \in A\times S}} \Big(x(h-1,s'',a'',s)\widetilde{P}(s' \mid s,a,h) - x'(h-1,s'',a'',s)P(s' \mid s,a,h) \Big) \right| \pi_h(a \mid s) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Eq. (19))} \\
&\le \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| \sum_{\substack{(a'',s'')\\ \in A\times S}} \big(x(h-1,s'',a'',s) - x'(h-1,s'',a'',s) \big) \right| \cdot \widetilde{P}(s' \mid s,a,h) \cdot \pi_h(a \mid s) \\
&\qquad + \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| \sum_{\substack{(a'',s'')\\ \in A\times S}} x'(h-1,s'',a'',s)\Big(\widetilde{P}(s' \mid s,a,h) - P(s' \mid s,a,h) \Big) \right| \cdot \pi_h(a \mid s) \\
&= \sum_{s\in S} \left| \sum_{\substack{(a'',s'')\\ \in A\times S}} \big(x(h-1,s'',a'',s) - x'(h-1,s'',a'',s) \big) \right| \\
&\qquad + \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| \sum_{\substack{(a'',s'')\\ \in A\times S}} x'(h-1,s'',a'',s)\Big(\widetilde{P}(s' \mid s,a,h) - P(s' \mid s,a,h) \Big) \right| \cdot \pi_h(a \mid s) \\
&\le \sum_{\substack{(s,a'',s'')\\ \in S\times A\times S}} \left| x(h-1,s'',a'',s) - x'(h-1,s'',a'',s) \right| \\
&\qquad + \sum_{\substack{(s,a,s',a'',s'')\\ \in S\times A\times S\times A\times S}} x'(h-1,s'',a'',s) \left| \widetilde{P}(s' \mid s,a,h) - P(s' \mid s,a,h) \right| \cdot \pi_h(a \mid s) \\
&\le \sum_{\substack{(s,a'',s'')\\ \in S\times A\times S}} \left| x(h-1,s'',a'',s) - x'(h-1,s'',a'',s) \right| \\
&\qquad + \sum_{\substack{(s,a,a'',s'')\\ \in S\times A\times A\times S}} x'(h-1,s'',a'',s) \cdot \frac{\epsilon_i(h,s,a)}{H} \cdot \pi_h(a \mid s) \qquad\qquad\qquad\qquad\qquad\qquad \text{(Eq. (20))} \\
&= \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| x(h-1,s,a,s') - x'(h-1,s,a,s') \right| + \sum_{\substack{(s,a)\\ \in S\times A}} x'(h,s) \cdot \frac{\epsilon_i(h,s,a)}{H} \cdot \pi_h(a \mid s) \quad \text{(Eq. (19))} \\
&= \sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left| x(h-1,s,a,s') - x'(h-1,s,a,s') \right| + \sum_{\substack{(s,a)\\ \in S\times A}} x'(h,s,a) \cdot \frac{\epsilon_i(h,s,a)}{H}.
\end{aligned}
$$

Applying this argument recursively, we obtain

$$\sum_{\substack{(s,a,s')\\ \in S\times A\times S}} \left|x(h,s,a,s') - x'(h,s,a,s')\right| \le \frac{1}{H} \sum_{\substack{(h,s,a)\\ \in[H]\times S\times A}} x'(h,s,a)\cdot\epsilon_i(s,a,h) = \frac{x'\cdot\epsilon_i}{H},$$

so that

$$\|x-x'\|_1 = \sum_{\substack{(h,s,a,s')\\ \in[H]\times S\times A\times S}} \left|x(h,s,a,s') - x'(h,s,a,s')\right| \le x'\cdot\epsilon_i.$$

A symmetric argument also provides $\|x-x'\|_1 \le x\cdot\epsilon_i$. ■

B.5. Proof of Theorem 5

In this section we prove:

