Proceedings of Machine Learning Research vol 178:1–50, 2022 35th Annual Conference on Learning Theory

Corralling a Larger Band of Bandits: A Case Study on Switching Regret for Linear Bandits

Haipeng Luo* HAIPENGL@USC.EDU
University of Southern California

Mengxiao Zhang* ZHAN147@USC.EDU
University of Southern California

Peng Zhao* ZHAOP@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University

Zhi-Hua Zhou* ZHOUZH @LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the problem of combining and learning over a set of adversarial bandit algorithms with the goal of adaptively tracking the best one on the fly. The CORRAL algorithm of Agarwal et al. (2017) and its variants (Foster et al., 2020a) achieve this goal with a regret overhead of order $\widetilde{\mathcal{O}}(\sqrt{MT})$ where M is the number of base algorithms and T is the time horizon. The polynomial dependence on M, however, prevents one from applying these algorithms to many applications where M is poly(T) or even larger. Motivated by this issue, we propose a new recipe to corral a larger band of bandit algorithms whose regret overhead has only *logarithmic* dependence on M as long as some conditions are satisfied. As the main example, we apply our recipe to the problem of adversarial linear bandits over a d-dimensional ℓ_p unit-ball for $p \in (1, 2]$. By corralling a large set of T base algorithms, each starting at a different time step, our final algorithm achieves the first optimal switching regret $\widetilde{\mathcal{O}}(\sqrt{dST})$ when competing against a sequence of comparators with S switches (for some known S). We further extend our results to linear bandits over a smooth and strongly convex domain as well as unconstrained linear bandits.

1. Introduction

We consider the problem of combining a set of bandit algorithms to learn the best one on the fly, which has many applications in dealing with uncertainty from the environment. Indeed, by combining a set of base algorithms, each dedicated for a certain type of environments, the final meta algorithm can then automatically adapt to and perform well in every problem instance encountered, as long as the price of such meta-level learning is small enough. While such ideas have a long history in online learning, doing so with partial information (that is, bandit feedback) is particularly challenging, and only recently have we seen success in various settings (Agarwal et al., 2017; Pacchiano et al., 2020; Foster et al., 2020a; Lee et al., 2020; Krishnamurthy et al., 2021; Wei and Luo, 2021; Zhao et al., 2021a; Wei et al., 2022).

We focus on an adversarial setting where the data are generated in an arbitrary and potentially malicious manner. The closest work is (Agarwal et al., 2017), where a generic algorithm called CORRAL is developed to learn over a set of M base algorithms with extra regret overhead $\widetilde{\mathcal{O}}(\sqrt{MT})$

[*] Authors are listed in alphabetical order.

after T rounds. In order to maintain $\widetilde{\mathcal{O}}(\sqrt{T})$ overall regret, which is often the optimal bound and the goal when designing bandit algorithms, CORRAL can thus at most tolerate $M = \text{poly}(\log T)$ base algorithms. However, there are many applications where M needs to be much larger to cover all possible scenarios of interest (we will soon provide an example where M needs to be of order T). Therefore, a natural question arises: *can we corral an even larger band of bandit algorithms, ideally with only logarithmic dependence on M in the regret?*

As an attempt to answer this question, we focus on the adversarial linear bandit problem and develop a new recipe to combine base algorithms, which reduces the problem to designing good unbiased loss estimators for the base algorithms and good optimistic loss estimators for the meta algorithm. As long as these estimators ensure certain properties, the resulting algorithm enjoys logarithmic dependence on M in the regret. We discuss this recipe in detail along with a warm-up example on the classic multi-armed bandit problem in Section 3.

Then, as a main example, in Section 4 we apply this recipe to develop the first optimal switching regret bound for adversarial linear bandits over a d-dimensional ℓ_p unit ball with $p \in (1, 2]$.[1] Switching regret measures the learner's performance against a sequence of changing comparators with S switches, and a standard technique to achieve so in the full-information setting is by combining T base algorithms, each of which starts at a different time step and is guaranteed to perform well against a fixed comparator starting from this step (that is, a standard static regret guarantee); see for example (Hazan and Seshadhri, 2007; Daniely et al., 2015; Luo and Schapire, 2015). Applying the same idea to bandit problems was not possible before because as mentioned, previous methods such as CORRAL cannot afford T base algorithms.[2] However, this is exactly where our approach shines. Indeed, by using our recipe to combine T instances of the algorithm of (Bubeck et al., 2018) together with carefully designed loss estimators, we manage to achieve logarithmic dependence on the number of base algorithms, resulting in the optimal (up to logarithmic factors) switching regret $\widetilde{\mathcal{O}}(\sqrt{dST})$ for this problem for any fixed S. As another example, in Appendix C we also generalize our results from ℓ_p balls to smooth and strongly convex sets.

Finally, in Section 5 we further generalize our results to the unconstrained linear bandit problem and obtain the first comparator-adaptive switching regret of order $\widetilde{\mathcal{O}}\big(\max_{k\in[S]} \|\mathring{u}_k\|_2 \cdot \sqrt{dST}\big)$ where $\mathring{u}_k$ is the k-th (arbitrary) comparator. The algorithm requires two components, one of which is exactly our new algorithm developed for ℓ_p balls, the other being a new parameter-free algorithm for unconstrained Online Convex Optimization with the first comparator-adaptive switching regret. We note that this latter algorithm/result might be of independent interest.

High-level ideas. For such as a meta learning framework, it is standard to decompose the overall regret as META-REGRET, which measures the regret of the meta algorithm to the best base algorithm, and BASE-REGRET, which measures the best base algorithm to the best elementary action. The main difficulty for bandit problems is that, it is hard to control BASE-REGRET in such a framework due to possible starvation of feedback for the base algorithm. The CORRAL algorithm of Agarwal et al. (2017) addresses this via a new meta algorithm based on Online Mirror Descent (OMD) with the log-barrier regularizer and an increasing learning rate schedule, which together provides a negative term in META-REGRET large enough to (approximately) cancel BASE-REGRET. However, the log-barrier regularizer unavoidably introduces poly(M) dependence in META-REGRET.

1. Switching regret is also known as tracking regret or shifting regret in many previous works (e.g., (Herbster and Warmuth, 1998; Cesa-Bianchi et al., 2012; György and Szepesvári, 2016)).

2. One can compromise and corral $o(T)$ base algorithms instead, which leads to suboptimal switching regret; see such an attempt in (Luo et al., 2018, Appendix G).

Our ideas to address this issue are two-fold. First, to make sure META-REGRET enjoys logarithmic dependence on M, we borrow the idea of the EXP4 algorithm (Auer et al., 2002), which combines M static experts (instead of learning algorithms) without paying polynomial dependence on M. This is achieved by OMD with the negative entropy regularizer, plus a better loss estimator with lower variance for each expert. In our case, this requires coming up with similar low-variance loss estimator for each base algorithm as well as updating each base algorithm no matter whether it is selected by the meta algorithm or not (in contrast, CORRAL only updates the selected base algorithm). Without the log-barrier regularizer, however, we now cannot use the same increasing learning rate schedule as CORRAL to generate a large enough negative term to cancel BASE-REGRET. To address this, our second main idea is to inject negative bias to the loss estimators (making term optimistic underestimators), with the goal of generating a reasonably small positive bias in the regret and at the same time a large enough negative bias to cancel BASE-REGRET. This idea is similar to that of Foster et al. (2020a), but they did not push it far enough and only improved CORRAL on logarithmic factors.

Related work. Since the work of Agarwal et al. (2017), there have been several follow-ups in the same direction, either for adversarial environments (Foster et al., 2020a) or stochastic environments (Pacchiano et al., 2020; Cutkosky et al., 2021; Arora et al., 2021; Krishnamurthy et al., 2021). The problem is also highly related to model selection in online learning with bandit feedback (Foster et al., 2019, 2020b; Marinov and Zimmert, 2021).

The optimal regret for adversarial linear bandits over a general d-dimensional set is $\widetilde{\mathcal{O}}(d\sqrt{T})$ (Dani et al., 2008; Bubeck et al., 2012), but it becomes $\widetilde{\mathcal{O}}(\sqrt{dT})$ for the special case of ℓ_p balls with $p \in [1, 2]$ (Bubeck et al., 2018). To the best of our knowledge, switching regret has not been studied for adversarial linear bandits, except for its special case of multi-armed bandits (Auer et al., 2002; Audibert and Bubeck, 2010). We discuss several natural attempts in Appendix A to extend existing methods to linear bandits, but the best we can get is $\widetilde{\mathcal{O}}(d\sqrt{ST})$ via combining the EXP2 algorithm (Bubeck et al., 2012) and the idea of uniform mixing (Herbster and Warmuth, 1998; Auer et al., 2002). On the other hand, our proposed approach achieves the optimal $\widetilde{\mathcal{O}}(\sqrt{dST})$ regret. In fact, our algorithm is also more computationally efficient as EXP2 requires log-concave sampling.

We assume a known and fixed S in most places. Achieving the same result for *all* S simultaneously is known to be impossible for adaptive adversaries (Marinov and Zimmert, 2021), and remains open for oblivious adversaries (our setting) even for the classic multi-armed bandit problem, so this is beyond the scope of this work. We mention that, however, without knowing S we can still achieve $\widetilde{\mathcal{O}}(S\sqrt{dT})$ regret via a slightly different parameter tuning of our algorithm, or $\widetilde{\mathcal{O}}(\sqrt{dST} + T^{3/4})$ regret via wrapping our algorithm with the generic Bandits-over-Bandits strategy of Cheung et al. (2019). On the other hand, for the easier piecewise stochastic environments, adapting to unknown S without price has been shown possible for different problems including multi-armed bandits (Auer et al., 2019), contextual bandits (Chen et al., 2019), and many more (Wei and Luo, 2021).

Regarding our extension to the unconstrained setting, while unconstrained online learning has been extensively studied in the full-information setting with gradient feedback since the work of Mcmahan and Streeter (2012) (see e.g., (Orabona, 2013; McMahan and Orabona, 2014; Foster et al., 2015; Cutkosky and Boahen, 2017; Cutkosky and Orabona, 2018)), as far as we know (van der Hoeven et al., 2020) is the only existing work considering the same with bandit feedback. They consider static regret and propose a black-box reduction approach, taking inspiration from a similar

Protocol 1 Combining M base algorithms in adversarial linear bandits

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

Each base algorithm $\mathcal{B}_i$ submits an action $\widetilde{a}_t^{(i)} \in \mathcal{X}$ to the meta algorithm, for all $i \in [M]$.
Meta algorithm selects x_t such that $\mathbb{E}_t[x_t] = \sum_{i\in[M]} p_{t,i}\widetilde{a}_t^{(i)}$ for some distribution $p_t \in \Delta_M$.
Play x_t and receive feedback $\ell_t^\top x_t$.
Construct base loss estimator $\widehat{\ell}_t \in \mathbb{R}^d$ and meta loss estimator $\widehat{c}_t \in \mathbb{R}^M$.
Base algorithms $\{\mathcal{B}_i\}_{i=1}^M$ update themselves based on the base loss estimator $\widehat{\ell}_t$.
Meta algorithm updates the weight p_{t+1} based on p_t and the meta loss estimator $\widehat{c}_t$.

end

reduction from the full-information setting (Cutkosky and Orabona, 2018). We consider the more general switching regret, and our algorithm is also built on a similar reduction.

