Parameter-Free Dynamic Regret for Unconstrained Linear Bandits

Alberto Rumi[*]
Università degli Studi di Milano

Andrew Jacobsen[*]
Università degli Studi di Milano
Politecnico di Milano

Fabio Vitale
CENTAI Institute

Nicolò Cesa-Bianchi
Università degli Studi di Milano
Politecnico di Milano

Abstract

We study dynamic regret minimization in online learning with an oblivious adversary and bandit feedback. In this setting, a learner must minimize the cumulative loss relative to an arbitrary sequence of comparators $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$ in $\mathcal{W} \subseteq \mathbb{R}^d$, but receives only *point-evaluation feedback* on each round. We provide a simple approach to combining the guarantees of several bandit algorithms, allowing us to design algorithms which optimally adapt to the path-length $P_T = \sum_t \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|$ or the number of switches $S_T = \sum_t \mathbb{I}\{\boldsymbol{u}_t \neq \boldsymbol{u}_{t-1}\}$ of an arbitrary comparator sequence. In particular, we provide the *first* algorithms for linear bandits which obtain the optimal regret guarantee of order $\mathcal{O}\left(\sqrt{(1+S_T)T}\right)$ up to poly-logarithmic terms *without prior knowledge of* S_T, resolving a long-standing open problem.

1 Introduction

Online learning is a framework that models sequential decision-making against adversarial environments (Shalev-Shwartz et al., 2012; Orabona, 2019). In this framework, a learner interacts with an environment over time, making decisions and receiving feedback which may be partial or incomplete. The classical way to evaluate the performance of online algorithms is through the notion of *regret*, which compares the learner's cumulative loss to that of the best fixed strategy chosen in hindsight. While this is a natural benchmark, it implicitly assumes that the environment is *stationary*, i.e., that the optimal action does not change over time. However, in many practical scenarios this is far from true: user preferences, market conditions, or system dynamics can evolve unpredictably over time, leading to situations where minimizing regret against a fixed comparator becomes meaningless.

Formally, we consider a bandit optimization setting characterized by a space of actions $\mathcal{W} \subseteq \mathbb{R}^d$ and an oblivious adversary privately selecting a sequence of convex loss functions $\ell_t : \mathcal{W} \to \mathbb{R}$. At each time step $t \in [T]$, the learner chooses $\boldsymbol{w}_t \in \mathcal{W}$ and observes the incurred loss $\ell_t(\boldsymbol{w}_t)$. A common special case we study is the linear bandit setting, in which losses are of the form $\ell_t(\boldsymbol{w}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \rangle$, where $\boldsymbol{\ell}_t \in \mathbb{R}^d$. The learner's performance is evaluated learners according to the expected *dynamic* regret

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)\right] = \mathbb{E}\left[\sum_{t\in[T]} \ell_t(\boldsymbol{w}_t)\right] - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t),$$

[*] Equal contribution

18th European Workshop on Reinforcement Learning (EWRL 2025).

where $(\boldsymbol{w}_t)_{t\in[T]}$ is the sequence of the learner's actions, $(\boldsymbol{u}_t)_{t\in[T]}$ is an arbitrary sequence of comparators, and the expectation is taken with respect to the internal randomness of the algorithm. The standard (static) regret is recovered when $\boldsymbol{u}_1 = \cdots = \boldsymbol{u}_T$.

Notice that the "complexity" of the comparator sequence contributes to characterizing the difficulty of a given problem. If the comparator is allowed to change arbitrarily on each round, there is no hope to achieve low dynamic regret since the comparator sequence can ensure $\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) = 0$ even against completely unpredictable losses. On the other hand, we know that in the special case $\boldsymbol{u}_1 = \ldots = \boldsymbol{u}_T$ we can ensure low regret, since this is just the usual static regret setting. Thus, we are typically interested in algorithms which make dynamic regret guarantees that gracefully *adapt* to some measure of complexity or *variability* of the comparator sequence. The classic way to quantify the variability of the comparator sequence is via its *path-length*

$$P_T = \sum_{t=2}^{T} \left\|\boldsymbol{u}_{t-1} - \boldsymbol{u}_t\right\|_2 .$$

A related classical benchmark is the *switching regret*, which bounds regret with respect to the number of times the comparator sequence changes over time $S_T = \sum_{t=2}^{T} \mathbb{I}[\boldsymbol{u}_t \neq \boldsymbol{u}_{t-1}]$, capturing a coarser notion of variability. We will consider a generalization of this notion in which the comparator at time t is sampled from a distribution $\boldsymbol{p}_t^*$ over $\mathcal{W}$, such that its expectation satisfies $\mathbb{E}_{\boldsymbol{w}\sim\boldsymbol{p}_t^*}[\boldsymbol{w}] = \boldsymbol{u}_t$. This generalizes to situations where the learner is compared against randomized strategies or smooth shifts in behavior. In this case, we define the *distributional path-length* as:

$$P_T^{\triangle} = \sum_{t=2}^{T} \left\|\boldsymbol{p}_t^* - \boldsymbol{p}_{t-1}^*\right\|_1 .$$

While dynamic or switching regret offers a more realistic benchmark in nonstationary settings, it also poses significant new challenges. In particular, achieving regret bounds that scale optimally with the path-length typically requires preliminary knowledge of it, or of at least a bound on it (Agarwal et al., 2017; Marinov & Zimmert, 2021; Luo et al., 2022). This is unrealistic because the comparator sequence can never be directly observed.

Parameter-free methods, which in dynamic regret settings eliminate the need for knowledge of the path length, are relatively well understood in the full-information setting. (Zhang et al., 2018; Cutkosky, 2020; Campolongo & Orabona, 2021; Jacobsen & Cutkosky, 2022, 2023; Zhang et al., 2023; Jacobsen & Cutkosky, 2024). In this work, we focus instead on the bandit setting, where the learner only receives point-evaluation feedback at the end of each round. With bandit feedback, the parameter-free techniques developed for the full information setting are either not directly applicable or result in a suboptimal dependence on the time horizon and path length.

Our contributions. We prove dynamic regret bounds under bandit feedback against an oblivious adversary and without requiring prior knowledge of the path length P_T. We obtain the first dynamic regret bounds of order $\widetilde{\mathcal{O}}\big(\sqrt{(1+P_T^{\triangle})T}\big)$ without prior knowledge of the distributional path length $P_T^{\triangle}$ (which includes the number of switches S_T as a special case), thus resolving a long-standing open problem (Marinov & Zimmert, 2021; Luo et al., 2022). Key to obtaining these results is a technique for combining the guarantees of *comparator-adaptive* base algorithms inspired by a clever result in the full information setting (Cutkosky, 2019), which we adapt to bandit feedback via a sampling trick. This simple trick enables us to easily combine the outputs of several bandit algorithms to achieve the best of their respective dynamic regret guarantees, effectively enabling us to "tune" hyperparameters on-the-fly. Such hyperparameter tuning arguments have been attempted by several prior works using sophisticated mixture-of-experts style arguments (Agarwal et al., 2017; Marinov & Zimmert, 2021; Luo et al., 2022), but have only achieved the optimal $\sqrt{P_T}$ dependence by leveraging *a priori* knowledge of P_T. We anticipate that our approach will also find applications in other settings where bandit-over-bandit ensembling strategies have failed in the past.

Related works. The study of dynamic regret was initiated by Herbster & Warmuth (1998, 2001). In the setting of online convex optimization (OCO) with Lipschitz losses, Zinkevich (2003) showed that online gradient descent achieves a bound of order $\mathcal{O}\big((1+P_T)\sqrt{T}\big)$. Yang et al. (2016) improved the rate to $\mathcal{O}(\sqrt{P_T T})$ by leveraging prior knowledge of P_T. This bound was later shown to be minimax optimal by Zhang et al. (2018), who also provided the first algorithm achieving a matching upper bound without prior knowledge of P_T. Since then, several works have achieved generalizations of $\sqrt{(1+P_T)T}$ rate by improving the T dependence with problem-dependent penalties such as $\sum_t \|\nabla\ell_t(\boldsymbol{w}_t)\|^2$ or $\sum_t \sup_{\boldsymbol{x}} |\ell_t(\boldsymbol{x}) - \ell_{t-1}(\boldsymbol{x})|$, and additionally adapting to $\max_t \|\boldsymbol{u}_t\|$ (Cutkosky, 2020; Campolongo & Orabona, 2021; Jacobsen & Cutkosky, 2023; Zhang et al., 2023). Various improvements in adaptivity can also be obtained under additional assumptions on the losses such as smoothness Mokhtari et al. (2016); Zhao et al. (2020, 2024).

Key to our results is the notion of *comparator adaptive* online learning, where the goal is to design algorithms that adapt to the complexity of the comparator sequence (e.g., its norm or a measure of its variability) without requiring prior knowledge about it. In the static comparator case, this idea has been extensively studied under full-information feedback, where the optimal guarantee $R_T(\boldsymbol{u}) = O(\|\boldsymbol{u}\|\sqrt{T})$, for any $\boldsymbol{u} \in \mathbb{R}^d$, can be obtained up to logarithmic terms (Mcmahan & Streeter, 2012; McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018). Notably, Cutkosky (2019) showed that algorithms making comparator-adaptive guarantees can be easily combined, obtaining regret proportional to the best among them; this observation will be crucial to our approach in Section 2. In both linear and convex bandit settings, comparator-adaptive bounds were studied by van der Hoeven et al. (2020) where they consider static regret and propose a black-box reduction approach, taking inspiration from the full information 1-dimensional reduction of Cutkosky & Orabona (2018).