Theorem 5 (restated). *Consider Algorithm 1 with $\eta = (2H\beta d)^{-1}\sqrt{\log|\mathcal{S}|/T}$ and $\gamma = 2H^2(H+\beta\sqrt{d})\eta/\lambda$. Then, given that $B \ge \sum_{t=1}^T(\hat{z}_t\cdot\epsilon_t)^2$ (almost surely), we have for any $y^\star\in\mathcal{S}$:*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t\cdot(\hat{z}_t - y^\star)\right] \le \left(4H\beta d + \frac{H^3}{\beta\lambda d} + \frac{H^2}{\lambda\sqrt{d}}\right)\sqrt{T\log|\mathcal{S}|} + 10\beta d\sqrt{BT},$$

provided that $\beta\ge 1$ and $T \ge (4H^2(H+\beta\sqrt{d})^2\log|\mathcal{S}|)/(\lambda^2\beta^2d^2)$.

The proof uses the following series of lemmas. The following lemma argues that the regret of Algorithm 1 is bounded by the regret of the multiplicative weights updates, plus an additive error term that scales with the perturbations ϵ_t.

Lemma 16. *Assume $\gamma\le\frac{1}{2}$. For all $y^\star\in\mathcal{S}$ it holds that*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t\cdot(\hat{z}_t - y^\star)\right] \le \mathbb{E}\left[\sum_{t=1}^T\sum_{y\in\mathcal{S}} p_t(y)\big(\tilde{\ell}_t(y) - \tilde{\ell}_t(y^\star)\big)\right] + \gamma HT + 5d\,\mathbb{E}\left[\sum_{t=1}^T \|\epsilon_t\|_{M_t}\right].$$

Proof. We prove that $\mathbb{E}_t[\sum_{y\in\mathcal{S}} p_t(y)\tilde{\ell}_t(y)] \ge \ell_t\cdot\hat{z}_t - 3d\,\|\epsilon_t\|_{M_t}$ which, together with Lemma 6, will imply the lemma by taking expectation and summing over $t=1,\ldots,T$. To see this, observe that by Eq. (7), for all $y\in\mathcal{S}$ one also has $\mathbb{E}_t[\tilde{\ell}_t(y)] \ge \ell_t\cdot y - 2\sqrt{d}\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}$, thus

$$\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} p_t(y)\tilde{\ell}_t(y)\right] \ge \ell_t\cdot\sum_{y\in\mathcal{S}} p_t(y)\,y - 2\sqrt{d}\,\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} p_t(y)\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}\right].$$

Now, $q_t = (1-\gamma)p_t + \gamma\mu$ together with $\gamma\le\frac{1}{2}$ implies $q_t - \gamma\mu \le p_t \le 2q_t$. Therefore, (defining $x_t = \sum_{y\in\mathcal{S}} q_t(y)\cdot y$)

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} p_t(y)\tilde{\ell}_t(y)\right] &\ge \ell_t\cdot x_t - \gamma\sum_{y\in\mathcal{S}}\mu(y)\,\ell_t\cdot y - 4\sqrt{d}\,\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} q_t(y)\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}\right]\\
&\ge \ell_t\cdot x_t - \gamma H - 4\sqrt{d}\,\|\epsilon_t\|_{M_t}\mathbb{E}_t\left[\|y_t\|_{M_t^{-1}}\right]\\
&\ge \ell_t\cdot x_t - \gamma H - 4d\,\|\epsilon_t\|_{M_t},
\end{aligned}$$

where the final inequality used $\mathbb{E}_t[\|y_t\|_{M_t^{-1}}] \le \sqrt{\mathbb{E}_t[y_t^\mathsf{T} M_t^{-1} y_t]} = \sqrt{d}$. Finally, observe that $(\mathbb{E}_t|y_t \cdot \epsilon_t|)^2 \le \mathbb{E}_t[(y_t \cdot \epsilon_t)^2] = \epsilon_t^\mathsf{T} \mathbb{E}_t[y_t y_t^\mathsf{T}]\epsilon_t = \|\epsilon_t\|_{M_t}^2$, so

$$\begin{aligned}\ell_t \cdot x_t &= \mathbb{E}_t[\ell_t \cdot \hat{z}_t] + \mathbb{E}_t[\ell_t \cdot (y_t - z_t)] \\ &\ge \mathbb{E}_t[\ell_t \cdot \hat{z}_t] - \mathbb{E}_t|y_t \cdot \epsilon_t| \\ &\ge \mathbb{E}_t[\ell_t \cdot \hat{z}_t] - \|\epsilon_t\|_{M_t}.\end{aligned}$$