2. Problem Setup and Notations

Problem setup. While our idea is applicable to more general setting, for ease of discussions we focus on the adversarial linear bandit problem throughout the paper. Specifically, at the beginning of a T-round game, an adversary (knowing the learner's algorithm) secretly chooses a sequence of linear loss functions parametrized by $\ell_1, \ldots, \ell_T \in \mathbb{R}^d$. Then, at each round $t \in [T]$, the learner makes a decision by picking a point (also called action) x_t from a known feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$, and subsequently suffers and observes the loss $\ell_t^\top x_t$. Note that $\ell_t^\top x_t$ is the only feedback on ℓ_t revealed to the learner. We measure the learner's performance via the *switching regret*, defined as

$$\textsc{Reg}(u_1, \ldots, u_T) \triangleq \sum_{t=1}^T \ell_t^\top x_t - \sum_{t=1}^T \ell_t^\top u_t = \sum_{k=1}^S \sum_{t\in\mathcal{I}_k} \ell_t^\top (x_t - \mathring{u}_k), \tag{1}$$

where $u_1, \ldots, u_T \in \mathcal{X}$ is a sequence of arbitrary comparators with $S-1$ switches for some known S (that is, $\sum_{t=2}^T \mathbb{1}\{u_{t-1} \neq u_t\} = S - 1$) and $\mathcal{I}_1, \ldots, \mathcal{I}_S$ denotes a partition of $[T]$ such that for each k, u_t remains the same (denoted by $\mathring{u}_k$) for all $t \in \mathcal{I}_k$. Except for comparator-adaptive bounds discussed in Section 5, our results have no explicit dependence on $\mathring{u}_1, \ldots, \mathring{u}_S$ other than the number S, so we often use $\textsc{Reg}_S$ as a shorthand for $\textsc{Reg}(u_1, \ldots, u_T)$. The classic static regret is simply $\textsc{Reg}_1$ (that is, competing with a fixed comparator throughout), which we also simply write as $\textsc{Reg}$.

Notations. For any integer n, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$, and Δ_n the simplex $\{p \in \mathbb{R}^n_{\geq 0} \mid \sum_{i=1}^n p_i = 1\}$. We use e_i to denote the standard basis vector (of appropriate dimension) with the i-th coordinate being 1 and others being 0. Given a vector $x \in \mathbb{R}^d$, its ℓ_p norm is defined by $\|x\|_p = (\sum_{n=1}^d |x_n|^p)^{1/p}$. $\mathbb{E}_t[\cdot]$ denotes the conditional expectation given the history before round t. The $\widetilde{\mathcal{O}}(\cdot)$ notation omits the logarithmic dependence on the time horizon T and the dimension d. For a differential convex function $\psi : \mathbb{R}^d \mapsto \mathbb{R}$, the induced Bregman divergence is defined by $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$.

3. Corralling a Larger Band of Bandits: A Recipe

In this section, we describe our general recipe to corral a large set of bandit algorithms. We start by showing a general and natural protocol of such a meta-base framework in Protocol 1. Specifically, suppose we maintain M base algorithms $\{\mathcal{B}_i\}_{i=1}^M$. At the beginning of each round, each base algorithm $\mathcal{B}_i$ submits its own action $\widetilde{a}_t^{(i)} \in \mathcal{X}$ to the meta algorithm, which then decides the final action x_t with expectation $\sum_{i\in[M]} p_{t,i}\widetilde{a}_t^{(i)}$ for some distribution $p_t \in \Delta_M$ specifying the importance/quality of each base algorithm. After playing x_t and receiving the feedback $\ell_t^\top x_t$, we construct *base loss estimator* $\widehat{\ell}_t \in \mathbb{R}^d$ and *meta loss estimator* $\widehat{c}_t \in \mathbb{R}^M$. As their name suggests, base loss estimator estimates ℓ_t and is used to update each base algorithm, while meta loss estimator estimates $A_t^\top \ell_t$, where the i-th column of $A_t \in \mathbb{R}^{d\times M}$ is $\widetilde{a}_t^{(i)}$, and is used to update the meta algorithm to obtain the next distribution $p_{t+1} \in \Delta_M$.

In the following, we formalize the high-level idea discussed in Section 1. For simplicity, we focus on the static regret REG in this discussion (that is, $S = 1$) and let u be the fixed comparator. The first step is to decompose the regret into two parts as mentioned in Section 1: as long as $\widehat{\ell}_t$ and $\widehat{c}_t$ are unbiased estimators (that is, $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$ and $\mathbb{E}_t[\widehat{c}_t] = A_t^\top \ell_t$), one can show:

$$\forall j \in [M], \quad \mathbb{E}[\textsc{Reg}] = \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle p_t - e_j, \widehat{c}_t\rangle\right]}_{\textsc{Meta-Regret}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \left\langle \widetilde{a}_t^{(j)} - u, \widehat{\ell}_t\right\rangle\right]}_{\textsc{Base-Regret}}. \tag{2}$$

Controlling BASE-REGRET is the key challenge. Indeed, even if the base algorithm enjoys a good regret guarantee when running on its own, it might not ensure the same guarantee any more in this meta-base framework because it cannot fully control the final action and observe the feedback it needs. At a technical level, this is reflected in a larger variance of $\widehat{\ell}_t$ due to the randomness from the meta algorithm, which then ruins the base algorithm's original regret guarantee.

As mentioned, the way CORRAL (Agarwal et al., 2017) addresses this issue is by using OMD with the log-barrier regularizer and increasing learning rates as the meta algorithm, which ensures that META-REGRET is at most $\widetilde{\mathcal{O}}(\sqrt{MT})$ plus some *negative* term large enough to cancel the prohibitively large part of BASE-REGRET. The poly(M) dependence in their approach is unavoidable because they treat the problem that the meta algorithm is facing as a classic multi-armed bandit problem and ignores the fact that information can be shared among different base algorithms. The recent follow-up (Foster et al., 2020a) shares the same issue.

Instead, we propose the following idea. We use OMD with entropy regularizer (a.k.a. multiplicative weights update) as the meta algorithm to update p_{t+1}, usually in the form $p_{t+1,i} \propto p_{t,i}e^{-\varepsilon\widehat{c}_{t,i}}$ where $\varepsilon > 0$ is some learning rate. This first ensures that the so-called regularization penalty term in META-REGRET is of order $\frac{\log M}{\varepsilon}$ instead of $\frac{M}{\varepsilon}$ as in CORRAL. To control the other so-called stability term in META-REGRET, the estimator $\widehat{c}_t$ has to be constructed in a way with low variance, but we defer the discussion and first look at how to control BASE-REGRET in this case. Since we are no longer using the log-barrier regularizer of CORRAL, a different way to generate a large negative term in META-REGRET to cancel BASE-REGRET is needed. To this end, we propose to inject a (negative) bias $b_t \in \mathbb{R}_+^M$ to the meta loss estimator $\widehat{c}_t$, making it an optimistic underestimator. More specifically, introduce another notation c_t for some unbiased estimator of $A_t^\top \ell_t$. Then the adjusted meta loss estimator is defined as $\widehat{c}_t = c_t - b_t$. Since $\widehat{c}_t$ is biased now, the decomposition (2) needs to be updated accordingly as

$$\mathbb{E}[\textsc{Reg}] = \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle p_t - e_j, \widehat{c}_t\rangle\right]}_{\textsc{Meta-Regret}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \left\langle \widetilde{a}_t^{(j)} - u, \widehat{\ell}_t\right\rangle\right]}_{\textsc{Base-Regret}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle p_t, b_t\rangle\right]}_{\textsc{Pos-Bias}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle e_j, b_t\rangle\right]}_{\textsc{Neg-Bias}}.$$

Based on this decomposition, our goal boils down to designing good base and meta loss estimators such that the following three terms are all well controlled:

$$\text{BASE-REGRET} - \text{NEG-BIAS} \leq \text{TARGET}, \tag{3}$$
$$\text{POS-BIAS} \leq \text{TARGET}, \tag{4}$$
$$\text{META-REGRET} \leq \text{TARGET}. \tag{5}$$

Here, TARGET represents the final targeted regret bound with logarithmic dependence on M and usually $\sqrt{T}$-dependence on T (such as $\widetilde{\mathcal{O}}(\sqrt{dST \log M})$ for our main application of switching regret discussed in Section 4).[3]

A recipe. We are now ready to summarize our recipe in the following three steps.

- **Step 1.** Start from designing $\widehat{\ell}_t$, which often follows similar ideas of the original base algorithm.
- **Step 2.** Then, by analyzing BASE-REGRET with such a base loss estimator, figure out what b_t needs to be in order to ensure Eq. (3) and Eq. (4) simultaneously.
- **Step 3.** Finally, design c_t to ensure Eq. (5). As mentioned in Section 1, this is a problem similar to combining static experts as in the EXP4 algorithm (Auer et al., 2002), and the key is to ensure that c_t allows information sharing among base algorithms and enjoys low variance. A natural choice is $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \widehat{\ell}_t \rangle$, which is exactly what EXP4 does and works in the toy example we show below, but sometimes one needs to replace $\widehat{\ell}_t$ with yet another better unbiased estimator of ℓ_t, which turns out to be indeed the case for our main example in Section 4.

A toy example. Now, we provide a warm-up example to show how to successfully apply our three-step recipe to the multi-armed bandit problem. We note that this example does *not* really lead to meaningful applications, as in the end we are simply combining different copies of the exact same algorithm. Nevertheless, this serves as a simple and illustrating exercise to execute our recipe, paving the way for the more complicated scenario to be discussed in the next section.

Specifically, in multi-armed bandit, we have $\mathcal{X} = \Delta_d$ and $\ell_t \in [0,1]^d$ for all $t \in [T]$, and we set the target to be TARGET $= \widetilde{\mathcal{O}}(\sqrt{dT \log M})$ (optimal up to logarithmic factors). The meta algorithm is as specified before (multiplicative weights update). For the base algorithm, we choose a slight variant of the classic EXP3 algorithm (Auer et al., 2002), so that $\widetilde{a}_{t+1}^{(i)} = \operatorname{argmin}_{a \in \Delta_d \cap [\eta,1]^d} \left\{ \langle a, \widehat{\ell}_t \rangle + \frac{1}{\eta} D_\psi(a, a_t^{(i)}) \right\}$, where $\eta > 0$ is a clipping threshold (and also a learning rate) and $\psi(a) = \sum_{n=1}^d a_n \log a_n$ is the negative entropy. Given $q_t = \sum_{i=1}^M p_{t,i} \widetilde{a}_t^{(i)} \in \Delta_d$, the meta algorithm naturally samples an arm $n_t \in [d]$ according to q_t, meaning $x_t = e_{n_t}$.

Step 1. With the feedback $\ell_t^\top x_t = \ell_{t,n_t}$, following EXP3 we let the base loss estimator be the standard importance-weighted estimator: $\widehat{\ell}_t = \frac{\ell_{t,n_t}}{q_{t,n_t}} x_t$, which is clearly unbiased with $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$.

3. We present our recipe with a targeted regret bound just because this is indeed the case for our main application (of getting switching regret bounds), but this is not necessary for our approach, and we can for example also achieve some of the applications discussed in (Agarwal et al., 2017) where the meta algorithm achieves different regret bounds (that is, not a single target) in different environments.