In the bandit setting, the study of the closely-related notion of switching regret was initiated by Auer (2002), where a bound of $\mathcal{O}(\sqrt{S_T T})$ was obtained with prior knowledge of the number of switches S_T. The optimal regret bound for the non-stationary *stochastic* bandit setting without *a priori* variational knowledge was first obtained in Auer et al. (2018). Interestingly, Marinov & Zimmert (2021) showed the impossibility of obtaining a $\mathcal{O}\left(\sqrt{S_T T}\right)$ bound in adaptive adversary settings, leaving the question open (until now) for the oblivious setting.

In the bandit convex optimization (BCO) literature, Zhao et al. (2021) showed that the classical Bandit Gradient Descent algorithm of Flaxman et al. (2004) can achieve dynamic regret bound of order $\mathcal{O}\big(T^{3/4}(1 + P_T^{1/4})\big)$ when the step-size is optimally tuned using knowledge of P_T. They also proposed a parameter-free version using a bandit-over-bandit ensemble strategy, but this leads to a suboptimal regret of $\mathcal{O}\big(T^{3/4}(1 + P_T^{1/2})\big)$ due to additional variance from the meta-learning layer. Later, Yan et al. (2023) improved the dependence on P_T but picked up an additive $T^{5/6}$ penalty, obtaining $O(P_T^{1/4}T^{3/4} + T^{5/6})$ overall.

More broadly, ensemble and meta-algorithmic strategies have been proposed to adapt to unknown environment parameters in bandits. One notable example is the CORRAL framework introduced by Agarwal et al. (2017); Luo et al. (2022), which maintains a pool of base bandit algorithms to hedge against misspecification. However, to obtain their result, they still need to assume a fixed and known number of switches (path length in our setting). Naive bandit-over-bandit schemes incur significant overhead: for instance, running EXP4 (Auer et al., 2002) as a master over EXP3 bases would yield $\mathcal{O}\left(T^{2/3}\right)$ regret due to the extra exploration needed (Odalric & Munos, 2011; Cheung et al., 2019). While these methods are flexible, they often pay a price in terms of worse regret bounds or higher variance, particularly in dynamic or tuning-free scenarios.

Notation and Assumptions. We assume the action set $\mathcal{W}$ is unconstrained, i.e., $\mathcal{W} = \mathbb{R}^d$. All the following results easily extend to action sets contained in a Euclidean ball.

In the following, let $D_\psi(x, y)$ be the Bregman divergence with respect to the strongly convex regularizer $\psi\colon \mathcal{W} \to \mathbb{R}$. We denote with $\|\cdot\|$ a norm; when the degree is not specified, it refers to the Euclidean norm. The corresponding dual norm of $\boldsymbol{w}$ is denoted as as $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{g}:\|\boldsymbol{g}\|\le 1} \langle \boldsymbol{g}, \boldsymbol{w}\rangle$. Given a norm $\|\cdot\|$ and $\rho \ge 0$, we denote by $\mathcal{B}_\rho := \{x : \|x\| \le \rho\}$ the closed ball of radius ρ, or the unit ball when the radius is not specified. We denote with $\mathcal{M}_1(\mathcal{W})$the set of distributions over $\mathcal{W}$. $\mathcal{O}(\cdot)$ hides constant factors and $\widetilde{\mathcal{O}}(\cdot)$ hides constant and logarithmic factors.

2 Combining guarantees with uniform sampling

Algorithm 1 Uniform Sampling Interface

```
Input: Domain $\mathcal{W} \subseteq \mathbb{R}^d$, base algorithms $\left(\mathcal{A}_n\right)_{n=1}^N$
for $t = 1, \ldots, T$ do
    Get $\boldsymbol{w}_t^{(i)}$ from $\mathcal{A}_i$ for all $i \in [N]$
    Sample $i_t \sim$ Uniform$(N)$ and play $\boldsymbol{w}_t = \boldsymbol{w}_t^{(i_t)}$
    Receive feedback $\phi(\boldsymbol{w}_t, \ell_t)$
    Send $\phi(\boldsymbol{w}_t, \ell_t)\mathbb{I}\{i_t = i\}$ to $\mathcal{A}_i$ for $i \in [N]$
end for
```

Our approach is inspired by a framework for combining guarantees of *comparator-adaptive* online learning algorithms proposed by Cutkosky (2019) for the simpler OCO setting. To illustrate the idea, suppose we have N OCO algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_N$, each guaranteeing regret $R_T^{\mathcal{A}_i}(\mathbf{0}) = \mathcal{O}(1)$— the fundamental feature characterizing *comparator-adaptive* algorithms (McMahan & Abernethy, 2013; McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018). The key insight of Cutkosky (2019) is that we can always obtain the best guarantee among them, $R_T(\boldsymbol{u}) = \mathcal{O}(\min_i R_T^{\mathcal{A}_i}(\boldsymbol{u}))$, by simply adding the iterates together. Indeed, letting $\boldsymbol{w}_t^{(i)}$ denote the output of $\mathcal{A}_i$ on round t, if we play $\boldsymbol{w}_t = \sum_{i \in [N]} \boldsymbol{w}_t^{(i)}$ then for any $j \in [N]$ we have

$$\begin{aligned} R_T(\boldsymbol{u}) = \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle &= \sum_{t=1}^T \left\langle \boldsymbol{g}_t, \boldsymbol{w}_t^{(j)} - \boldsymbol{u} \right\rangle + \sum_{i \neq j} \sum_{t=1}^T \left\langle \boldsymbol{g}_t, \boldsymbol{w}_t^{(i)} \right\rangle \\ &= R_T^{\mathcal{A}_j}(\boldsymbol{u}) + \sum_{i \neq j} R_T^{\mathcal{A}_i}(\mathbf{0}) = \mathcal{O}\left(R_T^{\mathcal{A}_j}(\boldsymbol{u}) + N\right), \end{aligned}$$

where the last step uses the fact that each algorithm guarantees $R_T^{\mathcal{A}_i}(\mathbf{0}) = \mathcal{O}(1)$. Moreover, since this holds for any $j \in [N]$, it must hold for the best among them.

While the above approach is particularly elegant for OCO, it unfortunately will not work under bandit feedback. Loosely speaking, the issue is that when playing $\boldsymbol{w}_t = \sum_{i \in [N]} \boldsymbol{w}_t^{(i)}$, the feedback received will be $\langle \boldsymbol{g}_t, \sum_i \boldsymbol{w}_t^{(i)} \rangle$. As a result, there is no way to precisely assign feedback to any individual learner, as the observed feedback includes contributions from all other learners' decisions. Instead, we make the following simple observation: if on each round we sample one of the algorithms uniformly at random and play *only its iterate* on round t, then *in expectation*, this is equivalent to playing $\sum_i \boldsymbol{w}_t^{(i)}/N$. As a result, we can still apply *nearly* the same iterate-adding argument outlined above, but we can now accurately assign feedback since only one learner's action is played on each round. The main difference from the iterate-adding approach is that we need to rescale the comparator to account for the $1/N$ factor that shows up in $\sum_i \boldsymbol{w}_t^{(i)}/N$.

The procedure described above is summarized in Algorithm 1. It generalizes the strategy to a generic feedback oracle, which returns a feedback signal $\phi(\boldsymbol{w}_t, \ell_t)$ given the loss ℓ_t and decision $\boldsymbol{w}_t$. Specifically, this model captures first-order feedback when $\phi(\boldsymbol{w}_t, \ell_t) = \boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{w}_t)$, bandit feedback when $\phi(\boldsymbol{w}_t, \ell_t) = \ell_t(\boldsymbol{w}_t)$, and full-information feedback when $\phi(\boldsymbol{w}_t, \ell_t) = \ell_t$. As a warm-up to build intuition, the following theorem shows how this simple strategy performs in the first-order feedback setting.