Thus we have

$$\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} p_t(y)\tilde{\ell}_t(y)\right] \ge \mathbb{E}_t[\ell_t \cdot \hat{z}_t] - \gamma H - (4d+1)\|\epsilon_t\|_{M_t} \ge \mathbb{E}_t[\ell_t \cdot \hat{z}_t] - \gamma H - 5d\|\epsilon_t\|_{M_t}.$$

This concludes the proof. ■

Next, we apply a standard second-order regret bound for the multiplicative weights method to obtain the following:

Lemma 17. *Provided that $\gamma \ge 2H^2 \max\{H, \beta\sqrt{d}\}\eta/\lambda$, the following regret bound holds for any $y^\star \in \mathcal{S}$:*

$$\sum_{t=1}^T \sum_{y\in\mathcal{S}} p_t(y)\big(\tilde{\ell}_t(y) - \tilde{\ell}_t(y^\star)\big) \le \frac{\log|\mathcal{S}|}{\eta} + \eta \sum_{y\in\mathcal{S}} p_t(y)(\tilde{\ell}_t(y))^2.$$

Proof. The claim would follow directly from the classical second-order bound for multiplicative weights (e.g., Cesa-Bianchi et al., 2007; Dani et al., 2008) once we establish that $|\tilde{\ell}_t(y)| \le 1/\eta$ for all t and $y \in \mathcal{S}$. Indeed, for all t and y we have

$$\begin{aligned}|\tilde{\ell}_t(y)| &= |(\ell_t \cdot \hat{z}_t) y^\mathsf{T} M_t^{-1} y_t - \sqrt{d}\,\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}| \\ &\le |\ell_t \cdot \hat{z}_t| \cdot |y^\mathsf{T} M_t^{-1} y_t| + \sqrt{d}\,\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t}.\end{aligned}$$

Recall that $|\ell_t \cdot \hat{z}_t| \le \|\ell_t\|_\infty \|\hat{z}_t\|_1 \le H$, $\|y\| \le H$ and $\|y_t\| \le H$ (see Section 3.2). Further, $\|\epsilon_t\|_{M_t}^2 = \mathbb{E}_t[(y_t \cdot \epsilon_t)^2] \le (\beta H)^2$. Hence, we obtain that $|\tilde{\ell}_t(y)| \le (H^3 + \beta H^2\sqrt{d})\|M_t^{-1}\|$. To conclude, recall that $M_t \ge \gamma\lambda I$ thanks to the added exploration, so $\|M_t^{-1}\| \le 1/(\lambda\gamma)$. Substituting this in the right-hand side and using the assumption that $\gamma \ge 2H^2 \max\{H, \beta\sqrt{d}\}\eta/\lambda$, the desired bound on $|\tilde{\ell}_t(y)|$ follows. ■

Finally, we establish a bound on the second-order variance term.

Lemma 18. *Assume $\beta^2 d \ge 1$. It holds that*

$$\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} q_t(y)\tilde{\ell}_t(y)^2\right] \le (2H\beta d)^2.$$

Proof. Using the inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\mathbb{E}_t[\tilde{\ell}_t(y)^2] = \mathbb{E}_t[(\hat{\ell}_t \cdot y - \sqrt{d}\,\|y\|_{M_t^{-1}}\|\epsilon_t\|_{M_t})^2] \le 2\mathbb{E}_t[(\hat{\ell}_t \cdot y)^2] + 2d\|y\|_{M_t^{-1}}^2\|\epsilon_t\|_{M_t}^2.$$