Step 2. By standard analysis (e.g., (Bubeck and Cesa-Bianchi, 2012, Theorem 3.1)), BASE-REGRET is at most $\eta dT + \frac{\log d}{\eta} + \eta\mathbb{E}\big[\sum_{t=1}^{T}\sum_{n=1}^{d}\widetilde{a}_{t,n}^{(j)}\widehat{\ell}_{t,n}^2\big]$. Since $\mathbb{E}_t[\widehat{\ell}_{t,n}^2] = \frac{\ell_{t,n}^2}{q_{t,n}}$, the last term is further bounded by $\eta\mathbb{E}\big[\sum_{t=1}^{T}\sum_{n=1}^{d}\widetilde{a}_{t,n}^{(j)}/q_{t,n}\big]$. This is exactly the problematic stability term that can be prohibitively large. We thus directly define the bias term $b_{t,j}$ as $\eta\sum_{n=1}^{d}\widetilde{a}_{t,n}^{(j)}/q_{t,n}$, so that BASE-REGRET − NEG-BIAS is simply bounded by $\eta dT + \frac{\log d}{\eta}$. Picking the optimal η ensures Eq. (3). On the other hand, POS-BIAS happens to be small as well: POS-BIAS $= \mathbb{E}\big[\sum_{t=1}^{T}\langle p_t, b_t\rangle\big] = \eta\mathbb{E}\big[\sum_{t=1}^{T}\sum_{i=1}^{M}p_{t,i}\sum_{n=1}^{d}\widetilde{a}_{t,n}^{(i)}/q_{t,n}\big] = \eta\mathbb{E}\big[\sum_{t=1}^{T}\sum_{n=1}^{d}q_{t,n}/q_{t,n}\big] = \eta dT$, ensuring Eq. (4).

Step 3. Finally, we use the natural meta loss estimator $c_{t,i} = \langle\widetilde{a}_t^{(i)}, \widehat{\ell}_t\rangle$. Since $q_{t,n} \geq \eta$ due to the clipping threshold and thus $0 \leq b_{t,i} \leq 1$ and $\widehat{c}_{t,i} \geq -1$ (that is, not too negative), standard analysis shows META-REGRET $\leq \frac{\log M}{\varepsilon} + \varepsilon\mathbb{E}\big[\sum_{t=1}^{T}\sum_{i=1}^{M}p_{t,i}\widehat{c}_{t,i}^2\big]$, with the last term further bounded by $2\varepsilon\mathbb{E}\big[\sum_{t=1}^{T}\sum_{i=1}^{M}(p_{t,i}c_{t,i}^2 + p_{t,i}b_{t,i}^2)\big] \leq 4\varepsilon dT$. Picking the optimal ε in the final bound META-REGRET $\leq \frac{\log M}{\varepsilon} + 4\varepsilon dT$ then ensures Eq. (5).

This concludes our example and shows that our recipe indeed enjoys logarithmic dependence on M in this case, which CORRAL fails to achieve. At a high level, our method avoids the poly(M) dependence by information sharing among different base algorithms: we update every base algorithm no matter whether it is selected by the meta algorithm or not, while in contrast, CORRAL only updates the selected base algorithm, which makes the poly(M) dependence unavoidable.

4. Optimal Switching Regret for Linear Bandits over ℓ_p Balls

As the main application in this work, we now discuss how to apply our recipe to achieve the optimal switching regret for adversarial linear bandits over ℓ_p balls. In this problem, the feasible domain is an ℓ_p unit ball for some $p \in (1, 2]$, namely, $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_p \leq 1\}$, and each ℓ_t is assumed to be from the dual ℓ_q unit ball with $q = p/(p-1)$, such that $|\ell_t^\top x| \leq 1$ for all $x \in \mathcal{X}$ and $t \in [T]$. Bubeck et al. (2018) show that the optimal regret in this case is $\Theta(\sqrt{dT})$, which is better than the general linear bandit problem by a factor of $\sqrt{d}$. This implies that the optimal switching regret for this problem is $\Omega(\sqrt{dST})$ — indeed, simply consider the case where $\mathcal{I}_1, \ldots, \mathcal{I}_S$ is an even partition of $[T]$ and the adversary forces the learner to suffer $\Omega(\sqrt{d|\mathcal{I}_k|}) = \Omega(\sqrt{dT/S})$ regret on each interval $\mathcal{I}_k$ by generating a new worst case instance regarding the static regret. Therefore, our target regret bound here is set to TARGET $= \widetilde{\mathcal{O}}(\sqrt{dST})$. We remind the reader that this problem has not been studied before and that in Appendix A, we discuss other potential approaches and why none of them is able to achieve this goal.

The pseudocode of our final algorithm is shown in Algorithm 2. At a high-level, it is simply following the standard idea in the literature on obtaining switching regret, that is, maintain a set of $M = T$ base algorithms with static regret guarantees, the t-th of which $\mathcal{B}_t$ starts learning from time step t (before time t, one pretends that $\mathcal{B}_t$ picks the same action as the meta algorithm). If the meta algorithm itself enjoys a switching regret guarantee,[4] then by competing with $\mathcal{B}_{j_k}$ on interval $\mathcal{I}_k$ where j_k is the first time step of $\mathcal{I}_k$ so that $\mathcal{B}_{j_k}$ enjoys a (static) regret guarantee on $\mathcal{I}_k$, the overall algorithm enjoys a switching regret for the original problem. While this is a standard and simple

4. We point out that in the full-information setting, even a certain static regret guarantee from the meta algorithm is enough, but a switching regret guarantee is needed in the bandit setting for technical reasons.

idea, applying it to the bandit setting was not possible before our work due to the large number of base algorithms (T) needed to be combined. Our approach, however, is able to overcome this with logarithmic dependence on M, making it the first successful execution of this long-standing idea in bandit problems.

Base algorithm overview. Our base algorithm is naturally the one proposed in (Bubeck et al., 2018) that achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ static regret.[5] Specifically, let $\mathcal{X}' = \{x \mid \|x\|_p \leq 1-\gamma\}$ for some clipping parameter γ be a slightly smaller ball. At each round t, each base algorithm $\mathcal{B}_i$ (for $i \leq t$) has a vector $a_t^{(i)} \in \mathcal{X}'$ at hand. Then, it generates a Bernoulli random variable $\xi_t^{(i)}$ with mean $\|a_t^{(i)}\|_p$. If $\xi_t^{(i)} = 0$, then its final decision $\widetilde{a}_t^{(i)}$ is uniformly sampled from $\{\pm e_n\}_{n=1}^d$; otherwise, $\widetilde{a}_t^{(i)} = a_t^{(i)}/\|a_t^{(i)}\|_p$. Next, $\mathcal{B}_i$ submits $(\widetilde{a}_t^{(i)}, a_t^{(i)}, \xi_t^{(i)})$ to the meta algorithm. After receiving the base loss estimator $\widehat{\ell}_t$ (to be specified later), $\mathcal{B}_i$ updates $a_{t+1}^{(i)}$ using OMD with the regularizer $R(a) = -\log(1-\|a\|_p^p)$, that is, $a_{t+1}^{(i)} = \operatorname{argmin}_{a\in\mathcal{X}'} \left\{\langle a, \widehat{\ell}_t\rangle + \frac{1}{\eta} D_R(a, a_t^{(i)})\right\}$ for some learning rate $\eta > 0$. We defer the pseudocode Algorithm 5 to Appendix B.1.

Meta algorithm overview. The meta algorithm maintains the distribution $p_t \in \Delta_T$ again via multiplicative weights update, but since a switching regret guarantee is required as mentioned, a slight variant studied in (Auer et al., 2002) is needed which mixes the multiplicative weights update with a uniform distribution: $p_{t+1,i} = (1-\mu)\frac{p_{t,i}\exp(-\varepsilon \widehat{c}_{t,i})}{\sum_{j=1}^T p_{t,j}\exp(-\varepsilon \widehat{c}_{t,j})} + \frac{\mu}{T}$ for some mixing rate μ, learning rate ε, and meta loss estimator $\widehat{c}_t$ (to be specified later). As mentioned, at time t, all base algorithm $\mathcal{B}_i$ with $i > t$ should be thought of as making the same decision as the meta algorithm, so in a sense we are looking for an action $\widetilde{x}_t$ such that $\widetilde{x}_t = \sum_{i=1}^t p_{t,i}\widetilde{a}_t^{(i)} + \sum_{i=t+1}^T p_{t,i}\widetilde{x}_t$, or equivalently $\widetilde{x}_t = \sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}$ with a distribution $\widehat{p}_t \in \Delta_t$ satisfying $\widehat{p}_{t,i} \propto p_{t,i}$. Combining this with some extra exploration for technical reasons, the final decision x_t of our algorithm is decided as follows: sample a Bernoulli random variable ρ_t with mean β (a small parameter); if $\rho_t = 1$, then x_t is uniformly sampled from $\{\pm e_n\}_{n=1}^d$, otherwise x_t is sampled from $\widetilde{a}_t^{(1)}, \ldots, \widetilde{a}_t^{(t)}$ according to the distribution $\widehat{p}_t$. See Line 3, Line 4, and Line 9. We are now ready to follow the three steps of our recipe to design the loss estimators.

Step 1. The design of the base loss estimator $\widehat{\ell}_t$ mostly follows (Bubeck et al., 2018), except for the extra consideration due to the sampling scheme of the meta algorithm (Line 4). The final form is shown in Eq. (6), and a direct calculation verifies its unbiasedness $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$ (see Lemma 5).

Step 2. With $\widehat{\ell}_t$ fixed, for an interval $\mathcal{I}_k$, we analyze the static regret of $\mathcal{B}_{j_k}$ on this interval (recall that j_k is the first time step of $\mathcal{I}_k$), mostly following the analysis of Bubeck et al. (2018). This corresponds to BASE-REGRET (since we have moved from static regret to switching regret). More concretely, in Lemma 7 we show for some universal constant $C_1 > 0$:

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}_k}\left\langle \widetilde{a}_t^{(j_k)} - \mathring{u}_k, \widehat{\ell}_t\right\rangle\right] \leq \frac{\log(1/\gamma)}{\eta} + \eta C_1 \sum_{t\in\mathcal{I}_k}\frac{1-\|a_t^{(j_k)}\|_p}{1-\sum_{j=1}^t \widehat{p}_{t,j}\|a_t^{(j)}\|_p}.$$

Again, the second term above is the prohibitively large term, and we thus define $b_{t,i}$ in the same form; see Eq. (8). As long as the parameters are chosen such that $\eta C_1 \leq \frac{1}{\lambda T(1-\beta)}$, BASE-REGRET $-$

5. To be more accurate, the version we present here is a slightly simpler variant with the same guarantee.

Algorithm 2 Algorithm for adversarial linear bandits over ℓ_p balls with optimal switching regret

Input: clipping parameter γ, base learning rate η, meta learning rate ε, mixing rate μ, exploration parameter β, bias coefficient λ, initial uniform distribution $p_1 \in \Delta_T$.

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

Start a new base algorithm $\mathcal{B}_t$, which is an instance of Algorithm 5 with learning rate η, clipping parameter γ, and initial round t.

Receive local decision $(\widetilde{a}_t^{(i)}, a_t^{(i)}, \xi_t^{(i)})$ from base algorithm $\mathcal{B}_i$ for each $i \leq t$.

Compute the renormalized distribution $\widehat{p}_t \in \Delta_t$ such that $\widehat{p}_{t,i} \propto p_{t,i}$ for $i \in [t]$.

Sample a Bernoulli random variable ρ_t with mean β. If $\rho_t = 1$, uniformly sample x_t from $\{\pm e_n\}_{n=1}^d$; otherwise, sample $i_t \in [t]$ according to $\widehat{p}_t$, and set $x_t = \widetilde{a}_t^{(i_t)}$ and $\xi_t = \xi_t^{(i_t)}$.

Make the final decision x_t and receive feedback $\ell_t^\top x_t$.