Proposition 2.1. *Let $\mathcal{A}_1, \ldots, \mathcal{A}_N$ be online learning algorithms and let $\boldsymbol{w}_t^{(i)}$ denote the output of $\mathcal{A}_i$ on round t. Suppose that for all i, $\mathcal{A}_i$ guarantees $R_T^{\mathcal{A}_i}(\mathbf{0}) = \sum_{t=1}^T \ell_t(\boldsymbol{w}_t^{(i)}) - \ell_t(\mathbf{0}) \leq G\epsilon$ for any sequence of G-Lipschitz loss functions. Then for any sequence $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$ in $\mathcal{W}$, Algorithm 1 guarantees $\mathbb{E}\left[R_T(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)\right] \leq \min_{n=1,\ldots,N} \mathbb{E}\left[R_T^{\mathcal{A}_n}(N\boldsymbol{u}_1, \ldots, N\boldsymbol{u}_T)\right] + (N-1)G\epsilon$.*

Proof. Denote $\widehat{\boldsymbol{g}}_t^{(n)} = \mathbb{I}\{i_t = n\}\, \boldsymbol{g}_t$ and observe that for any $n \in [N]$ we have

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{w}_t)\right] - \sum_{t=1}^T \ell_t(\boldsymbol{u}_t) \leq \mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{u}_t \rangle\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\boldsymbol{g}}_t^{(n)}, \boldsymbol{w}_t^{(n)} - N\boldsymbol{u}_t \right\rangle + \sum_{n'\neq n}\sum_{t=1}^{T} \left\langle \widehat{\boldsymbol{g}}_t^{(n)}, \boldsymbol{w}_t^{(n')} \right\rangle\right] \\
&\leq \mathbb{E}\left[R_T^{\mathcal{A}_n}(N\boldsymbol{u}_1,\dots,N\boldsymbol{u}_T) + \sum_{n'\neq n} R_T^{\mathcal{A}_{n'}}(\mathbf{0})\right] \\
&\leq \mathbb{E}\left[R_T^{\mathcal{A}_n}(N\boldsymbol{u}_1,\dots,N\boldsymbol{u}_T)\right] + (N-1)G\epsilon \qquad \square
\end{aligned}
$$

This theorem illustrates how uniform sampling of comparator-adaptive algorithms acts as a rudimentary coordination mechanism: by ensuring each learner individually adapts to the complexity of an arbitrary comparator, then on average we can compare our total loss to the total loss of any individual algorithm. As we will see in the following sections, this will enable us to "tune" hyperparameters on-the-fly, allowing us to adapt to the unknown problem parameters such as the path-length P_T without any prior knowledge of it.

Note that, unlike the iterate-adding approach of Cutkosky (2019), this method for combining guarantees ends up increasing the comparator norm by a factor of N, and so in the context of OCO the uniform sampling approach is a worse option than the iterate-adding approach in most situations. Indeed, this trick is primarily of interest in settings where the iterate-adding approach can't be applied. In the following sections, we will see that Algorithm 1 lets us obtain new guarantees for bandit feedback, where the iterate-adding can not be applied in general, at the expense of mild poly-logarithmic penalties.

3 Linear Bandits

After illustrating the intuition behind using uniform sampling to combine regret guarantees, we turn our attention to the linear bandit feedback setting.

Recall from the previous section that the key property that we need to apply the uniform sampling strategy is that the base algorithms $\mathcal{A}_i$ guarantee a comparator-adaptive property, $R_T(\mathbf{0}) = \mathcal{O}(1)$. To obtain guarantees of this form in the bandit setting, we use a reduction introduced by van der Hoeven et al. (2020). The idea is to decompose $\boldsymbol{w}_t$ into a *scale* $v_t \in \mathbb{R}$ and *direction* $\boldsymbol{\beta}_t \in \mathcal{B}$, which will be learned by separate online learning algorithms. In particular, observe that for $M = \max_t \|\boldsymbol{u}_t\|$ we have

$$
\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t - \boldsymbol{u}_t \rangle &= \sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle v_t - \langle \boldsymbol{\ell}_t, \boldsymbol{u}_t \rangle \\
&= \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle (v_t - M)}_{=:R_T^{\mathcal{A}_{\mathcal{V}}}(M)} + M \underbrace{\sum_{t=1}^{T} \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t - \boldsymbol{u}_t/M \rangle}_{=:R_T^{\mathcal{A}_{\mathcal{B}}}(\boldsymbol{u}_1/M,\dots,\boldsymbol{u}_T/M)},
\end{aligned}
$$

so to ensure $R_T(\mathbf{0}) = \mathcal{O}(1)$ it suffices to provide a scale learner which can guarantee $R_T^{\mathcal{A}_{\mathcal{V}}}(0) = \mathcal{O}(1)$ against the losses $v \mapsto \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle v$. Yet this is simply a 1-dimensional *static regret* OLO problem, so we can apply any of the existing comparator-adaptive algorithms as the scale-learner to ensure this property (Orabona & Pál, 2016; Cutkosky & Orabona, 2018; Mhammedi & Koolen, 2020; Jacobsen & Cutkosky, 2022). The following proposition shows that this scale/direction decomposition extends easily to a collection of algorithms combined via uniform sampling.

Proposition 3.1. *Pick N base algorithms as described in Algorithm 2, where the scale learner $\mathcal{A}_{\mathcal{V}}$ is over $\mathbb{R}$ and the direction learner is over $\mathcal{B}_N$. Suppose that for any sequence of G-Lipschitz linear losses $g_1, \dots, g_T$ in $\mathbb{R}$ and for any $\epsilon > 0$, the regret of $\mathcal{A}_{\mathcal{V}}$ satisfies $R_T^{\mathcal{A}_{\mathcal{V}}}(0) = \sum_{t=1}^{T} g_t(v_t^{(n)} - 0) \leq G\epsilon$. Then, for any $n \in [N]$ and any $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T$ in $\mathbb{R}^d$, Algorithm 1 guarantees*

$$
\mathbb{E}\left[R_T(\boldsymbol{u}_1,\dots,\boldsymbol{u}_T)\right] \leq \mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{V}}^{(n)}}(MN)\right] + MN\mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{B}}^{(n)}}\Big(\frac{\boldsymbol{u}_1}{M},\dots,\frac{\boldsymbol{u}_T}{M}\Big)\right] + (N-1)G\epsilon
$$

where $M = \max_t \|\boldsymbol{u}_t\|$.

Algorithm 2 Scale and Direction Decomposition (van der Hoeven et al., 2020)

```
Input: Domain $\mathcal{W}$, scale learner $\mathcal{A}_{\mathcal{V}}$, direction learner $\mathcal{A}_{\mathcal{B}}$
for $t = 1, \dots, T$ do
    Get scale prediction $v_t$ from $\mathcal{A}_{\mathcal{V}}$ and direction prediction $\boldsymbol{\beta}_t$ from $\mathcal{A}_{\mathcal{B}}$
    Play $\boldsymbol{w}_t = v_t \boldsymbol{\beta}_t$
    Observe $\langle \boldsymbol{w}_t, \boldsymbol{\ell}_t \rangle$ and compute $g_t = \langle \boldsymbol{w}_t, \boldsymbol{\ell}_t \rangle / v_t$
    Send $\langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle = \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \rangle / v_t$ to $\mathcal{A}_{\mathcal{B}}$ as the feedback on round $t$
    Send $v \mapsto \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle v$ to $\mathcal{A}_{\mathcal{V}}$ as the $t^{\text{th}}$ loss
end for
```

Algorithm 3 Continuous Exponential Weight for Linear Bandits

```
Input: Action set $\mathcal{W}$, initial exploration distr. $\pi$, learning rate $\eta > 0$, exploration param. $\beta > 0$
Initialize: $\widetilde{p}_1 = \pi$
for $t = 1 : T$ do
    Sample $\boldsymbol{w}_t \sim \widetilde{p}_t$, observe $\langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \rangle$
    Let $\boldsymbol{Q}_t = \mathbb{E}_{\boldsymbol{w} \sim \widetilde{p}_t}[\boldsymbol{w}\boldsymbol{w}^\top]$ and estimate $\widehat{\boldsymbol{\ell}}_t = \boldsymbol{Q}_t^{-1} \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \rangle \boldsymbol{w}_t$
    Update $p_{t+1} = \arg\min_{p \in \mathcal{M}_1(\mathcal{W})} \mathbb{E}_{\boldsymbol{w} \sim p}[\langle \boldsymbol{\ell}_t, \boldsymbol{w} \rangle] + D_\psi(p | \widetilde{p}_t)$
    Update $\widetilde{p}_{t+1} = (1 - \beta) p_{t+1} + \beta \pi$
end for
```

Proof. The result is immediate by applying Proposition 2.1 followed by the scale / direction decomposition:

$$
\begin{aligned}
&\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t - \boldsymbol{u}_t \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, \boldsymbol{\beta}_t^{(n)} \right\rangle v_t^{(n)} - \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, N\boldsymbol{u}_t \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, \boldsymbol{\beta}_t^{(n)} \right\rangle v_t^{(n)} \pm \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, \boldsymbol{\beta}_t^{(n)} \right\rangle NM - \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, N\boldsymbol{u}_t \right\rangle\right] \\
&\leq \underbrace{\mathbb{E}\left[\sum_{t=1}^T \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, \boldsymbol{\beta}_t^{(n)} \right\rangle (v_t - MN)\right]}_{R_T^{\mathcal{A}_{\mathcal{V}}^{(n)}}(MN)} + MN\mathbb{E}\underbrace{\left[\sum_{t=1}^T \left\langle \widehat{\boldsymbol{\ell}}_t^{(n)}, \boldsymbol{\beta}_t^{(n)} - \frac{\boldsymbol{u}_t}{M} \right\rangle\right]}_{R_T^{\mathcal{A}_{\mathcal{B}}^{(n)}}\left(\frac{\boldsymbol{u}_1}{M}, \dots, \frac{\boldsymbol{u}_T}{M}\right)} + (N-1)G\epsilon \\
&= \mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{V}}^{(n)}}(MN)\right] + MN\mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{B}}^{(n)}}\left(\frac{\boldsymbol{u}_1}{M}, \dots, \frac{\boldsymbol{u}_T}{M}\right)\right] + (N-1)G\epsilon \qquad \square
\end{aligned}
$$

As observed above, the key insight of van der Hoeven et al. (2020) is that the feedback received by the scale learner is actually full-information feedback; indeed, a scale learner's loss function $v \mapsto \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle v$ can be precisely recovered from $\langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \rangle = \langle \boldsymbol{\ell}_t, \boldsymbol{\beta}_t \rangle v_t$ by dividing by v_t, so no tricky loss estimation is required for the scale learner. Moreover, observe that the scale learner faces a *static* regret problem, so overall to meet the condition of Proposition 3.1,

it will suffice to apply any comparator-adaptive OCO algorithm for static regret as the scale learner, leading to a guarantee of the form $R_T^{\mathcal{A}_{\mathcal{V}}^{(n)}}(M) = \widetilde{\mathcal{O}}(M\sqrt{T})$. Thus all that remains is to design a bandit algorithm which can handle dynamic regret on a ball of radius N. To achieve this, we use the Continuous Exponential Weights algorithm (Algorithm 3) and derive the following dynamic regret guarantee, proven in Appendix A.