Now, for the first term we have

$$\mathbb{E}_t[(\hat{\ell}_t \cdot y)^2] = \mathbb{E}_t\big[(\ell_t \cdot \hat{z}_t)^2\, y^\top M_t^{-1} y_t\, y_t^\top M_t^{-1} y\big] \le H^2\, y^\top M_t^{-1}\mathbb{E}_t[y_t y_t^\top] M_t^{-1} y = H^2 y^\top M_t^{-1} y = H^2 \|y\|^2_{M_t^{-1}}.$$

For the second term, notice that

$$\|\epsilon_t\|^2_{M_t} = \mathbb{E}_t[(y_t \cdot \epsilon_t)^2] \le (\beta H)^2.$$

Hence $\mathbb{E}_t[\tilde{\ell}_t(y)^2] \le 2H^2(1+d\beta^2)\|y\|^2_{M_t^{-1}} \le 4H^2\beta^2 d\|y\|^2_{M_t^{-1}}$, thus we can bound

$$\mathbb{E}_t\left[\sum_{y\in\mathcal{S}} q_t(y)(\tilde{\ell}_t(y))^2\right] \le 4H^2\beta^2 d \sum_{y\in\mathcal{S}} q_t(y)\|y\|^2_{M_t^{-1}}.$$

To conclude, observe that

$$\sum_{y\in\mathcal{S}} q_t(y)\|y\|^2_{M_t^{-1}} = \mathrm{Tr}\left(M_t^{-1}\sum_{y\in\mathcal{S}} q_t(y) y y^\top\right) = \mathrm{Tr}(M_t^{-1}M_t) = d. \qquad \blacksquare$$

We can now prove Theorem 5.

Proof. Combining Lemmas 16 and 17 and using Lemma 18, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t \cdot (\hat{z}_t - y^\star)\right] \le \frac{\log|\mathcal{S}|}{\eta} + 4(H\beta d)^2\eta T + \gamma H T + 5d\,\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|_{M_t}\right]. \tag{21}$$

To bound the final term, we use two applications of Jensen's inequality,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|_{M_t}\right] \le \sqrt{T\sum_{t=1}^{T}\mathbb{E}\|\epsilon_t\|^2_{M_t}} = \sqrt{T\sum_{t=1}^{T}\mathbb{E}[(y_t \cdot \epsilon_t)^2]}.$$

Further, observe that since $\|\epsilon_t\|_\infty \le \beta$ and $\|y_t - z_t\|_1 \le |z_t \cdot \epsilon_t|$, we have

$$\begin{aligned}\mathbb{E}[(y_t \cdot \epsilon_t)^2] &\le 2\mathbb{E}[(z_t \cdot \epsilon_t)^2] + 2\mathbb{E}[((y_t - z_t)\cdot\epsilon_t)^2] \\ &\le 2\mathbb{E}[(z_t \cdot \epsilon_t)^2] + 2\mathbb{E}[\|y_t - z_t\|_1^2\|\epsilon_t\|_\infty^2] \\ &\le 2(1+\beta^2)\mathbb{E}[(z_t \cdot \epsilon_t)^2],\end{aligned}$$

and by Jensen's inequality we obtain

$$\mathbb{E}[(z_t \cdot \epsilon_t)^2] = \mathbb{E}[(\mathbb{E}_t[\hat{z}_t \mid z_t] \cdot \epsilon_t)^2] \le \mathbb{E}[(\hat{z}_t \cdot \epsilon_t)^2].$$

Thus,

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\epsilon_t\|_{M_t}\right] \le \sqrt{T \cdot 4\beta^2\sum_{t=1}^{T}\mathbb{E}[(\hat{z}_t \cdot \epsilon_t)^2]} \le 2\beta\sqrt{BT}.$$

Plugging this into Eq. (21), and using the choices of η and γ, the statement follows. $\blacksquare$

B.6. Proof of Theorem 7

Here we prove:

Theorem 7 (restated). *Consider Algorithm 2 with*

$$\eta_0 = \min\left\{\sqrt{\frac{\vartheta \log(HT)}{d^2H^2T}}, \frac{1}{4d\sqrt{BT}}\right\}.$$

Then, for any $y^\star \in \mathcal{S}$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T}(\hat{z}_t - y^\star)\cdot \ell_t\right] = O\Big(d\beta H\vartheta \log(HT) + d\vartheta\sqrt{BT}\,\log(HT) + dH\sqrt{\vartheta T \log(HT)}\Big),$$

provided that $B \geq \max\{\sum_{t=1}^{T}(\hat{z}_t\cdot\epsilon_t)^2, H\}$ (almost surely).