Construct the base loss estimator $\widehat{\ell}_t \in \mathbb{R}^d$ as follows and send it to all base algorithms $\{\mathcal{B}_i\}_{i=1}^t$:

$$\widehat{\ell}_t = \frac{\mathbb{1}\{\rho_t = 0\}\mathbb{1}\{\xi_t = 0\}}{1-\beta} \cdot \frac{d(\ell_t^\top x_t)}{1 - \sum_{i=1}^t \widehat{p}_{t,i} \|a_t^{(i)}\|_p} \cdot x_t. \tag{6}$$

Construct another loss estimator $\bar{\ell}_t \in \mathbb{R}^d$ as

$$\bar{\ell}_t = \widetilde{M}_t^{-1} x_t x_t^\top \ell_t, \tag{7}$$

where $\widetilde{M}_t = \frac{\beta}{d}\sum_{n=1}^d e_n e_n^\top + (1-\beta)\sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}\widetilde{a}_t^{(i)\top}$.

Construct the meta loss estimator $\widehat{c}_t \in \mathbb{R}^T$ as:

$$\widehat{c}_{t,i} = \begin{cases} \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle - b_{t,i}, & i \leq t, \\ \sum_{j=1}^t \widehat{p}_{t,j}\widehat{c}_{t,j}, & i > t, \end{cases} \quad \text{where } \ b_{t,i} = \frac{1}{\lambda T(1-\beta)} \frac{1 - \|a_t^{(i)}\|_p}{1 - \sum_{j=1}^t \widehat{p}_{t,j}\|a_t^{(j)}\|_p}. \tag{8}$$

Meta algorithm updates the weight $p_{t+1} \in \Delta_T$ according to

$$p_{t+1,i} = (1-\mu)\frac{p_{t,i}\exp(-\varepsilon \widehat{c}_{t,i})}{\sum_{j=1}^T p_{t,j}\exp(-\varepsilon \widehat{c}_{t,j})} + \frac{\mu}{T}, \quad \forall i \in [T]. \tag{9}$$

end

NEG-BIAS is simply bounded by $\frac{\log(1/\gamma)}{\eta}$, and Eq. (3) can be ensured. Direct calculation shows that with such a bias term $b_{t,i}$, POS-BIAS is also small enough to ensure Eq. (4); see Appendix B.4.

Step 3. Finally, it remains to design unbiased loss estimator $c_{t,i}$ and finalize the meta loss estimator $\widehat{c}_{t,i}$. As mentioned, a natural choice would be $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \widehat{\ell}_t \rangle$. However, despite its unbiasedness, it turns out to suffer a large variance in this case and cannot lead to a favorable guarantee for META-REGRET. To address this issue, we introduce yet another unbiased loss estimator $\bar{\ell}_t$ for ℓ_t, defined in Eq. (7), which follows standard idea from the general linear bandit literature (see for example the EXP2 algorithm of Bubeck et al. (2012)). With that, $c_{t,i}$ is defined as $\langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle$ instead, which now has a small enough variance. We find it intriguing that using different unbiased loss estimators ($\widehat{\ell}_t$ for base algorithms and $\bar{\ell}_t$ for the meta algorithm) for the same quantity ℓ_t appears to be necessary for this problem. As the final piece of the puzzle, we set $\widehat{c}_{t,i} = c_{t,i} - b_{t,i}$ for $i \le t$ as our recipe describes, and for $i > t$, recall that these base algorithms are thought of as making the same prediction of the meta algorithm, thus we set $\widehat{c}_{t,i} = \sum_{j=1}^{t} \widehat{p}_{t,j} \widehat{c}_{t,j}$; see Eq. (8). This ensures an important property $\langle p_t, \widehat{c}_t \rangle = \sum_{i \le t} \widehat{p}_{t,i} \widehat{c}_{t,i}$, which we use to finally prove that META-REGRET is small enough to ensure Eq. (5) (see Lemma 10).

This concludes the description of our entire algorithm. Algorithm 2 summarizes the main update procedures. We briefly discuss the computational and space complexity. Note that each base algorithm performs a simple OMD update with a barrier regularizer, and it suffices to obtain an approximate solution with $\frac{1}{T}$ precision, which takes $\mathcal{O}(\mathrm{poly}(d))$ space and $\mathcal{O}(\mathrm{poly}(d \log T))$ time via for example the Interior Point Method. On the other hand, the computational/space complexity of the meta algorithm is clearly $\mathcal{O}(T \,\mathrm{poly}(d))$ per round.

We formally prove in Appendix B that our algorithm enjoys the following switching regret.

Theorem 1 *Define $C = \sqrt{p-1} \cdot 2^{-\frac{2}{p-1}}$. With parameters $\gamma = 4C\sqrt{\frac{dS}{T}}$, $\eta = C\sqrt{\frac{S}{dT}}$, $\varepsilon = \min\left\{\sqrt{\frac{S}{dT}}, \frac{1}{16d}, \frac{C^2}{2}\right\}$, $\mu = \frac{1}{T}$, $\beta = 8d\varepsilon$, and $\lambda = \frac{C}{\sqrt{dST}}$, Algorithm 2 guarantees*

$$\mathbb{E}[\mathrm{REG}_S] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^\top x_t - \sum_{t=1}^{T} \ell_t^\top u_t\right] = \widetilde{\mathcal{O}}\left(\sqrt{dST}\right),$$

where $u_1, \ldots, u_T \in \mathcal{X}$ are arbitrary comparators such that $\sum_{t=2}^{T} \mathbb{1}\{u_{t-1} \neq u_t\} \le S - 1$.

We point out again that this is the first optimal switching regret guarantee for linear bandits over ℓ_p balls with $p \in (1, 2]$, demonstrating the importance of our new corralling method.

Extensions to smooth and strongly convex domain. Our ideas and results can be generalized to adversarial linear bandits over any smooth and strongly convex set, a setting studied in (Kerdreux et al., 2021). Specifically, for a smooth and strongly convex set containing the ℓ_p unit ball and contained by the dual ℓ_q unit ball (for some $p \in (1, 2]$), our algorithm achieves $\widetilde{\mathcal{O}}\big(d^{1/p}\sqrt{ST}\big)$ switching regret. We defer all details to Appendix C.

5. Extension to Unconstrained Linear Bandits

In this section, we further extend our results on linear bandits to the *unconstrained* setting, that is, $\mathcal{X} = \mathbb{R}^d$, which means both the learner's decisions $\{x_t\}_{t=1}^T$ and the comparators $\{u_t\}_{t=1}^T$ can be chosen arbitrarily in $\mathbb{R}^d$. The loss vectors are assumed to have bounded ℓ_2 norm: $\|\ell_t\|_2 \le 1$ for all $t \in [T]$. As mentioned, (van der Hoeven et al., 2020) is the only existing work considering the same setting. They study static regret and achieve a comparator-adaptive bound $\mathbb{E}[\mathrm{REG}] = \widetilde{\mathcal{O}}(\|u\|_2\sqrt{dT})$ simultaneously for all u (the fixed comparator).[6] Building on our results in Section 4, we generalize their static regret bound to switching regret and achieve a similar comparator-adaptive bound $\mathbb{E}[\mathrm{REG}(u_1, \ldots, u_T)] = \widetilde{\mathcal{O}}\big(\max_{k \in [S]} \|\mathring{u}_k\|_2 \cdot \sqrt{dST}\big)$ simultaneously for all $u_1, \ldots, u_T$ with $S - 1$ switches.

6. The actual bound stated in (van der Hoeven et al., 2020) is actually $\widetilde{\mathcal{O}}(\|u\|_2 d\sqrt{T})$, but it is straightforward to see that it can be improved to $\widetilde{\mathcal{O}}(\|u\|_2\sqrt{dT})$ by picking the optimal linear bandit algorithm over ℓ_2 balls in their reduction.

Algorithm 3 Comparator-adaptive algorithm for unconstrained linear bandits

Input: subroutine $\mathcal{A}_{\mathcal{V}}$ (unconstrained OCO algorithm), subroutine $\mathcal{A}_{\mathcal{Z}}$ (constrained linear bandits algorithm), $\mathcal{Z} = \{z \mid \|z\|_2 \leq 1\}$.

for $t = 1$ ***to*** T **do**

Receive the direction $z_t \in \mathcal{Z}$ from subroutine $\mathcal{A}_{\mathcal{Z}}$.

Receive the magnitude $v_t \in \mathbb{R}$ from subroutine $\mathcal{A}_{\mathcal{V}}$.

Submit $x_t = z_t \cdot v_t$ and receive and observe the loss $\ell_t^\top x_t$.

Send $\ell_t^\top z_t = \ell_t^\top x_t / v_t$ as the feedback for subroutine $\mathcal{A}_{\mathcal{Z}}$.

Construct linear function $f_t(v) \triangleq v \cdot \ell_t^\top z_t$ as the feedback for subroutine $\mathcal{A}_{\mathcal{V}}$.

end

Instead of using our recipe and starting from scratch to solve this problem, we directly make use of the reduction of van der Hoeven et al. (2020) that reduces the unconstrained problem to the constrained counterpart (already solved by our Algorithm 2) plus another one-dimensional unconstrained problem; see Section 5.1. To solve the latter problem, in Section 5.2 we design a new unconstrained algorithm for general Online Convex Optimization (OCO) that enjoys a comparator-adaptive switching regret guarantee and might be of independent interest. Finally, we summarize the overall algorithm and provide the formal guarantees in Section 5.3.

5.1. Black-box reduction for switching regret of unconstrained linear bandits

The reduction of van der Hoeven et al. (2020) takes heavy inspiration from (Cutkosky and Orabona, 2018). Specifically, suppose that we have two subroutines denoted by $\mathcal{A}_{\mathcal{Z}}$ and $\mathcal{A}_{\mathcal{V}}$: $\mathcal{A}_{\mathcal{Z}}$ is a constrained linear bandit algorithm over the ℓ_2 ball $\mathcal{Z} = \{z \in \mathbb{R}^d \mid \|z\|_2 \leq 1\}$ and $\mathcal{A}_{\mathcal{V}}$ is an unconstrained and one-dimensional online linear optimization algorithm with full-information feedback (in fact, in the one-dimensional linear case, there is no difference between full-information and bandit feedback). Then, one can solve an unconstrained linear bandit problem as follows: at each round $t \in [T]$, the learner makes the decision $x_t = v_t \cdot z_t$, where $z_t \in \mathcal{Z}$ is the direction returned by the constrained bandit algorithm $\mathcal{A}_{\mathcal{Z}}$, and $v_t \in \mathbb{R}$ is the scalar returned by the one-dimensional algorithm $\mathcal{A}_{\mathcal{V}}$. After observing the loss $\ell_t^\top x_t$, the learner then feeds $\ell_t^\top z_t = \frac{\ell_t^\top x_t}{v_t}$ to both $\mathcal{A}_{\mathcal{Z}}$ and $\mathcal{A}_{\mathcal{V}}$ so they can update themselves. See Algorithm 3 for the pseudocode.

van der Hoeven et al. (2020) show that the static regret of such a reduction can be expressed using the regret of the two subroutines. This can be directly generalized to switching regret, formally described below (see Appendix D.1 for the proof).