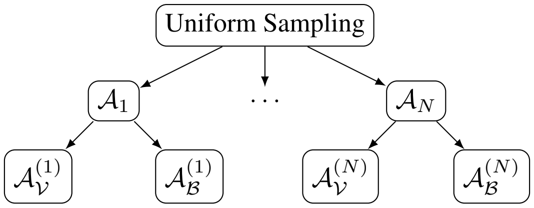
Uniform Sampling
$\mathcal{A}_1$
$\dots$
$\mathcal{A}_N$
$\mathcal{A}_{\mathcal{V}}^{(1)}$
$\mathcal{A}_{\mathcal{B}}^{(1)}$
$\mathcal{A}_{\mathcal{V}}^{(N)}$
$\mathcal{A}_{\mathcal{B}}^{(N)}$

Figure 1: Illustration of how the Uniform Sampling interface interacts with each base algorithm $\mathcal{A}_i$. Each base algorithm internally applies the direction and scale decomposition, using its own hyperparameters.

Lemma 3.2. *Let $(\|\cdot\|, \|\cdot\|_*)$ be a dual-norm pair and suppose that $\|\boldsymbol{w}\| \leq D$ for all $\boldsymbol{w} \in \mathcal{W}$. Then, for any sequence $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T$ in $\mathcal{W}$ and for any sequence of distributions $\boldsymbol{p}_1^*, \dots, \boldsymbol{p}_T^*$ satisfying $\mathbb{E}_{\boldsymbol{w} \sim \boldsymbol{p}_t^*}[\boldsymbol{w}] = \boldsymbol{u}_t$, Algorithm 3 with regularizer $\psi(\boldsymbol{w}) = \frac{1}{\eta} \sum_{i=1}^d (w_i \log(w_i) - w_i)$ and exploration distribution uniform over $\mathcal{W}$ guarantees*

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] \leq \frac{1 + \log(d)\,(1 + P_T^{\triangle})}{\eta} + \eta d D^2 \sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2 \,.$$

Note that, in this specific setting, we obtain a generalization of the notion of *switching* regret, which measures the number of times the comparator action changes, $S_T = \sum_t^T \mathbb{I}\{\boldsymbol{u}_t \neq \boldsymbol{u}_{t-1}\}$. This results from the fact that Continuous Exponential Weights works with the distributions over actions rather than the actions themselves.

With this guarantee at hand, we can see that the optimal choice of learning rate would be $\eta^* \sim \sqrt{\frac{P_T^{\triangle} \log d}{dT}}$. However, choosing this step-size would require knowledge of $P_T^{\triangle}$. Instead, we will use the guarantee-combining argument suggested above to consider a grid of exponentially spaced candidate step-sizes:

$$\mathcal{S} = \left\{ \frac{2^i}{GT} \wedge \frac{1}{G} \;:\; i = 0, 1, \dots \right\} . \tag{1}$$

By running a base algorithm for each $\eta_i \in \mathcal{S}$, we guarantee that at least one of them achieves regret within a constant factor of the optimal choice η^* (see Appendix B). Combining this with the guarantees of the scale and direction learners above and applying Proposition 3.1 we obtain the following guarantee (proof in Appendix A.2).

Theorem 3.3. *Let $\mathcal{S}$ as defined in Equation (1). For all $i \in |\mathcal{S}|$, let $\mathcal{A}_i$ be an instance of Algorithm 2 such that the direction learner $\mathcal{A}_{\mathcal{B}}^{(i)}$ is an instance of Algorithm 3 with step-size η_i on domain $\mathcal{B}_N$, and $\mathcal{A}_{\mathcal{V}}^{(i)}$ is a comparator-adaptive OLO routine. Then for any sequence $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T$ in $\mathcal{W}$ and sequence of distributions $\boldsymbol{p}_1^*, \dots, \boldsymbol{p}_T^*$ satisfying $\mathbb{E}_{\boldsymbol{w} \sim p_t^*}[\boldsymbol{w}] = \boldsymbol{u}_t$, Algorithm 1 guarantees*

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] = \widetilde{O}\left(G\epsilon + dMG(1 + P_T^{\triangle}) + dM\sqrt{N(1 + P_T^{\triangle}) \sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2} \right),$$

where $M = \max_t \|\boldsymbol{u}_t\|$.

As noted above, $P_T^{\triangle}$ is a generalization of the number of switches S_T. Hence, this result implies the optimal switching regret guarantee $\mathcal{O}(dM\sqrt{(S_T + 1)T})$ up to poly-logarithmic terms.

4 Conclusions and Future Work

The approach proposed in this paper is remarkably simple and allows hyperparameters to be tuned on-the-fly, enabling adaptivity to problem parameters that cannot be directly observed or estimated, such as the path-length. This leads to the first optimal parameter-free dynamic regret bound for linear bandits. This technique can easily be extended to other settings involving hard-to-estimate quantities, but it crucially relies on the scale and direction decomposition, which may not apply when the action set lacks such structure. Extending this to more general or irregular action sets—e.g., combinatorial sets—is non-trivial, as these often lack a clear notion of direction or have an irregular geometry that complicates such decoupling.

A natural question is whether it is possible to recover guarantees that scale with the standard path length P_T in linear bandits. One possible approach is to directly optimize in the action space—e.g., via mirror descent—and then construct a distribution p_t such that $\mathbb{E}_{p_t}[\boldsymbol{w}] = \boldsymbol{w}_t$, playing $\boldsymbol{w} \sim p_t$. However, the feasibility and variance control of such sampling methods remain challenging.

References

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017. URL `http://proceedings.mlr.press/v65/agarwal17b/agarwal17b.pdf`.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

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

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning*, volume 14, pp. 375, 2018.

Campolongo, N. and Orabona, F. A closer look at temporal variability in dynamic online learning, 2021. URL `https://arxiv.org/abs/2102.07666`.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087. PMLR, 2019.

Cutkosky, A. Combining online learning guarantees. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 895–913, Phoenix, USA, 2019. PMLR. URL `http://proceedings.mlr.press/v99/cutkosky19b.html`.

Cutkosky, A. Parameter-free, dynamic, and strongly-adaptive online learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2250–2259, Virtual, 2020. PMLR. URL `http://proceedings.mlr.press/v119/cutkosky20a.html`.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Herbster, M. and Warmuth, M. K. Tracking the best regressor. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 24–31, 1998. URL `https://dl.acm.org/doi/pdf/10.1145/279943.279949`.

Herbster, M. and Warmuth, M. K. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(281-309):10–1162, 2001. URL `https://www.jmlr.org/papers/volume1/herbster01a/herbster01a.pdf`.

Jacobsen, A. and Cutkosky, A. Parameter-free mirror descent. In *Conference on Learning Theory*, pp. 4160–4211. PMLR, 2022.

Jacobsen, A. and Cutkosky, A. Unconstrained online learning with unbounded losses. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14590–14630. PMLR, 2023. URL `https://proceedings.mlr.press/v202/jacobsen23a.html`.

Jacobsen, A. and Cutkosky, A. Online linear regression in dynamic environments via discounting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21083–21120. PMLR, 2024. URL `https://proceedings.mlr.press/v235/jacobsen24a.html`.

Luo, H., Zhang, M., Zhao, P., and Zhou, Z.-H. Corralling a larger band of bandits: A case study on switching regret for linear bandits. In *Conference on Learning Theory*, pp. 3635–3684. PMLR, 2022.

Marinov, T. V. and Zimmert, J. The pareto frontier of model selection for general contextual bandits, 2021. URL `https://arxiv.org/abs/2110.13282`.

McMahan, B. and Abernethy, J. Minimax optimal algorithms for unconstrained linear optimization. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/e00406144c1e7e35240afed70f34166a-Paper.pdf`.

Mcmahan, B. and Streeter, M. No-regret algorithms for unconstrained online convex optimization. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/hash/38ca89564b2259401518960f7a06f94b-Abstract.html`.

McMahan, H. B. and Orabona, F. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In Balcan, M. F., Feldman, V., and Szepesvári, C. (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 1020–1039, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v35/mcmahan14.html`.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2858–2887. PMLR, 09–12 Jul 2020. URL `http://proceedings.mlr.press/v125/mhammedi20a.html`.

Mokhtari, A., Shahrampour, S., Jadbabaie, A., and Ribeiro, A. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7195–7201. IEEE, 2016.