To prove the theorem, we first prove a few lemmas that will aid in the main proof. Our first lemma shows some necessary technical results, the first of which is that indeed $y_t \in \mathcal{S}$ for all $t = 1,\ldots,T$.

Lemma 19. *For all $t = 1,\ldots,T$: $\;y_t \in \mathcal{S}$; $\;\|\nabla^2 R(x_t)^{1/2}u_t\|^\star_{x_t} = 1$; and $\;\|\tilde{\ell}_t\|^\star_{x_t} \leq dH$.*

Proof. Since R is a self-concordant barrier function over a compact set $\mathcal{S}$, following Eq. (3), it suffices to show that for all t, $\|y_t - x_t\|_{x_t} \leq 1$, and indeed

$$\|y_t - x_t\|^2_{x_t} = \|\nabla^2 R(x_t)^{-1/2}u_t\|^2_{x_t} = u_t^\top \nabla^2 R(x_t)^{-1/2}\nabla^2 R(x_t)\nabla^2 R(x_t)^{-1/2}u_t = 1.$$

Similarly,

$$\left(\left\|\nabla^2 R(x_t)^{1/2}u_t\right\|^\star_{x_t}\right)^2 = u_t^\top \nabla^2 R(x_t)^{1/2}\nabla^2 R(x_t)^{-1}\nabla^2 R(x_t)^{1/2}u_t = 1,$$

and

$$\|\tilde{\ell}_t\|^\star_{x_t} = d\,|\ell_t\cdot\hat{z}_t|\,\|\nabla^2 R(x_t)^{1/2}u_t\|^\star_{x_t} \leq dH. \qquad \blacksquare$$

Lemma 20. *Suppose $\eta_0 \leq 1/4d\sqrt{BT}$, then $\eta_0 \leq \eta_t \leq 2\eta_0, \quad \forall t = 1,\ldots,T$.*

Proof. $\eta_0 \leq \eta_t$ holds by definition. The other direction is because

$$\eta_t^{-1} = \eta_0^{-1} - 2d\sum_{s=1}^{t}|\hat{z}_t\cdot\epsilon_t| \geq \eta_0^{-1} - 2d\sqrt{T\cdot\sum_{s=1}^{t}(\hat{z}_t\cdot\epsilon_t)^2} \geq \eta_0^{-1} - 2d\sqrt{BT} \geq \frac{1}{2}\eta_0^{-1}. \qquad \blacksquare$$

Finally, we combine the lemmas above with the guarantee of OMD to yield the main theorem.

Proof of Theorem 7. Observe the three summands of Eq. (8). For the first summand, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(z_t - x_t)\cdot\ell_t\right] &= \mathbb{E}\left[\sum_{t=1}^{T}(z_t - y_t)\cdot\ell_t\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}|z_t\cdot\epsilon_t|\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}|\hat{z}_t\cdot\epsilon_t|\right] \leq \sqrt{T\,\mathbb{E}\left[\sum_{t=1}^{T}(\hat{z}_t\cdot\epsilon_t)^2\right]} \leq \sqrt{BT},\end{aligned}$$

where the first inequality uses that $\|\ell_t\|_\infty \leq 1$ and the assumption that $\|z_t - y_t\|_1 \leq |z_t\cdot\epsilon_t|$, the second inequality is by Jensen's inequality, and the third inequality is due to Cauchy-Schwartz. For the last inequality we recall that $\sum_{t=1}^T (\hat{z}_t \cdot \epsilon_t)^2 \le B$ by the assumptions of the DLB setting (see Section 3.2).