Lemma 2 *For an interval $\mathcal{I} \subseteq [T]$, let $\textsc{Reg}_{\mathcal{I}}^{\mathcal{V}}(v) = \sum_{t\in\mathcal{I}}(v_t - v)\langle z_t, \ell_t\rangle$ be the regret of the unconstrained one-dimensional algorithm $\mathcal{A}_{\mathcal{V}}$ against a comparator $v \in \mathbb{R}$ on this interval, and similarly $\textsc{Reg}_{\mathcal{I}}^{\mathcal{Z}}(z) = \sum_{t\in\mathcal{I}}\langle z_t - z, \ell_t\rangle$ be the regret of the constrained linear bandits algorithm $\mathcal{A}_{\mathcal{Z}}$ against a comparator $z \in \mathcal{Z} = \{z \in \mathbb{R}^d \mid \|z\|_2 \leq 1\}$ on this interval. Then Algorithm 3 (with decision $x_t = z_t \cdot v_t$) satisfies*

$$\textsc{Reg}(u_1, \ldots, u_T) = \sum_{k=1}^{S} \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{V}}(\|\mathring{u}_k\|_2) + \sum_{k=1}^{S} \|\mathring{u}_k\|_2 \cdot \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{Z}}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right), \tag{10}$$

where we recall that $\mathcal{I}_1, \ldots, \mathcal{I}_S$ denotes a partition of $[T]$ such that for each k, u_t remains the same (denoted by $\mathring{u}_k$) for all $t \in \mathcal{I}_k$.

One can see that the first term in Eq. (10) is clearly the switching regret of $\mathcal{A}_{\mathcal{V}}$, while the second term, after upper bounded by $\max_{k\in[S]}\|\mathring{u}_k\|_2 \sum_{k=1}^{S} \text{REG}^{\mathcal{Z}}_{\mathcal{I}_k}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right)$, is the switching regret of $\mathcal{A}_{\mathcal{Z}}$ scaled by the maximum comparator norm. Therefore, to control the second term, we simply apply our Algorithm 2 as the subroutine $\mathcal{A}_{\mathcal{Z}}$, making it at most $\widetilde{\mathcal{O}}\big(\max_{k\in[S]}\|\mathring{u}_k\|_2 \cdot \sqrt{dST}\big)$. On the other hand, to the best of our knowledge, there are no existing unconstrained algorithms with switching regret guarantees. To this end, we design one such algorithm in the next section. In fact, for full generality, we do so for the more general unconstrained OCO problem of arbitrary dimension *without the knowledge of* S, which might be of independent interest.

5.2. Subroutine: switching regret of unconstrained online convex optimization

As a slight detour, in this section we consider a general unconstrained OCO problem: at round $t \in [T]$, the learner makes a decision $v_t \in \mathbb{R}^d$ and simultaneously the adversary chooses a loss function $f_t : \mathbb{R}^d \mapsto \mathbb{R}$, then the algorithm suffers loss $f_t(v_t)$ and observes the gradient $\nabla f_t(v_t)$ as feedback. Notably, the feasible domain is $\mathbb{R}^d$ (that is, no constraints). The goal of the learner is to minimize the switching regret

$$\text{REG}(u_1, \ldots, u_T) = \sum_{t=1}^{T} f_t(v_t) - \sum_{t=1}^{T} f_t(u_t) = \sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\Big(f_t(v_t) - f_t(\mathring{u}_k)\Big), \tag{11}$$

where the notations $\mathcal{I}_1, \ldots, \mathcal{I}_S$ and $\mathring{u}_1, \ldots, \mathring{u}_S \in \mathbb{R}^d$ are defined similarly as in Section 2. Without loss of generality, it is assumed that $\max_x \|\nabla f_t(x)\|_2 \le 1$ for all t. Note that this setup is a strict generalization of what we need for the one-dimensional subroutine $\mathcal{A}_{\mathcal{V}}$ discussed in Section 5.1.

Our idea is once again via a meta-base framework, which is in fact easier than our earlier discussions because now we have gradient feedback. There are two quantities that we aim to adapt to: the number of switches S and the comparator norm $\|\mathring{u}_k\|_2$ (although the latter can be unbounded, it suffices to consider a maximum norm of 2^T as (Chen et al., 2021, Appendix D.5) shows). Therefore, we create an exponential grid for these two quantities, and maintain one base algorithm for each possible configuration. These base algorithms only need to satisfy some mild conditions specified in Requirement 1 of Appendix D.2, and many existing algorithms such as (Daniely et al., 2015; Jun et al., 2017; Zhang et al., 2019; Cutkosky, 2020) indeed meet the requirements.

The design of the meta algorithm requires some care to ensure the desirable adaptive guarantees, and we achieve so by building upon the recent progress in the classic expert problem (Chen et al., 2021). In short, our meta algorithm is OMD with a multi-scale entropy regularizer and certain important correction terms. We defer the details to Appendix D.2 and only present the pseudocode of the full algorithm in Algorithm 4. Below we present the main comparator-adaptive switching regret guarantee of this algorithm.

Theorem 3 *Algorithm 4 with a base algorithm satisfying Requirement 1 guarantees that for any S, any partition $\mathcal{I}_1, \ldots, \mathcal{I}_S$ of $[T]$, and any comparator sequence $\mathring{u}_1, \ldots, \mathring{u}_S \in \mathbb{R}^d$, we have*

$$\sum_{k=1}^{S}\left(\sum_{t\in\mathcal{I}_k} f_t(v_t) - \sum_{t\in\mathcal{I}_k} f_t(\mathring{u}_k)\right) \le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S}\|\mathring{u}_k\|_2\sqrt{|\mathcal{I}_k|}\right) \le \widetilde{\mathcal{O}}\left(\max_{k\in[S]}\|\mathring{u}_k\|_2 \cdot \sqrt{ST}\right).$$

We emphasize again that in contrast to our other results on bandit problems, the guarantee above is achieved for all S simultaneously (in other words, the algorithm does not need the knowledge

Algorithm 4 Comparator-adaptive algorithm for unconstrained OCO

Input: base algorithm $\mathfrak{B}$.

Define: $H = \lceil \log_2 T \rceil + T + 1$ and $R = \lceil \log_2 T \rceil$.

Define: clipped domain $\Omega = \{w \mid w \in \Delta_N \text{ and } w_{t,(i,r)} \geq \frac{1}{T^2 \cdot 2^{2i}}, \forall i \in [H], r \in [R]\}$.

Define: weighted entropy regularizer $\psi(w) \triangleq \sum_{(i,r)\in[H]\times[R]} \frac{c_i}{\eta_r} w_{(i,r)} \log w_{(i,r)}$ with $c_i = T^{-1} \cdot 2^{i-1}$ for $i \in [H]$ and $\eta_r = \frac{1}{32 \cdot 2^r}$ for $r \in [R]$.

Initialization: for $(i,r) \in [H] \times [R]$, initiate base algorithm $\mathcal{B}_{i,r} \leftarrow \mathfrak{B}(\mathcal{X}_i)$ with $\mathcal{X}_i = \{x \mid \|x\|_2 \leq D_i\}$, which is an instance of $\mathfrak{B}$, and prior distribution $w_{1,(i,r)} \propto \eta_r^2 / c_i^2$.

for $t = 1$ ***to*** T **do**

  Each base learner $\mathcal{B}_{(i,r)}$ returns the local decision $v_{t,(i,r)}$ for each $i \in [H]$ and $r \in [R]$.

  Make the final decision $v_t = \sum_{(i,r)\in[H]\times[R]} w_{t,(i,r)} v_{t,(i,r)}$ and receive feedback $\nabla f_t(v_t)$.

  Construct feedback loss $\ell_t \in \mathbb{R}^N$ and correction term $a_t \in \mathbb{R}^N$ for meta algorithm : $\ell_{t,j} \triangleq \langle \nabla f_t(v_t), v_{t,j} \rangle$, $a_{t,j} \triangleq 32 \frac{\eta_r}{c_i} \ell_{t,j}^2, \forall j = (i,r) \in [H] \times [R]$.

  Meta algorithm updates the weight by $w_{t+1} = \operatorname{argmin}_{w\in\Omega} \langle w, \ell_t + a_t \rangle + D_\psi(w, w_t)$.

end

of S). It also adapts to the norm of the comparator $\|\mathring{u}_k\|_2$ on each interval $\mathcal{I}_k$, instead of only the maximum norm $\max_{k\in[S]} \|\mathring{u}_k\|_2$. As another remark, if the base algorithms further guarantee a data-dependent regret (this is satisfied by for example the algorithm of Zhang et al. (2019)), the switching regret can be further improved to $\widetilde{\mathcal{O}}\Big(\sum_{k=1}^S \|\mathring{u}_k\|_2 \sqrt{\sum_{t\in\mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2} \Big) \leq \widetilde{\mathcal{O}}\Big(\max_{k\in[S]} \|\mathring{u}_k\|_2 \cdot \sqrt{S \sum_{t=1}^T \|\nabla f_t(v_t)\|_2^2} \Big)$, replacing the dependence on T by the cumulative gradient norm square. This results holds even if the algorithm is required to make decisions from a bounded domain, thus strictly improving the $\widetilde{\mathcal{O}}\Big(D_{\max} \sqrt{S \sum_{t=1}^T \|\nabla f_t(v_t)\|_2^2} \Big)$ result of prior works (Cutkosky, 2020; Zhao et al., 2020, 2021b) where $D_{\max}$ is the diameter of the domain. See Appendix D.4 for details.

5.3. Summary: comparator-adaptive switching regret for unconstrained linear bandits

Combining all previous discussions, we now present the final result on unconstrained linear bandits.

Theorem 4 *Using Algorithm 2 (with $p = 2$) as the subroutine $\mathcal{A}_\mathcal{Z}$ and Algorithm 4 as the subroutine $\mathcal{A}_\mathcal{V}$ in the black-box reduction Algorithm 3, the overall algorithm enjoys the following comparator-adaptive switching regret against any partition $\mathcal{I}_1, \ldots, \mathcal{I}_S$ of $[T]$ and any corresponding comparators $\mathring{u}_1, \ldots, \mathring{u}_S \in \mathbb{R}^d$:*

$$\mathbb{E}[\textsc{Reg}_S] \leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^S \|\mathring{u}_k\|_2 \left(\sqrt{\frac{dT}{S}} + \sqrt{\frac{dS}{T}} |\mathcal{I}_k| \right) \right) \leq \widetilde{\mathcal{O}}\left(\max_{k\in[S]} \|\mathring{u}_k\|_2 \cdot \sqrt{dST} \right).$$

The proof can be found in Appendix D.5. Again, this is the first switching regret for unconstrained linear bandits, and it strictly generalizes the static regret results of van der Hoeven et al. (2020). Although we are not directly using our new corralling recipe to achieve this result, it clearly serves as an indispensable component for this result due to the usage of Algorithm 2.

6. Conclusion and Discussions

In this paper, we propose a new mechanism for combining a collection of bandit algorithms with regret overhead only logarithmically depending on the number of base algorithms. As a case study, we provide a set of new results on switching regret for adversarial linear bandits using this recipe. One future direction is to extend our switching regret results to linear bandits with general domains or even to general convex bandits, which appears to require additional new ideas to execute our recipe. Another interesting direction is to find more applications for our corralling mechanism beyond obtaining switching regret, as we know that logarithmic dependence on the number of base algorithms is possible.

Acknowledgments

Peng Zhao and Zhi-Hua Zhou are supported by National Science Foundation of China (61921006). Haipeng Luo and Mengxiao Zhang are supported by NSF Award IIS-1943607. The authors thank Chen-Yu Wei for helpful discussions on the idea of negative bias injection in the meta algorithm design.

References

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 12–38, 2017.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2116–2124, 2021.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

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

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, pages 138–158, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multiarmed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 41.1–41.14, 2012.