Odalric, M. and Munos, R. Adaptive bandits: Towards the best history-dependent strategy. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 570–578. JMLR Workshop and Conference Proceedings, 2011.

Orabona, F. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL `http://arxiv.org/abs/1912.13213`.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

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

van der Hoeven, D., Cutkosky, A., and Luo, H. Comparator-adaptive convex bandits. *CoRR*, abs/2007.08448, 2020. URL `https://arxiv.org/abs/2007.08448`.

Yan, Y.-H., Zhao, P., and Zhou, Z.-H. Online non-stochastic control with partial feedback. *Journal of Machine Learning Research*, 24(273):1–50, 2023.

Yang, T., Zhang, L., Jin, R., and Yi, J. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pp. 449–457. PMLR, 2016.

Zhang, L., Lu, S., and Zhou, Z.-H. Adaptive online learning in dynamic environments. *Advances in neural information processing systems*, 31, 2018.

Zhang, Z., Cutkosky, A., and Paschalidis, Y. Unconstrained dynamic regret via sparse coding. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74636–74670. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/ec2833cda146c277cdaa39066764f25c-Paper-Conference.pdf`.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Dynamic regret of convex and smooth functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12510–12520. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/939314105ce8701e67489642ef4d49e8-Paper.pdf`.

Zhao, P., Wang, G., Zhang, L., and Zhou, Z.-H. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1–52, 2024. URL `http://jmlr.org/papers/v25/21-0748.html`.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A Linear Bandits

In this section, we show how to derive the unconstrained linear bandit result stated in Theorem 3.3. We begin by establishing a base regret guarantee for the continuous exponential weights learner. To that end, we first prove the base guarantee.

Lemma 3.2. *Let $(\|\cdot\|, \|\cdot\|_*)$ be a dual-norm pair and suppose that $\|\boldsymbol{w}\| \le D$ for all $\boldsymbol{w} \in \mathcal{W}$. Then, for any sequence $\boldsymbol{u}_1, \dots, \boldsymbol{u}_T$ in $\mathcal{W}$ and for any sequence of distributions $\boldsymbol{p}_1^*, \dots, \boldsymbol{p}_T^*$ satisfying $\mathbb{E}_{\boldsymbol{w}\sim\boldsymbol{p}_t^*}[\boldsymbol{w}] = \boldsymbol{u}_t$, Algorithm 3 with regularizer $\psi(\boldsymbol{w}) = \frac{1}{\eta}\sum_{i=1}^d (w_i \log(w_i) - w_i)$ and exploration distribution uniform over $\mathcal{W}$ guarantees*

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] \le \frac{1 + \log(d)\,(1 + P_T^{\Delta})}{\eta} + \eta d D^2 \sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2 \,.$$

Proof. First observe that we can relate the dynamic regret *w.r.t.* the chosen actions to the dynamic regret *w.r.t.* the action sampling distributions as follows

$$\begin{aligned}
\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] &= \mathbb{E}\left[\sum_{t=1}^T \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t - \boldsymbol{u}_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \left\langle \boldsymbol{\ell}_t, \int \boldsymbol{w}\widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w} - \boldsymbol{u}_t\right\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \int \langle \boldsymbol{\ell}_t, \boldsymbol{w}\rangle\, (\widetilde{p}_t(\boldsymbol{w}) - p_t^*(\boldsymbol{w}))d\boldsymbol{w}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \int \left\langle \widehat{\boldsymbol{\ell}}_t, \boldsymbol{w}\right\rangle (\widetilde{p}_t(\boldsymbol{w}) - p_t^*(\boldsymbol{w}))d\boldsymbol{w}\right]
\end{aligned}$$

where $p_t^*(\boldsymbol{w})$ is the distribution supported on $\mathcal{W}$ such that $\mathbb{E}_{p_t^*}[\boldsymbol{w}] = \boldsymbol{u}_t$. Equivalently, we will slightly abuse notation by writing $\widehat{\boldsymbol{g}}_t(\boldsymbol{w}) = \left\langle \widehat{\boldsymbol{\ell}}_t, \boldsymbol{w}\right\rangle$ and $\langle \widehat{\boldsymbol{g}}_t, \boldsymbol{p}\rangle = \int \widehat{\boldsymbol{g}}_t(\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$, we have

$$\mathbb{E}\left[R_T((\boldsymbol{u}_1, \dots, \boldsymbol{u}_T))\right] \le \mathbb{E}\left[\sum_{t=1}^T \langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_t^*\rangle\right]. \tag{2}$$

Now we apply the following regret guarantee, proven in Appendix A.1:

Proposition A.1. *Let $\beta = 1/(T+1)$, let $\eta > 0$ satisfy $\eta\,|\widehat{g}_t(\boldsymbol{w})| \le 1$ for all $\boldsymbol{w} \in \mathcal{W}$, and let $\psi(p) = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\frac{\log(p(\boldsymbol{x}))}{\eta}\right] = \int \frac{p(\boldsymbol{x})\log(p(\boldsymbol{x}))}{\eta}d\boldsymbol{x}$. Set $\widetilde{p}_1$ = Uniform($\mathcal{W}$) and on each round suppose we play $\widetilde{p}_t = (1-\beta)p_t + \beta\widetilde{p}_1$ and update $p_{t+1} = \arg\min_{p\in\mathcal{M}_1(\mathcal{W})} \mathbb{E}_{\boldsymbol{w}\sim p}[\widehat{\boldsymbol{g}}_t(\boldsymbol{w})] + D_\psi(p|\widetilde{p}_t)$. Then for any sequence of distributions $p_1^*, \dots, p_T^*$ in $\mathcal{M}_1(\mathcal{W})$, it holds that*

$$\sum_{t=1}^T \langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_t^*\rangle \le \frac{1 + \log(\mathrm{Vol}(\mathcal{W}))\,(1 + P_T^{\Delta})}{\eta} + \eta\sum_{t=1}^T \mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}[\widehat{\boldsymbol{g}}_t(\boldsymbol{w})^2],$$

where $P_T^{\Delta} = \sum_{t=2}^T \left\|\boldsymbol{p}_t^* - \boldsymbol{p}_{t-1}^*\right\|_1$.

Hence, plugging this back into Equation (2), the regret of Algorithm 3 is bounded as

$$\begin{aligned}
\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] &\le \mathbb{E}\left[\frac{1 + \log(\mathrm{Vol}(\mathcal{W}))\,(1 + P_T^{\Delta})}{\eta} + \eta\sum_{t=1}^T \mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}[\widehat{\boldsymbol{g}}_t(\boldsymbol{w})^2]\right] \\
&= \mathbb{E}\left[\frac{1 + \log(\mathrm{Vol}(\mathcal{W}))\,(1 + P_T^{\Delta})}{\eta} + \eta\sum_{t=1}^T \mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}\left[\left\langle \widehat{\boldsymbol{\ell}}_t, \boldsymbol{w}\right\rangle^2\right]\right].
\end{aligned}$$

To bound the last term, define $\boldsymbol{Q}_t = \mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}[\boldsymbol{w}\boldsymbol{w}^\top]$ and observe that

$$\mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}\left[\left\langle \widehat{\boldsymbol{\ell}}_t, \boldsymbol{w}\right\rangle^2\right] = \int \langle \boldsymbol{\ell}_t, \boldsymbol{w}_t\rangle^2\, \boldsymbol{w}_t^\top \boldsymbol{Q}_t^{-1}\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{Q}_t^{-1}\boldsymbol{w}_t^\top\widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w}$$

$$
\begin{aligned}
&= \left\langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \right\rangle^2 \boldsymbol{w}_t^\top \boldsymbol{Q}_t^{-1} \underbrace{\mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}[\boldsymbol{w}\boldsymbol{w}^\top]}_{\boldsymbol{Q}_t} \boldsymbol{Q}_t^{-1}\boldsymbol{w}_t \\
&= \left\langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \right\rangle^2 \left\|\boldsymbol{w}_t\right\|^2_{\boldsymbol{Q}_t^{-1}}
\end{aligned}
$$

Therefore, for $\boldsymbol{w}_t \sim \widetilde{p}_t$, we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\int \left\langle \widehat{\boldsymbol{\ell}}_t, \boldsymbol{w} \right\rangle^2 \widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \boldsymbol{\ell}_t, \boldsymbol{w}_t \right\rangle^2 \left\|\boldsymbol{w}_t\right\|^2_{\boldsymbol{Q}_t^{-1}}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \left\|\boldsymbol{\ell}_t\right\|_* D^2 \mathbb{E}\left[\left\|\boldsymbol{w}_t\right\|^2_{\boldsymbol{Q}_t^{-1}} \middle| \widetilde{p}_t\right]\right] \\
&= \mathbb{E}\left[D^2 \sum_{t=1}^{T} \left\|\boldsymbol{\ell}_t\right\|_*^2 \operatorname{Tr}\left(\int \boldsymbol{w}\boldsymbol{w}^\top \widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w}\boldsymbol{Q}_t^{-1}\right)\right] \\
&= \mathbb{E}\left[D^2 d \sum_{t=1}^{T} \left\|\boldsymbol{\ell}_t\right\|_*^2\right],
\end{aligned}
$$

for $\|\boldsymbol{w}_t\| \leq D$. Plugging this back in above, we have

$$
\mathbb{E}\left[R_T(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)\right] \leq \frac{1 + \log\left(\operatorname{Vol}\left(\mathcal{W}\right)\right)\left(1 + P_T^{\Delta}\right)}{\eta} + \eta d D^2 \sum_{t=1}^{T} \left\|\boldsymbol{\ell}_t\right\|_*^2 .
$$

□

A.1 Switching Regret on the Simplex

For completeness, in this section we show how to derive the dynamic regret of the base OLO algorithm. The result follows a standard mirror descent argument.