For the second summand in Eq. (8), since $B \ge H$, by our choice of η_0, and by Lemma 20 we have $\eta_t \le 2\eta_0 \le 1/2dH$, so $\eta_t \|\tilde{\ell}_t\|_{x_t}^\star \le \frac{1}{2}$ by Lemma 19. We can therefore apply Lemma 2 to get

$$\mathbb{E}\left[\sum_{t=1}^T (x_t - y_\gamma^\star) \cdot \tilde{\ell}_t\right] \le \mathbb{E}\left[\underbrace{\frac{1}{\eta_1} B_R(y_\gamma^\star \,\|\, x_1)}_{(1)} - \underbrace{\sum_{t=2}^T \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) B_R(y_\gamma^\star \,\|\, x_t)}_{(2)} + \underbrace{\sum_{t=1}^T \eta_t (\|\tilde{\ell}_t\|_{x_t}^\star)^2}_{(3)}\right].$$

We now bound each of the three terms (1), (2), and (3). We have $(1) \le \eta_0^{-1} \vartheta \log(\gamma^{-1})$ by Lemma 1 and as $\eta_1 \ge \eta_0$ (Lemma 20). For term (2), we have

$$\begin{aligned}
(2) &= 2d \sum_{t=2}^T |\hat{z}_t \cdot \epsilon_t| B_R(y_\gamma^\star \,\|\, x_t) \\
&= 2d \sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| B_R(y_\gamma^\star \,\|\, x_t) - 2d \underbrace{|\hat{z}_1 \cdot \epsilon_1|}_{\le \|\hat{z}_1\|_1 \|\epsilon_1\|_\infty \le H\beta} B_R(y_\gamma^\star \,\|\, x_1) \\
&\ge 2d \sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| \left(\frac{1}{2} \|x_t - y_\gamma^\star\|_{x_t} - 1\right) - 2d\beta H \cdot \vartheta \log \frac{1}{\gamma} && \text{(Lemmas 1 and 9)} \\
&= d \sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| \|x_t - y_\gamma^\star\|_{x_t} - 2d \sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| - 2d\beta H \cdot \vartheta \log \frac{1}{\gamma} \\
&\ge d \sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| \|x_t - y_\gamma^\star\|_{x_t} - 2d\sqrt{BT} - 2d\beta H \cdot \vartheta \log \frac{1}{\gamma},
\end{aligned}$$

where the last inequality is since $\sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| \le \sqrt{T \sum_{t=1}^T (\hat{z}_t \cdot \epsilon_t)^2}$ by Cauchy-Schwartz and as $\sum_{t=1}^T (\hat{z}_t \cdot \epsilon_t)^2 \le B$ by assumption. We lastly employ Lemma 19 and that $\eta_t \le 2\eta_0$ by Lemma 20 to bound $(3) \le 2\eta_0 d^2 H^2 T$. All in all, this obtains us Eq. (10).

We sum Eq. (9) over all t and take expectation. Together with Eq. (10) this replaces the perceived losses, $\tilde{\ell}_t$, by the real losses, ℓ_t. The terms $d\,\mathbb{E}[\sum_{t=1}^T |\hat{z}_t \cdot \epsilon_t| \, \|x_t - y^\star\|_{x_t}]$ in Eq. (9) and in Eq. (10) cancel out, and we get

$$\mathbb{E}\left[\sum_{t=1}^T (x_t - y_\gamma^\star) \cdot \ell_t\right] \le \left(\frac{1}{\eta_0} + 2d\beta H\right) \vartheta \log \frac{1}{\gamma} + 2d\sqrt{BT} + 2\eta_0 d^2 H^2 T.$$

Finally, for the third summand in Eq. (8), we have

$$\sum_{t=1}^T (y_\gamma^\star - y^\star) \cdot \ell_t = \gamma \sum_{t=1}^T (x_1 - y^\star) \cdot \ell_t \le 2\gamma H T.$$

Combining the bounds on all three summands and setting γ, η_0 as in the theorem's statement yields the final regret bound. ■

B.7. Proof of Corollary 10

Lemma 21. *For option 1 in Section 5, the following holds:*

(i) $\|y_k - x_k\|_{x_k} = 1$.