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of the 29th International Algorithmic Learning Theory (ALT)*, pages 111–127, 2018.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 989–997, 2012.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 1216–1259, 2021.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, pages 696–726, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1079–1087, 2019.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2250–2259, 2020.

Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 643–677, 2017.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1493–1529, 2018.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and RL. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 2276–2285, 2021.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 345–352, 2008.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1405–1411, 2015.

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3375–3383, 2015.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 11478–11489, 2020a.

Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Open problem: Model selection for contextual bandits. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, pages 3842–3846, 2020b.

András György and Csaba Szepesvári. Shifting regret, mirror descent, and matrices. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2943–2951, 2016.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. *Electronic colloquium on computational complexity (ECCC)*, 14(088), 2007.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 943–951, 2017.

Thomas Kerdreux, Christophe Roux, Alexandre d'Aspremont, and Sebastian Pokutta. Linear bandits on uniformly convex sets. *Journal of Machine Learning Research*, 22(284):1–23, 2021.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5805–5814, 2021.

Chung-Wei Lee, Haipeng Luo, and Mengxiao Zhang. A closer look at small-loss bounds for bandits with graph feedback. In *Proceedings of the 33th Conference on Learning Theory (COLT)*, pages 2516–2564, 2020.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of the 28th Annual Conference Computational Learning Theory (COLT)*, pages 1286–1304, 2015.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 1739–1776, 2018.

Teodor Vanislavov Marinov and Julian Zimmert. The pareto frontier of model selection for general contextual bandits. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Brendan Mcmahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 1020–1039, 2014.

Francesco Orabona. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1806–1814, 2013.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvári. Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 10328–10337, 2020.

Dirk van der Hoeven, Ashok Cutkosky, and Haipeng Luo. Comparator-adaptive convex bandits. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 4300–4354, 2021.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *Proceedings of the 33rd International Algorithmic Learning Theory (ALT)*, pages 1043–1096, 2022.

Lijun Zhang, Tie-Yan Liu, and Zhi-Hua Zhou. Adaptive regret of convex and smooth functions. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7414–7423, 2019.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 12510–12520, 2020.

Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021a.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *ArXiv preprint*, arXiv:2112.14368, 2021b.

Appendix A. Potential Approaches for Switching Regret of Linear Bandits

As mentioned in the main paper, to the best of our knowledge, we are not aware of any paper with switching regret for adversarial linear bandits. In this section, we present two potential approaches to achieve switching regret for adversarial linear bandits with ℓ_p-ball feasible domain, however, the regret bounds are suboptimal.

Method 1. Periodical Restart. The first generic method for tackling the switching regret of linear bandits is by running a classic linear bandits algorithm with a periodical restart. Specifically, suppose we employ an algorithm $\mathcal{A}$ as the base algorithm and restart it for every $\Delta > 0$ rounds. Then, the switching regret of the overall algorithm satisfies:

$$\mathbb{E}[\textsc{Reg}_S] \leq S \cdot \Delta + \left(\frac{T}{\Delta} - S\right) \cdot \textsc{Reg}(\mathcal{A}; \Delta) \leq \widetilde{\mathcal{O}}\left(S\Delta + \frac{T}{\sqrt{\Delta}}\right) = \widetilde{\mathcal{O}}\Big(S^{\frac{1}{3}} T^{\frac{2}{3}}\Big), \tag{12}$$

where the first inequality holds because there are at most S periods that contains a shift of comparators and we bound the regret in those periods trivially by $S\Delta$, and for the other periods the regret is controlled by the base algorithm $\mathcal{A}$. The second inequality is by chosen base algorithm $\mathcal{A}$ such that the regret is of order $\widetilde{\mathcal{O}}(\sqrt{\Delta})$, which can be satisfied by for example SCRIBLE (Abernethy et al., 2008). The last equality is by set the period optimally as $\Delta = \lceil (T/S)^{\frac{1}{3}} \rceil$. To summarize, the restarting algorithm applies to general adversarial linear bandits and attains a suboptimal switching regret of order $\widetilde{\mathcal{O}}(S^{\frac{1}{3}}T^{\frac{2}{3}})$, given the knowledge of S.

Method 2. EXP2 with Fixed-share Update. The second method is by using the EXP2 algorithm (Dani et al., 2008) with a uniform mixing update (Herbster and Warmuth, 1998; Auer et al., 2002), which can give an $\widetilde{\mathcal{O}}(d\sqrt{ST})$ switching regret for adversarial linear bandits with a general convex and compact domain. Note that the method is based on continuous exponential weights and thus requires log-concave sampling (Lovász and Vempala, 2007), which is theoretically efficient but usually time-consuming in practice. More importantly, the dimensional dependence is linear and hence not optimal when the feasible domain is an ℓ_p ball, $p \in (1, 2]$.

Beyond the above two methods, one may wonder whether we can simply use FTRL/OMD with some barrier regularizer (such as SCRIBLE (Abernethy et al., 2008)) along with either a uniform mixing update (Herbster and Warmuth, 1998; Auer et al., 2002) or a clipped domain (Herbster and Warmuth, 2001) to achieve switching regret for linear bandits. However, the attempt fails to work as the regularization term in the regret bound will become too large to control due to the property of barrier regularizer. Indeed, this method cannot even achieve switching regret guarantees for MAB due to the same reason.

Appendix B. Omitted Details for Section 4

In this section, we provide the omitted details for Section 4, including the pseudocode of the base algorithm (in Appendix B.1) and the proof of Theorem 1 (in Appendix B.2 – B.7). To prove Theorem 1, we first prove the unbiasedness of loss estimators in Appendix B.2, then decompose the regret in Appendix B.3, and subsequently upper bound each term in Appendix B.4, Appendix B.5, and Appendix B.6. We finally put everything together and present the proof in Appendix B.7.

B.1. Pseudocode of Base Algorithm

Algorithm 5 shows the pseudocode of the base algorithm for linear bandits with ℓ_p unit-ball feasible domain, which is the same as the one proposed in (Bubeck et al., 2018).

B.2. Unbiasedness of Loss Estimators

The following lemma shows the unbiasedness of the constructed loss estimators for both meta and base algorithms.

Lemma 5 *The meta loss estimator $\bar{\ell}_t$ defined in Eq. (7) and the base loss estimator $\widehat{\ell}_t$ defined in Eq. (6) satisfy that $\mathbb{E}_t[\bar{\ell}_t] = \ell_t$ and $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$ for all $t \in [T]$.*

Proof We first show the unbiasedness of the meta loss estimator $\bar{\ell}_t$. According to the definition in Eq. (7), we have

$$\mathbb{E}_t[\bar{\ell}_t] = \mathbb{E}_t[\widetilde{M}_t^{-1} x_t x_t^\top \ell_t]$$

Algorithm 5 Base algorithm for linear bandits on ℓ_p ball

Input: learning rate η, clipping parameter γ, initial round t_0.
Define: clipped feasible domain $\mathcal{X}' = \{x \mid \|x\|_p \le 1-\gamma\}$.
Initialize: $a_{t_0}^{(t_0)} = \mathrm{argmin}_{x\in\mathcal{X}'} R(x)$ and $\xi_{t_0}^{(t_0)} = 0$.
Draw $\widetilde{a}_{t_0}^{(t_0)}$ uniformly randomly from $\{\pm e_n\}_{n=1}^d$.
for $t = t_0$ ***to*** T **do**

Send $(\widetilde{a}_t^{(t_0)}, a_t^{(t_0)}, \xi_t^{(t_0)})$ to the meta algorithm.
Receive a loss estimator $\widehat{\ell}_t$.
Update the strategy based on OMD with regularizer $R(x) = -\log(1-\|x\|_p^p)$:

$$a_{t+1}^{(t_0)} = \underset{a\in\mathcal{X}'}{\mathrm{argmin}} \left\{ \left\langle a, \widehat{\ell}_t \right\rangle + \frac{1}{\eta} D_R(a, a_t^{(t_0)}) \right\}. \tag{13}$$

Generate a random variable $\xi_{t+1}^{(t_0)} \sim \mathbf{Ber}(\|a_{t+1}^{(t_0)}\|_p)$ and set

$$\widetilde{a}_{t+1}^{(t_0)} = \begin{cases} a_{t+1}^{(t_0)}/\|a_{t+1}^{(t_0)}\|_p & \text{if } \xi_{t+1}^{(t_0)} = 1, \\ \delta e_n & \text{if } \xi_{t+1}^{(t_0)} = 0, \end{cases}$$

where n is uniformly chosen from $\{1,\ldots,d\}$ and δ is a uniform random variable over $\{-1,+1\}$.

end

$$\begin{aligned}
&= \left(\frac{\beta}{d}\sum_{n=1}^d e_n e_n^\top + (1-\beta)\sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}\widetilde{a}_t^{(i)\top} \right)^{-1} \cdot \mathbb{E}_t[x_t x_t^\top]\ell_t \\
&= \left(\frac{\beta}{d}\sum_{n=1}^d e_n e_n^\top + (1-\beta)\sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}\widetilde{a}_t^{(i)\top} \right)^{-1} \cdot \left(\frac{\beta}{d}\sum_{n=1}^d e_n e_n^\top + (1-\beta)\sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}\widetilde{a}_t^{(i)\top} \right)\ell_t \\
&= \ell_t.
\end{aligned} \tag{14}$$

Next, we show the unbiasedness of the base loss estimator $\widehat{\ell}_t$. According to the definition in Eq. (6), we have

$$\begin{aligned}
\mathbb{E}_t[\widehat{\ell}_t] &= \mathbb{E}_t\left[\frac{1-\xi_t}{1-\beta} \cdot \frac{d}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \cdot (\ell_t^\top x_t)\cdot x_t \cdot \mathbb{1}\{\rho_t = 0\} \right] \\
&= \mathbb{E}_t\left[\sum_{j=1}^t \widehat{p}_{t,j} \cdot \frac{d(1-\xi_t^{(j)})}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \cdot (\ell_t^\top \widetilde{a}_t^{(j)})\cdot \widetilde{a}_t^{(j)} \right] \\
&= \sum_{j=1}^t \widehat{p}_{t,j}(1-\|a_t^{(j)}\|_p) \cdot \frac{d}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \cdot \frac{1}{d}\sum_{n=1}^d e_n e_n^\top \ell_t \\
&= \ell_t.
\end{aligned}$$

In above derivations, the first step simply substitutes the definition of loss estimator, the second step holds due to the sampling scheme of Algorithm 2 (see Line 4), and the third step is because of the sampling mechanism in base algorithm (see Algorithm 5). This finishes the proof. ∎

B.3. Regret Decomposition

We introduce shifted comparators $u'_t = (1-\gamma)u_t$ and $\mathring{u}'_k = (1-\gamma)\mathring{u}_k$ to ensure that $u'_t \in \mathcal{X}'$ for $t \in [T]$ and $\mathring{u}'_k \in \mathcal{X}'$ for $k \in [S]$, where $\mathcal{X}' = \{x \mid \|x\|_p \le 1-\gamma\}$. Based on the unbiasedness of $\widehat{\ell}_t$ and $\bar{\ell}_t$, the expected regret can be decomposed as

$$
\begin{aligned}
&\mathbb{E}[\textsc{Reg}_S] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t, \ell_t \rangle - \sum_{t=1}^{T} \langle u_t, \ell_t \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t, \ell_t \rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left\langle u'_t, \widehat{\ell}_t \right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle\right] \\
&= (1-\beta)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left\langle u'_t, \widehat{\ell}_t \right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \right\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left\langle u'_t, \widehat{\ell}_t \right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle - \beta \sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} c_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T} \left\langle u'_t, \widehat{\ell}_t \right\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle - \beta \sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right],
\end{aligned}
$$

where the third equation holds because of the sampling scheme of x_t: with probability β, the action x_t is uniformly sampled from $\{\pm e_n\}$, $n \in [d]$; with probability $1-\beta$, the action is sampled from $\{(\widetilde{a}_t^{(i)}, \xi_t^{(i)})\}_{i=1}^t$ according to $\widehat{p}_t$. In the last step, we recall that the notation $c_t \in \mathbb{R}^t$ is defined by $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle$ for all $i \in [t]$.