Throughout this section, we will use the short-hand notation $\langle \boldsymbol{g}, \boldsymbol{p}\rangle = \int g(\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$, where g is a function with domain $\mathcal{W}$, p is a the density of a distribution of $\mathcal{W}$. We let $\mathcal{M}_1(\mathcal{W})$ denote the set of distributions over $\mathcal{W}$.

Proposition A.1. *Let $\beta = 1/(T+1)$, let $\eta > 0$ satisfy $\eta\left|\widehat{g}_t(\boldsymbol{w})\right| \leq 1$ for all $\boldsymbol{w} \in \mathcal{W}$, and let $\psi(p) = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\frac{\log(p(\boldsymbol{x}))}{\eta}\right] = \int \frac{p(\boldsymbol{x})\log(p(\boldsymbol{x}))}{\eta}d\boldsymbol{x}$. Set $\widetilde{p}_1 = \text{Uniform}(\mathcal{W})$ and on each round suppose we play $\widetilde{p}_t = (1-\beta)p_t + \beta\widetilde{p}_1$ and update $p_{t+1} = \arg\min_{p\in\mathcal{M}_1(\mathcal{W})} \mathbb{E}_{\boldsymbol{w}\sim p}[\widehat{\boldsymbol{g}}_t(\boldsymbol{w})] + D_\psi(p|\widetilde{p}_t)$. Then for any sequence of distributions $p_1^*, \ldots, p_T^*$ in $\mathcal{M}_1(\mathcal{W})$, it holds that*

$$
\sum_{t=1}^{T} \left\langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_t^* \right\rangle \leq \frac{1 + \log\left(\operatorname{Vol}\left(\mathcal{W}\right)\right)\left(1 + P_T^{\Delta}\right)}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{w}\sim\widetilde{p}_t}[\widehat{\boldsymbol{g}}_t(\boldsymbol{w})^2],
$$

where $P_T^{\Delta} = \sum_{t=2}^{T} \left\|\boldsymbol{p}_t^* - \boldsymbol{p}_{t-1}^*\right\|_1$.

Proof. Using a standard dynamic regret decomposition for mirror descent updates (see, e.g., Jacobsen & Cutkosky (2023, Lemma A.1)), we have

$$
\begin{aligned}
\sum_{t=1}^{T} \left\langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_t^* \right\rangle \leq \sum_{t=1}^{T} \underbrace{D_\psi(p_t^*|\widetilde{p}_t) - D_\psi(p_t^*|\widetilde{p}_{t+1})}_{=:\mathcal{P}_t} + \sum_{t=1}^{T} \underbrace{D_\psi(p_t^*|p_{t+1}) - D_\psi(p_t^*|\widetilde{p}_{t+1})}_{=:\xi_t} \\
+ \sum_{t=1}^{T} \underbrace{\left\langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_{t+1} \right\rangle - D_\psi(p_{t+1}|\widetilde{p}_t)}_{=:\delta_t}.
\end{aligned}
$$

Observe that for any $u, p, q \in \mathcal{M}_1(\mathcal{W})$ we have

$$
D_\psi(u \mid p) - D_\psi(u \mid q) = \frac{1}{\eta}\int u(\boldsymbol{w})\log\left(\frac{u(\boldsymbol{w})}{p(\boldsymbol{w})}\right)d\boldsymbol{w} - \frac{1}{\eta}\int u(\boldsymbol{w})\log\left(\frac{u(\boldsymbol{w})}{q(\boldsymbol{w})}\right)d\boldsymbol{w}
$$

$$= \frac{1}{\eta} \int u(\boldsymbol{w}) \log \left(\frac{q(\boldsymbol{w})}{p(\boldsymbol{w})} \right) d\boldsymbol{w} \tag{3}$$

Thus, the terms ξ_t can be bound as

$$\begin{aligned}
\sum_{t=1}^{T} \xi_t &= \sum_{t=1}^{T} D_\psi(p_t^* \mid \widetilde{p}_{t+1}) - D_\psi(p_t^* \mid p_{t+1}) \\
&= \frac{1}{\eta} \sum_{t=1}^{T} \int p_t^*(\boldsymbol{w}) \log \left(\frac{p_{t+1}(\boldsymbol{w})}{\widetilde{p}_{t+1}(\boldsymbol{w})} \right) d\boldsymbol{w} \\
&= \frac{1}{\eta} \sum_{t=1}^{T} \int p_t^*(\boldsymbol{w}) \log \left(\frac{p_{t+1}(\boldsymbol{w})}{(1-\beta) p_{t+1}(\boldsymbol{w}) + \beta \widetilde{p}_1(\boldsymbol{w})} \right) d\boldsymbol{w} \\
&\leq \frac{1}{\eta} \sum_{t=1}^{T} \int p_t^*(\boldsymbol{w}) \log \left(\frac{1}{(1-\beta)} \right) d\boldsymbol{w} \\
&\leq \frac{T}{\eta} \log \left(\frac{1}{(1-\beta)} \right).
\end{aligned}$$

Now, focusing on the path terms $\mathcal{P}_t$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathcal{P}_t &= D_\psi(p_1^* | \widetilde{p}_1) - D_\psi(p_T^* | \widetilde{p}_{T+1}) + \sum_{t=2}^{T} D_\psi(p_t^* | \widetilde{p}_t) - D_\psi(p_{t-1}^* | \widetilde{p}_t) \\
&= D_\psi(p_1^* | \widetilde{p}_1) - D_\psi(p_1^* | \widetilde{p}_{T+1}) \\
&\qquad + D_\psi(p_T^* | \widetilde{p}_1) - D_\psi(p_1^* | \widetilde{p}_1) + \sum_{t=2}^{T} \frac{1}{\eta} \int (p_{t-1}^*(\boldsymbol{w}) - p_t^*(\boldsymbol{w})) \left(\log (\widetilde{p}_t(\boldsymbol{w})) - \log (\widetilde{p}_1(\boldsymbol{w})) \right) d\boldsymbol{w} \\
&= D_\psi(p_T^* | \widetilde{p}_1) - D_\psi(p_T^* | \widetilde{p}_{T+1}) + \sum_{t=2}^{T} \frac{1}{\eta} \int \left(p_{t-1}^*(\boldsymbol{w}) - p_t^*(\boldsymbol{w}) \right) \log \left(\frac{\widetilde{p}_t(\boldsymbol{w})}{\widetilde{p}_1(\boldsymbol{w})} \right) d\boldsymbol{w} \\
&\leq D_\psi(p_T^* | \widetilde{p}_1) - D_\psi(p_T^* | \widetilde{p}_{T+1}) + \sum_{t=2}^{T} \frac{1}{\eta} \left\| \boldsymbol{p}_t^* - \boldsymbol{p}_{t-1}^* \right\|_1 \sup_{\boldsymbol{w} \in \mathcal{W}} \left| \log \left(\frac{\widetilde{p}_t(\boldsymbol{w})}{\widetilde{p}_1(\boldsymbol{w})} \right) \right| \\
&\leq D_\psi(p_T^* | \widetilde{p}_1) - D_\psi(p_T^* | \widetilde{p}_{T+1}) + \sum_{t=2}^{T} \frac{1}{\eta} \left\| \boldsymbol{p}_t^* - \boldsymbol{p}_{t-1}^* \right\|_1 \left| \log \left(\mathrm{Vol} \left(\mathcal{W} \right) \right) \right|,
\end{aligned}$$

where the first inequality applies Hölder inequality. Moreover, using Equation (3), the first two terms can be bound as

$$\begin{aligned}
D_\psi(p_T^* | \widetilde{p}_1) - D_\psi(p_T^* | \widetilde{p}_{T+1}) &= \frac{1}{\eta} \int p_T^*(\boldsymbol{w}) \log \left(\frac{\widetilde{p}_{T+1}(\boldsymbol{w})}{\widetilde{p}_1(\boldsymbol{w})} \right) d\boldsymbol{w} \\
&= \frac{1}{\eta} \int p_T^*(\boldsymbol{w}) \log \left(\frac{(1-\beta) p_{T+1}(\boldsymbol{w}) + \beta \widetilde{p}_1(\boldsymbol{w})}{\widetilde{p}_1(\boldsymbol{w})} \right) d\boldsymbol{w} \\
&\leq \frac{1}{\eta} \int p_T^*(\boldsymbol{w}) \log \left(\frac{(1-\beta) p_{T+1}(\boldsymbol{w})}{\widetilde{p}_1(\boldsymbol{w})} + \beta \right) d\boldsymbol{w} \\
&\leq \frac{1}{\eta} \log \left(\mathrm{Vol} \left(\mathcal{W} \right) (1-\beta) + \beta \right) \leq \frac{\log_+ \left(\mathrm{Vol} \left(\mathcal{W} \right) \right)}{\eta}
\end{aligned}$$

where we've recalled $\widetilde{p}_t = (1-\beta) p_t + \beta \widetilde{p}_1$ for any t and used $(1-\beta)a + \beta b \leq \max \{a, b\}$.