(ii) $\left(\|WW^\top \nabla^2 R(x_t)^{1/2} W u_t\|^\star_{x_t}\right) = 1$.

(iii) $\|\tilde{\ell}_t\|^\star_{x_t} \leq pH$.

Proof. We have,

$$\begin{aligned}\|y_k - x_k\|^2_{x_k} &= (y_k - x_k)^\top \nabla^2 R(x_k)(y_k - x_k)\\ &= u_k^\top W^\top \nabla^2 R(x_k)^{-1/2} W W^\top \nabla^2 R(x_k) W W^\top \nabla^2 R(x_k)^{-1/2} W u_k\\ &= u_k^\top (W^\top \nabla^2 R(x_k) W)^{-1/2} (W^\top \nabla^2 R(x_k) W)(W^\top \nabla^2 R(x_k) W)^{-1/2} u_k = 1.\end{aligned}$$

(*W* is orthogonal)

In addition,

$$\begin{aligned}\left(\|WW^\top \nabla^2 R(x_t)^{1/2} W u_t\|^\star_{x_t}\right)^2 &= u_t^\top W^\top \nabla^2 R(x_t)^{1/2} W W^\top \nabla^2 R(x_t)^{-1} W W^\top \nabla^2 R(x_t)^{1/2} W u_t\\ &= u_t^\top (W^\top \nabla^2 R(x_t) W)^{1/2} (W^\top \nabla^2 R(x_t) W)^{-1} (W^\top \nabla^2 R(x_t) W)^{1/2} u_t\\ &= u_t^\top u_t = 1,\end{aligned}$$

where the second equality is as W is orthogonal.

Finally,

$$\|\tilde{\ell}_t\|^\star_{x_t} = p\,|\ell_t \cdot \hat{z}_t|\, \|WW^\top \nabla^2 R(x_t)^{1/2} W u_t\|^\star_{x_t} \leq pH.$$

■

Proof of Corollary 10. The proof follows that of Theorem 3 but uses the the regret bound obtained from Theorem 7 with ϑ-self-concordant barrier for $\vartheta = O(|S|^2|A|H)$. To prove Theorem 7, we replace the results of Lemma 19 by these of Lemma 21. The computational efficiency of the algorithm stems from the discussion in Section 5. ■

Appendix C. Self-concordant Barriers: definitions and basic properties

For a k-array tensor $U \in \mathbb{R}^{d \times k}$, we define

$$U[h_1, \ldots, h_k] = \sum_{i_1,\ldots,i_k \in [d]} U(i_1, \ldots, i_k) \prod_{j=1}^{k} h_j(i_j).$$

For $k = 2$ we have that U is a matrix, h_1 and h_2 are vectors, and $U[h_1, h_2] = h_1^\top U h_2$.

Definition 22. For a convex set $\mathcal{S} \subset \mathbb{R}^n$, a self-concordant function $R : \mathrm{int}(\mathcal{S}) \mapsto \mathbb{R}$ is a C^3-convex function such that

$$\left|D^3 R(x)[h,h,h]\right| \leq 2\left(D^2 R(x)[h,h]\right)^{3/2}.$$

In words: the third derivative of R at x in direction h is upper bounded by a constant times the second derivative of R at x in direction h, raised to the 3/2 power.

Definition 23. A self-concordant function R is a ϑ-self-concordant barrier if

$$\left|DR(x)[h]\right| \leq \vartheta^{1/2}\left(D^2 R(x)[h,h]\right)^{1/2}.$$

We have the following upper bound on the difference a ϑ-self-concordant barrier R at two points $x, y \in \mathcal{K}$:

$$R(y) - R(x) \le \vartheta \log \frac{1}{1 - \pi_x(y)}, \tag{22}$$

where $\pi_x(y)$ is the Minkowski function of $\mathcal{S}$ w.r.t. x: $\pi_x(y) = \inf\{t > 0 : x + t^{-1}(y - x) \in \mathcal{S}\}$.