We further decompose the above regret into several intervals. To this end, we split the horizon to a partition $\mathcal{I}_1, \ldots, \mathcal{I}_S$. Let j_k be the start time stamp of $\mathcal{I}_k$. Note again that we use $\mathring{u}_k \in \mathcal{X}$ to denote the comparator in $\mathcal{I}_k$ for $k \in [S]$, which means that $u_t = \mathring{u}_k$ for all $t \in \mathcal{I}_k$. Then we have

$$
\begin{aligned}
&\mathbb{E}[\textsc{Reg}_S] \\
&\le \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\sum_{i=1}^{t} \widehat{p}_{t,i} c_{t,i} - \left\langle \mathring{u}'_k, \widehat{\ell}_t \right\rangle\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle - \beta \sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k} \langle \widehat{p}_t - e_{j_k}, c_t \rangle + \sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\langle e_{j_k}, c_t \rangle - \left\langle \mathring{u}'_k, \widehat{\ell}_t \right\rangle\right)\right] \\
&\qquad + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle - \beta \sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k} \langle \widehat{p}_t - e_{j_k}, c_t - b_t \rangle + \sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\langle e_{j_k}, c_t \rangle - \left\langle \mathring{u}'_k, \widehat{\ell}_t \right\rangle\right)\right] \\
&\qquad + \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k} \langle \widehat{p}_t - e_{j_k}, b_t \rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle u'_t - u_t, \ell_t \rangle - \beta \sum_{t=1}^{T}\sum_{i=1}^{t} \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t \right\rangle\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}-\widehat{c}_{t,j_k}\right)+\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\langle e_{j_k},c_t\rangle-\left\langle \mathring{u}'_k,\widehat{\ell}_t\right\rangle\right)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\widehat{c}_{t,i}=c_{t,i}-b_{t,i} \text{ for } i\in[t]) \\
&\quad + \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\sum_{i=1}^{t}(\widehat{p}_{t,i}-e_{j_k})b_{t,i}\right]+\mathbb{E}\left[\sum_{t=1}^{T}\langle u'_t-u_t,\ell_t\rangle-\beta\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}\left\langle \widetilde{a}_t^{(i)},\ell_t\right\rangle\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\sum_{i\in[t]}\langle p_t-e_{j_k},\widehat{c}_t\rangle+\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left(\left\langle \widetilde{a}_t^{(j_k)},\bar{\ell}_t\right\rangle-\left\langle \mathring{u}'_k,\widehat{\ell}_t\right\rangle\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\sum_{i=1}^{t}(\widehat{p}_{t,i}-e_{j_k})b_{t,i}\right]+\mathbb{E}\left[\sum_{t=1}^{T}\langle u'_t-u_t,\ell_t\rangle-\beta\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}\left\langle \widetilde{a}_t^{(i)},\ell_t\right\rangle\right] \\
&= \mathbb{E}\Bigg[\underbrace{\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\langle p_t-e_{j_k},\widehat{c}_t\rangle}_{\textsc{Meta-Regret}}+\underbrace{\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\left\langle a_t^{(j_k)}-\mathring{u}'_k,\widehat{\ell}_t\right\rangle}_{\textsc{Base-Regret}}+\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}b_{t,i}}_{\textsc{Pos-Bias}} \\
&\qquad\quad \underbrace{-\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}b_{t,j_k}}_{\textsc{Neg-Bias}}+\underbrace{\sum_{t=1}^{T}\langle u'_t-u_t,\ell_t\rangle-\beta\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}\left\langle \widetilde{a}_t^{(i)},\ell_t\right\rangle}_{\textsc{Deviation}}\Bigg], \qquad (15)
\end{aligned}
$$

where the second-last equality is due to the constructions of $\widehat{p}_t$ and $\widehat{c}_t$ (see Line 8 in Algorithm 2),

$$
\begin{aligned}
\langle p_t,\widehat{c}_t\rangle &= \sum_{i\in[t]}p_{t,i}\widehat{c}_{t,i}+\sum_{i>t}p_{t,i}\widehat{c}_{t,i}=\sum_{i\in[t]}\widehat{p}_{t,i}\left(\sum_{j\in[t]}p_{t,j}\right)\widehat{c}_{t,i}+\sum_{i>t}p_{t,i}\sum_{j\in[t]}\widehat{p}_{t,j}\widehat{c}_{t,j} \\
&= \sum_{i\in[t]}\widehat{p}_{t,i}\left(\sum_{j\in[t]}p_{t,j}+\sum_{i>t}p_{t,i}\right)\widehat{c}_{t,i}=\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i},
\end{aligned}
$$

and the last equality is based on the definition of $\widetilde{a}_t^{(i)}$ and $a_t^{(i)}$ and the following equation:

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle \widetilde{a}_t^{(i)},\bar{\ell}_t\right\rangle-\left\langle u,\widehat{\ell}_t\right\rangle\right]=\mathbb{E}\left[\left\langle \widetilde{a}_t^{(i)},\mathbb{E}_{\widehat{p}_t}[\bar{\ell}_t]\right\rangle-\left\langle u,\widehat{\ell}_t\right\rangle\right]=\mathbb{E}\left[\left\langle \widetilde{a}_t^{(i)},\ell_t\right\rangle-\left\langle u,\widehat{\ell}_t\right\rangle\right] \\
&=\mathbb{E}\left[\left\langle \mathbb{E}_{\xi_t^{(i)}}[\widetilde{a}_t^{(i)}],\ell_t\right\rangle-\left\langle u,\widehat{\ell}_t\right\rangle\right]=\mathbb{E}\left[\left\langle a_t^{(i)},\mathbb{E}[\widehat{\ell}_t]\right\rangle-\left\langle u,\widehat{\ell}_t\right\rangle\right]=\mathbb{E}\left[\left\langle a_t^{(i)}-u,\widehat{\ell}_t\right\rangle\right].
\end{aligned}
$$

As a consequence, we upper bound the expected switching regret by five terms as shown in Eq. (15), including: META-REGRET, BASE-REGRET, POS-BIAS, NEG-BIAS, and DEVIATION. In the following, we will bound each term respectively.

B.4. Bounding DEVIATION and POS-BIAS

DEVIATION. DEVIATION can be simply bounded by $(\beta+\gamma)T$ as

$$
\sum_{t=1}^{T}\langle u'_t-u_t,\ell_t\rangle-\beta\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}\left\langle \widetilde{a}_t^{(i)},\ell_t\right\rangle
$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} ((1-\gamma)-1)\langle u_t, \ell_t\rangle + \beta T \\
&\leq \sum_{t=1}^{T} (1-(1-\gamma)) + \beta T \\
&\leq (\beta+\gamma)T.
\end{aligned} \tag{16}
$$

where the first and second inequalities hold because we have $|\ell_t^\top x| \leq 1$ for any $x \in \mathcal{X}$ and $t \in [T]$.

POS-BIAS. According to the definition of $b_{t,i}$, we show that POS-BIAS is at most

$$
\frac{1}{\lambda T(1-\beta)} \sum_{t=1}^{T}\sum_{i=1}^{t} \frac{\widehat{p}_{t,i}(1-\|a_t^{(i)}\|_p)}{1-\sum_{j=1}^{t}\widehat{p}_{t,j}\|a_t^{(j)}\|_p} = \frac{1}{\lambda(1-\beta)} \leq \frac{2}{\lambda}. \tag{17}
$$

Hence, it remains to evaluate BASE-REGRET and META-REGRET, and in the following two subsections we present their upper bounds, respectively.

B.5. Bounding BASE-REGRET

In order to bound BASE-REGRET, we need to introduce the following lemma proven in (Bubeck et al., 2018), which shows that the dual local norm with respect to the regularizer $R(x) = -\log(1-\|x\|_p^p)$ is well bounded. This will later be shown to be crucial in controlling the stability of a_t updated by the online mirror descent shown in Algorithm 5.

Lemma 6 (Lemma 2 in (Bubeck et al., 2018)) *Let $x, \ell \in \mathbb{R}^d$ such that $\|x\|_p < 1$, $\|\ell\|_0 = 1$ and $\|\ell\|_2 \leq 1$. Let $y \in \mathbb{R}^d$ such that $\nabla R(y) \in [\nabla R(x), \nabla R(x)+\ell]$, $R(x) = -\log(1-\|x\|_p^p)$. Then, we have for $p \in (1,2]$,*

$$
\|\ell\|_{y,*}^2 \leq \frac{2^{\frac{3}{p-1}}(1-\|x\|_p^p)}{p(p-1)} \sum_{n=1}^{d}\left(|x_n|^{2-p} + |\ell_n|^{\frac{2-p}{p-1}}\right)\ell_n^2.
$$

In above, for a vector $h \in \mathbb{R}^d$, $\|h\|_0 \triangleq \#\{n \mid h_n \neq 0\}$ denotes the number of non-zero entries, $\|h\|_x \triangleq \sqrt{h^\top \nabla^2 R(x) h}$ denotes the local norm induced by R at x, and $\|h\|_{x,*} \triangleq \sqrt{h^\top(\nabla^2 R(x))^{-1}h}$ denotes the dual local norm.

Then we are ready to bound BASE-REGRET for each $k \in [S]$. Note that for each $k \in [S]$, as j_k is the start time stamp of interval $\mathcal{I}_k$, and base algorithm $\mathcal{B}_t$ starts at round t, we know that $\sum_{t\in\mathcal{I}_k}\langle a_t^{(j_k)} - \mathring{u}_k, \widehat{\ell}_t\rangle$ is in fact the (estimated) static regret against comparator $\mathring{u}_k$ for $\mathcal{B}_{j_k}$.