Returning to our regret bound, we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \widehat{\boldsymbol{g}}_t, \widetilde{\boldsymbol{p}}_t - \boldsymbol{p}_t^* \rangle &\leq \frac{1}{\eta} \left| \log \left(\mathrm{Vol} \left(\mathcal{W} \right) \right) \right| + \frac{1}{\eta} \left| \log \left(\mathrm{Vol} \left(\mathcal{W} \right) \right) \right| \sum_{t=2}^{T} \left\| \boldsymbol{p}_{t-1}^* - \boldsymbol{p}_t^* \right\|_1 + \frac{T}{\eta} \log \left(\frac{1}{(1-\beta)} \right) + \sum_{t=1}^{T} \delta_t \\
&= \frac{\left| \log \left(\mathrm{Vol} \left(\mathcal{W} \right) \right) \right| (1 + P_T^{\Delta})}{\eta} + \frac{T}{\eta} \log \left(\frac{1}{(1-\beta)} \right) + \sum_{t=1}^{T} \delta_t \qquad (4)
\end{aligned}$$

We now have to take care of the terms $\sum_{t=1}^{T}\delta_t$. Let p_{t+1}^{+} denote the unconstrained minimizer $p_{t+1}^{+}=\arg\min_p \langle\widehat{\boldsymbol{g}}_t,\boldsymbol{p}\rangle+D_\psi(p|\widetilde{p}_t)$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\delta_t &= \sum_{t=1}^{T}\langle\widehat{\boldsymbol{g}}_t,\widetilde{\boldsymbol{p}}_t-\boldsymbol{p}_{t+1}\rangle-D_\psi(p_{t+1}|\widetilde{p}_t)\le\sum_{t=1}^{T}\langle\widehat{\boldsymbol{g}}_t,\widetilde{\boldsymbol{p}}_t-\boldsymbol{p}_{t+1}^{+}\rangle-D_\psi(p_{t+1}^{+}|\widetilde{p}_t)\\
&=\sum_{t=1}^{T}\left\langle\nabla\psi(\widetilde{\boldsymbol{p}}_t)-\nabla\psi(\boldsymbol{p}_{t+1}^{+}),\widetilde{\boldsymbol{p}}_t-\boldsymbol{p}_{t+1}^{+}\right\rangle-D_\psi(p_{t+1}^{+}|\widetilde{p}_t)\\
&=\sum_{t=1}^{T}D_\psi(\widetilde{p}_t|p_{t+1}^{+})=\sum_{t=1}^{T}\int\frac{1}{\eta}\Big(\widetilde{p}_t(\boldsymbol{w})\log\big(\widetilde{p}_t(\boldsymbol{w})/p_{t+1}^{+}(\boldsymbol{w})\big)-\widetilde{p}_t(\boldsymbol{w})+p_{t+1}^{+}(\boldsymbol{w})\Big)d\boldsymbol{w}\\
&=\sum_{t=1}^{T}\int\frac{1}{\eta}\Big(\widetilde{p}_t(\boldsymbol{w})\log\Big(e^{\eta\widehat{g}_t(\boldsymbol{w})}\Big)-\widetilde{p}_t(\boldsymbol{w})+\widetilde{p}_t(\boldsymbol{w})e^{-\eta\widehat{g}_t(\boldsymbol{w})}\Big)d\boldsymbol{w}\\
&=\sum_{t=1}^{T}\int\frac{\widetilde{p}_t(\boldsymbol{w})}{\eta}\Big(\eta\widehat{g}_t(\boldsymbol{w})-1+e^{-\eta\widehat{g}_t(\boldsymbol{w})}\Big)\,d\boldsymbol{w}\\
&\le\sum_{t=1}^{T}\int\eta\widehat{g}_t(\boldsymbol{w})^2\widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w},
\end{aligned}$$

where we've used $p_{t+1}^{+}(\boldsymbol{w})=\widetilde{p}_t(\boldsymbol{w})e^{-\eta\widehat{g}_t(\boldsymbol{w})}$ via the first-order optimality condition for p_{t+1}^{+} and $e^{-x}-1+x\le x^2$ for $x\ge-1$. Hence,

$$\begin{aligned}
\sum_{t=1}^{T}\langle\widehat{\boldsymbol{g}}_t,\widetilde{\boldsymbol{p}}_t-\boldsymbol{p}_t^*\rangle\le\frac{|\log(\mathrm{Vol}(\mathcal{W}))|}{\eta}&+\frac{|\log(\mathrm{Vol}(\mathcal{W}))|\sum_{t=2}^{T}\left\|\boldsymbol{p}_t^*-\boldsymbol{p}_{t-1}^*\right\|_1}{\eta}\\
&+\frac{T}{\eta}\log\left(\frac{1}{(1-\beta)}\right)+\eta\sum_{t=1}^{T}\int\widehat{g}_t(\boldsymbol{w})^2\widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w}.
\end{aligned}$$

Setting $\beta=1/(T+1)$, and using the following simple inequalities:

$$\log\left(\frac{1}{1-\beta}\right)=\log\left(\frac{1}{1-1/(T+1)}\right)=\log\left(\frac{T+1}{T}\right)=\log\left(1+\frac{1}{T}\right)\le\frac{1}{T},$$

we have an overall regret bound of

$$\sum_{t=1}^{T}\langle\widehat{\boldsymbol{g}}_t,\widetilde{\boldsymbol{p}}_t-\boldsymbol{p}_t^*\rangle\le\frac{1+|\log(\mathrm{Vol}(\mathcal{W}))|(1+P_T^{\Delta})}{\eta}+\eta\sum_{t=1}^{T}\int\widehat{g}_t(\boldsymbol{w})^2\widetilde{p}_t(\boldsymbol{w})d\boldsymbol{w}.$$

□

A.2 Optimal Switching Regret without Prior Knowledge

Now combining the results from Proposition 3.1 and Lemma 3.2, we are ready to prove Theorem 3.3.

Theorem 3.3. *Let $\mathcal{S}$ as defined in Equation* (1). *For all $i\in|\mathcal{S}|$, let $\mathcal{A}_i$ be an instance of Algorithm 2 such that the direction learner $\mathcal{A}_{\mathcal{B}}^{(i)}$ is an instance of Algorithm 3 with step-size η_i on domain $\mathcal{B}_N$, and $\mathcal{A}_{\mathcal{V}}^{(i)}$ is a comparator-adaptive OLO routine. Then for any sequence $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_T$ in $\mathcal{W}$ and sequence of distributions $\boldsymbol{p}_1^*,\ldots,\boldsymbol{p}_T^*$ satisfying $\mathbb{E}_{\boldsymbol{w}\sim p_t^*}[\boldsymbol{w}]=\boldsymbol{u}_t$, Algorithm 1 guarantees*

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_T)\right]=\widetilde{O}\left(G\epsilon+dMG(1+P_T^{\Delta})+dM\sqrt{N(1+P_T^{\Delta})\sum_{t=1}^{T}\|\boldsymbol{\ell}_t\|_*^2}\right),$$

where $M=\max_t\|\boldsymbol{u}_t\|$.

Proof. Observe that $\eta_{\max}/\eta_{\min} = T$, so $N = \lceil \log_2(\eta_{\max}/\eta_{\min}) \rceil = \lceil \log_2(T) \rceil$. By Proposition 3.1, for any $\eta_i \in \mathcal{S}$ we have

$$\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] \le \mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{V}}^{(i)}}(MN) + MNR_T\left(\frac{\boldsymbol{u}_1}{M}, \dots, \frac{\boldsymbol{u}_T}{M}\right) + (N-1)G\epsilon\right],$$

where $M = \max_t \|\boldsymbol{u}_t\|$. The regret guarantee of $\mathcal{A}_{\mathcal{V}}^{(i)}$ ensures that

$$\begin{aligned}\mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{V}}^{(i)}}(NM)\right] &\le G\epsilon + MN\mathbb{E}\left[\sqrt{\sum_t \left\langle \widehat{\boldsymbol{\ell}}_t^{(i)}, \boldsymbol{\beta}_t^{(i_t)} \right\rangle^2 \log\left(\frac{MN\Lambda_T}{\epsilon G} + 1\right)} \vee G\log\left(\frac{NM\Lambda_T}{\epsilon G} + 1\right)\right] \\ &= \widetilde{O}\left(\epsilon G + M\sqrt{N\sum_t \|\boldsymbol{\ell}_t\|_*^2}\right),\end{aligned}$$

since $\left\|\boldsymbol{\beta}_t^{(i_t)}\right\| \le 1$ for all t and $\mathbb{E}\left[\sqrt{\sum_{t=1}^T \left\|\widehat{\boldsymbol{\ell}}_t^{(i)}\right\|_*^2}\right] \le \sqrt{\sum_{t=1}^T \mathbb{E}\left[\|\boldsymbol{\ell}_t\|_*^2 \mathbb{I}\{i_t = i\}\right]} = \sqrt{\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2 / N}$ via Jensen's inequality. Moreover, letting $\widetilde{\boldsymbol{u}}_t = \frac{\boldsymbol{u}_t}{M}$, each of the algorithms $\mathcal{A}_{\mathcal{B}}^{(i)}$ guarantees

$$\begin{aligned}\mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{B}}^{(i)}}(\widetilde{\boldsymbol{u}}_1, \dots, \widetilde{\boldsymbol{u}}_T)\right] &\le \frac{1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right)(1 + P_T^{\Delta})}{\eta_i} + \eta_i d \sum_{t=1}^T \mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_t^{(i)}\right\|_*^2\right] \\ &= \frac{1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right)(1 + P_T^{\Delta})}{\eta_i} + \frac{\eta_i d}{N}\sum_{t=1}^T \mathbb{E}\left[\|\boldsymbol{\ell}_t\|_*^2\right],\end{aligned}$$

where we've used the fact that $\mathbb{E}\left[\left\|\widehat{\boldsymbol{\ell}}_t^{(i)}\right\|_*^2\right] = \mathbb{E}\left[\|\boldsymbol{\ell}_t\|_*^2 \mathbb{I}\{i_t = i\}\right] = \|\boldsymbol{\ell}_t\|_*^2 / N$. Hence, by Lemma B.1, there is an $i \in [N]$ such that

$$\begin{aligned}NM\mathbb{E}\left[R_T^{\mathcal{A}_{\mathcal{B}}^{(i)}}(\widetilde{\boldsymbol{u}}_1, \dots, \widetilde{\boldsymbol{u}}_T)\right] &\le 3M\sqrt{dN(1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right)(1 + P_T^{\Delta}))\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2} \\ &\quad + NM(1 + \log\left(d\right)(1 + P_T^{\Delta}))G + \frac{dM}{GT}\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2 \\ &\le 3\sqrt{dN(1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right)(1 + P_T^{\Delta}))\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2} \\ &\quad + NM(1 + \log\left(d\right)(1 + P_T^{\Delta}))G + MGd\end{aligned}$$

Hence, combining with the previous display we have

$$\begin{aligned}\mathbb{E}\left[R_T(\boldsymbol{u}_1, \dots, \boldsymbol{u}_T)\right] &= \widetilde{O}\Bigg(NG(\epsilon + dM) + NM(1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right))(1 + P_T^{\Delta})G \\ &\qquad + M\sqrt{N\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2} + 3M\sqrt{dN(1 + \log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right)(1 + P_T^{\Delta}))\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2}\Bigg) \\ &= \widetilde{O}\left(G\epsilon + dMG(1 + P_T^{\Delta}) + dM\sqrt{N(1 + P_T^{\Delta})\sum_{t=1}^T \|\boldsymbol{\ell}_t\|_*^2}\right),\end{aligned}$$

where the last line hides polylog factors and bounds $\log\left(\mathrm{Vol}\left(\mathcal{W}\right)\right) \le \widetilde{\mathcal{O}}(d)$.

□

B Supporting Lemmas

In this section, we present the supporting lemmas used to combine the guarantees of the base algorithms.

Lemma B.1. *Let $b > 1$, $0 < \eta_{\min} \le \eta_{\max}$ and let $\mathcal{S} = \left\{\eta_i = \eta_{\min} b^i \wedge \eta_{\max} : i = 0, 1, \ldots\right\}$ Then for any $P, V \in \mathbb{R}_{\ge 0}$, there is an $\eta \in \mathcal{S}$ such that*

$$R(\eta) := \frac{P}{\eta} + \eta V \le (b+1)\sqrt{PV} + \frac{P}{\eta_{\max}} + \eta_{\min} V$$

Proof. Define

$$R(\eta) = \frac{P}{\eta} + \eta V.$$

Clearly the bound is optimized by $\eta^* = \arg\min_\eta \frac{P}{\eta} + \eta V = \sqrt{P/V}$. Suppose that $\eta^* \in [\eta_{\min}, \eta_{\max}]$. Then there is an $i \in 0, 1, \ldots$ such that $\eta_i \le \eta^* \le \eta_{i+1} = b\eta_i$, hence setting $\eta = \eta_i$ yields

$$R(\eta) = \frac{P}{\eta_i} + \eta_i V \le b\frac{P}{\eta^*} + \eta^* V = (b+1)\sqrt{PV}. \tag{5}$$

Otherwise, $\eta^* \notin [\eta_{\min}, \eta_{\max}]$ and there are two cases to consider. First suppose that $\eta^* \ge \eta_{\max}$, then setting $\eta = \eta_{\max}$ we have

$$R(\eta) = \frac{P}{\eta_{\max}} + \eta_{\max} V \le \frac{P}{\eta_{\max}} + \eta^* V = \frac{P}{\eta_{\max}} + \sqrt{PV}. \tag{6}$$

Likewise, if $\eta^* \le \eta_{\min}$, then setting $\eta = \eta_{\min}$ yields

$$R(\eta) = \frac{P}{\eta_{\min}} + \eta_{\min} V \le \frac{P}{\eta^*} + \eta_{\min} V = \sqrt{PV} + \eta_{\min} V. \tag{7}$$

Overall, combining Equations (5) to (7), we have that there is an $\eta \in \mathcal{S}$ such that

$$R(\eta) \le (b+1)\sqrt{PV} + \frac{P}{\eta_{\max}} + \eta_{\min} V.$$

□

Lemma B.2. *Let $b > 1, c > 1$, $0 < \eta_{\min} \le \eta_{\max}$, $0 < \delta_{\min} \le \delta_{\max}$ and let $\mathcal{S} = \left\{\eta_i = \eta_{\min} b^i \wedge \eta_{\max} : i = 0, 1, \ldots\right\}$ and $\left\{\delta_j = \delta_{\min} c^j \wedge \delta_{\max} : j = 0, 1, \ldots\right\}$ Then for any $P, V_1, V_2 \in \mathbb{R}_{\ge 0}$, there is an $\eta \in \mathcal{S}$ such that*

$$R(\eta, \delta) := \frac{P}{\eta} + \eta\frac{V_1}{\delta^2}T + \delta V_2 T \le \frac{P}{\eta_{\max}} + \eta_{\min}\frac{V_1}{\delta_{\max}^2}T + \delta_{\min} V_2 T + \frac{1+b+c^2}{\sqrt{2}} V_2^{1/2}(PV_1)^{1/4} T^{3/4}$$

Proof. Define:

$$R(\eta, \delta) = \frac{P}{\eta} + \eta\frac{V_1}{\delta^2}T + \delta V_2 T$$

To obtain the optimal η^* and δ^*:

$$\frac{\partial R}{\partial \eta} = -\frac{P}{\eta^2} + \frac{V_1 T}{\delta^2} = 0 \implies \eta = \sqrt{\frac{P\delta^2}{V_1 T}}.$$

Then,

$$\frac{\partial R}{\partial \delta} = TV_2 - 2\frac{\eta V_1 T}{\delta^3} = 0 \implies V_2 T - 2\frac{\left(\sqrt{\frac{P\delta^2}{V_1 T}}\right) V_1 T}{\delta^3} = 0$$

$$\implies V_2 T - 2\frac{\sqrt{PV_1 T}}{\delta^2} = 0 \implies \delta = \left(\frac{2}{V_2}\sqrt{\frac{PV_1}{T}}\right)^{1/2}$$

So, $\delta^* = \left(\frac{2}{V_2}\sqrt{\frac{PV_1}{T}}\right)^{1/2}$ and $\eta^* = \sqrt{\frac{P\frac{2}{V_2}\sqrt{\frac{PV_1}{T}}}{V_1 T}} = \sqrt{\frac{2P^{3/2}}{V_2 V_1^{1/2} T^{3/2}}}$ yields

$$R(\eta^*, \delta^*) \leq 2\sqrt{2} V_2^{1/2} (PV_1)^{1/4} T^{3/4} \,.$$

Now suppose that $\eta^* \in [\eta_{\min}, \eta_{\max}]$ and $\delta^* \in [\delta_{\min}, \delta_{\max}]$. Then there are $i, j \in 0, 1, \dots$ such that $\eta_i \leq \eta^* \leq \eta_{i+1} = b\eta_i$ and $\delta_j \leq \delta^* \leq \delta_{j+1} = c\delta_j$, hence setting $\eta = \eta_i$ and $\delta = \delta_j$ yields

$$R(\eta, \delta) \leq \frac{P}{\eta} + \eta\frac{V_1}{\delta^2}T + \delta V_2 T \leq b\frac{P}{\eta^*} + c^2\eta^*\frac{V_1}{\delta^{*2}}T + \delta^* V_2 T \leq \frac{1+b+c^2}{\sqrt{2}} V_2^{1/2}(PV_1)^{1/4}T^{3/4}$$

Now we need to analyze the other cases, note that we can proceed as in the previous lemma for both the parameters, if $\eta^* \geq \eta_{\max}$ then the second term can be optimized with η^*, while in the other case the first term will be. The same reasoning is used for the δ parameter, yielding:

$$R(\eta, \delta) \leq \frac{P}{\eta_{\max}} + \eta_{\min}\frac{V_1}{\delta_{\max}^2}T + \delta_{\min} V_2 T + \frac{1+b+c^2}{\sqrt{2}} V_2^{1/2}(PV_1)^{1/4}T^{3/4}$$

□