Lemma 7 *For an arbitrary interval $\mathcal{I}$ started at round j, if $\gamma = 4d\eta_{j'}$ for all $j' \in [T]$, Algorithm 2 ensures that the base regret of $\mathcal{B}_j$ with learning rate η for any comparator $u \in \mathcal{X}'$ is at most*

$$
\mathbb{E}\left[\sum_{t\in\mathcal{I}}\left\langle a_t^{(j)} - u, \widehat{\ell}_t\right\rangle\right] \leq \frac{\log(1/\gamma)}{\eta} + \frac{2^{\frac{4}{p-1}}d\eta}{(p-1)(1-\beta)} \sum_{t\in\mathcal{I}} \frac{1-\|a_t^{(j)}\|_p}{1-\sum_{i=1}^{t}\widehat{p}_{t,i}\|a_t^{(i)}\|_p}. \tag{18}
$$

Proof Since the base algorithm $\mathcal{B}_j$ performs the online mirror descent over loss $\widehat{\ell}_t$ with learning rate η, see update in Eq. (13), according to the standard analysis of OMD (see Lemma 27) we have

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}}\left\langle a_t^{(j)}-u,\widehat{\ell}_t\right\rangle\right] \leq \frac{R(u)-R(a_j^{(j)})}{\eta}+\frac{1}{\eta}\sum_{t\in\mathcal{I}}\mathbb{E}\left[D_{R^*}\left(\nabla R(a_t^{(j)})-\eta\widehat{\ell}_t,\nabla R(a_t^{(j)})\right)\right].$$

Consider the first term. As $a_j^{(j)}=\operatorname{argmin}_{x\in\mathcal{X}}R(x)$ and $u\in\mathcal{X}'=\{x\mid \|x\|_p\leq 1-\gamma\}$, we have

$$R(u)-R(a_j^{(j)})\leq -\log(1-(1-\gamma))\leq -\log\gamma. \tag{19}$$

For the second term, in the following we will employ Lemma 6 to show that

$$\mathbb{E}_t\left[D_{R^*}\left(\nabla R(a_t^{(j)})-\eta\widehat{\ell}_t,\nabla R(a_t^{(j)})\right)\right]\leq \eta^2\cdot\frac{d\cdot 2^{\frac{4}{p-1}}}{(p-1)(1-\beta)}\cdot\frac{1-\|a_t^{(j)}\|_p}{1-\sum_{i=1}^t\widehat{p}_{t,i}\|a_t^{(i)}\|_p}. \tag{20}$$

To this end, we need to verify the condition of Lemma 6. In fact, according to the definition of the base loss estimator in Eq. (6), $\widehat{\ell}_t$ is a non-zero vector only when Algorithm 2 samples from one of the base algorithm instances and $\xi_t=0$, meaning that $x_t=\pm e_n$ for some $n\in[d]$ according to Algorithm 5. Using the fact that $a_t^{(i)}\in\mathcal{X}'$ and $\beta\leq\frac{1}{2}$, we have $\|a_t^{(i)}\|_p\leq 1-\gamma$ and

$$\|\eta\widehat{\ell}_t\|_2\leq\frac{\eta d}{(1-\beta)(1-\sum_{i=1}^t\widehat{p}_{t,i}(1-\gamma))}\leq\frac{\eta d}{\gamma(1-\beta)}\leq\frac{2\eta d}{\gamma}\leq\frac{1}{2},$$

where the last inequality is because of the choice of $\gamma=4d\eta$. In addition, based on the definition of $\widehat{\ell}_t$, we have $\|\eta\widehat{\ell}_t\|_0=1$. Therefore, we can apply Lemma 6 and obtain that

$$\begin{aligned}
&\mathbb{E}_t\left[D_{R^*}\left(\nabla R(a_t^{(j)})-\eta\widehat{\ell}_t,\nabla R(a_t^{(j)})\right)\right]\\
&=\mathbb{E}_t\left[\|\eta\widehat{\ell}_t\|_{y_t,*}^2\right]\\
&\leq\eta^2\cdot\frac{2^{\frac{3}{p-1}}(1-\|a_t^{(j)}\|_p^p)}{p(p-1)}\sum_{n=1}^d\mathbb{E}_t\left[\left(|a_{t,n}^{(j)}|^{2-p}+|\eta\widehat{\ell}_{t,n}|^{\frac{2-p}{p-1}}\right)\widehat{\ell}_{t,n}^2\right]\\
&=\eta^2\cdot\frac{2^{\frac{3}{p-1}}(1-\|a_t^{(j)}\|_p^p)}{p(p-1)}\left(\underbrace{\sum_{n=1}^d\mathbb{E}_t\left[|a_{t,n}^{(j)}|^{2-p}\cdot\widehat{\ell}_{t,n}^2\right]}_{\text{TERM (A)}}+\underbrace{\sum_{n=1}^d\mathbb{E}_t\left[|\eta\widehat{\ell}_{t,n}|^{\frac{2-p}{p-1}}\cdot\widehat{\ell}_{t,n}^2\right]}_{\text{TERM (B)}}\right),
\end{aligned}$$

where the first equality holds for some $y_t\in[\nabla R(a_t^{(j)})-\eta\widehat{\ell}_t,\nabla R(a_t^{(j)})]$ by the definition of Bregman divergence and the mean value theorem, the second inequality is by Lemma 6. The last equality splits the desired quantity into two terms, and we upper bound TERM (A) and TERM (B) respectively.

For TERM (A), substituting the definition of loss estimator $\widehat{\ell}_t$ (see definition in Eq. (6)) yields

$$\begin{aligned}
&\sum_{n=1}^d\mathbb{E}_t\left[|a_{t,n}^{(j)}|^{2-p}\cdot\widehat{\ell}_{t,n}^2\right]\\
&=\frac{d^2}{(1-\beta)(1-\sum_{i=1}^t\widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2}\sum_{n=1}^d|a_{t,n}^{(j)}|^{2-p}\cdot\sum_{\tau=1}^t\widehat{p}_{t,\tau}\mathbb{E}_t\left[(1-\xi_t^{(\tau)})^2\widetilde{a}_{t,n}^{(\tau)^2}\langle\widetilde{a}_t^{(\tau)},\ell_t\rangle^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{d^2)}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2} \sum_{n=1}^d |a_{t,n}^{(j)}|^{2-p} \cdot \sum_{\tau=1}^t \widehat{p}_{t,\tau}(1-\|a_t^{(\tau)}\|_p) \cdot \frac{1}{d}\sum_{n'=1}^d \left[\mathbb{1}\{n'=n\}\ell_{t,n'}^2\right] \\
&= \frac{d}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)} \sum_{n=1}^d |a_{t,n}^{(j)}|^{2-p}\ell_{t,n}^2 \\
&\le \frac{d}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)},
\end{aligned}
$$

where the last inequality is because of Hölder's inequality, $\|\ell_t\|_q \le 1$ and $\|a_t^{(j)}\|_p \le 1$.

For TERM (B), again by definition of the loss estimator, we have

$$
\begin{aligned}
&\sum_{n=1}^d \mathbb{E}_t\left[|\eta\widehat{\ell}_{t,n}|^{\frac{2-p}{p-1}} \cdot \widehat{\ell}_{t,n}^2\right] \\
&\le \sum_{n=1}^d \mathbb{E}_t\left[\left|\frac{\eta(1-\xi_t)d(x_t x_t^\top \ell_t)_n}{(1-\beta)\gamma}\right|^{\frac{2-p}{p-1}} \cdot \widehat{\ell}_{t,n}^2\right] \\
&= \sum_{n=1}^d \mathbb{E}_t\left[\left|\frac{\eta(1-\xi_t)d(x_t x_t^\top \ell_t)_n}{(1-\beta)\gamma}\right|^{\frac{2-p}{p-1}} \cdot \frac{(1-\xi_t)^2 d^2 (x_t x_t^\top \ell_t)_n^2 \cdot \mathbb{1}\{\rho_t=0\}}{(1-\beta)^2(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2}\right] \\
&\le \sum_{n=1}^d \mathbb{E}_t\left[\left|\frac{(x_t x_t^\top \ell_t)_n}{2}\right|^{\frac{2-p}{p-1}} \cdot \frac{(1-\xi_t)^2 d^2 (x_t x_t^\top \ell_t)_n^2 \cdot \mathbb{1}\{\rho_t=0\}}{(1-\beta)^2(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2}\right] \qquad (\gamma = 4d\eta,\ 1-\beta \ge \tfrac{1}{2}) \\
&\le \frac{d^2}{(1-\beta)^2(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2} \sum_{n=1}^d \mathbb{E}_t\left[(1-\xi_t)^2 (x_t x_t^\top \ell_t)_n^q \cdot \mathbb{1}\{\rho_t=0\}\right] \\
&\hspace{20em} \left(\text{note that } \tfrac{2-p}{p-1}+2=q\right) \\
&\le \frac{1}{(1-\beta)^2} \cdot \frac{d^2}{(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)^2} \cdot \sum_{n=1}^d (1-\beta)\sum_{\tau=1}^t \widehat{p}_{t,\tau}(1-\|a_t^{(\tau)}\|_p) \cdot \frac{1}{d} \cdot \ell_{t,n}^q \\
&\le \frac{d}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p)}.
\end{aligned}
$$

Combining the above upper bounds for TERM (A) and TERM (B), we obtain

$$
\begin{aligned}
\mathbb{E}_t\left[D_{R^*}\left(\nabla R(a_t^{(j)}) - \eta\widehat{\ell}_t, \nabla R(a_t^{(j)})\right)\right] &\le \frac{\eta^2}{1-\beta} \cdot \frac{2d \cdot 2^{\frac{3}{p-1}}}{p(p-1)} \cdot \frac{1-\|a_t^{(j)}\|_p^p}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \\
&\le \frac{\eta^2}{1-\beta} \cdot \frac{d \cdot 2^{\frac{4}{p-1}}}{p(p-1)} \cdot \frac{1-\|a_t^{(j)}\|_p^p}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \\
&\le \frac{\eta^2}{1-\beta} \cdot \frac{d \cdot 2^{\frac{4}{p-1}}}{p-1} \cdot \frac{1-\|a_t^{(j)}\|_p}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p}.
\end{aligned}
$$

Note that the last step is true because we have $1-\|x\|_p^p \le p(1-\|x\|_p)$ by the following inequality

$$
1 + p \cdot \frac{\|x\|_p^p - 1}{p} \le \left(1 + \frac{\|x\|_p^p - 1}{p}\right)^p,
$$

which holds due to $p \in (1, 2]$ and $0 \le \|x\|_p \le 1$ as well as the Bernoulli's inequality that $1 + r\theta \le (1 + \theta)^r$ for any $r \ge 1$ and $\theta \ge -1$.

Therefore, we finish proving the desired upper bound in Eq. (20). Further combining it with the upper bound in Eq. (19) finishes the proof of Lemma 7. ■

We will show later that the second term in the bound shown in Eq. (18) can in fact be cancelled by the NEG-BIAS. Finally, we bound the term META-REGRET.

B.6. Bounding META-REGRET

We prove the following lemma to bound the META-REGRET.

Lemma 8 *For an arbitrary interval $\mathcal{I} \subseteq [T]$ started at round j, setting $\frac{\varepsilon}{\lambda\gamma T} \le \frac{1}{8}$, $\beta = 8d\varepsilon$ and $\mu = \frac{1}{T}$, Algorithm 2 guarantees that*

$$\sum_{t \in \mathcal{I}} \langle p_t - e_j, \widehat{c}_t \rangle \le \frac{2\log T}{\varepsilon} + \varepsilon \sum_{t \in \mathcal{I}} \sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 + \mathcal{O}\left(\frac{|\mathcal{I}|}{\varepsilon T}\right). \tag{21}$$

Proof Note that the meta algorithm essentially performs the exponential weights with a fixed-share update and sleeping expert. Define $v_{t+1,i} \triangleq \frac{p_{t,i}\exp(-\varepsilon \widehat{c}_{t,i})}{\sum_{t=1}^{T} p_{t,i}\exp(-\varepsilon \widehat{c}_{t,i})}$ for all $i \in [T]$. Then $p_{t+1,i} = \frac{\mu}{T} + (1 - \mu)v_{t+1,i}$. Note that

$$\begin{aligned}
&\langle p_t, \widehat{c}_t \rangle + \frac{1}{\varepsilon} \log \left(\sum_{i=1}^{T} p_{t,i} \exp(-\varepsilon \widehat{c}_{t,i}) \right) \\
&\le \langle p_t, \widehat{c}_t \rangle + \frac{1}{\varepsilon} \log \left(\sum_{i=1}^{T} p_{t,i} (1 - \varepsilon \widehat{c}_{t,i} + \varepsilon^2 \widehat{c}_{t,i}^2) \right) \\
&= \langle p_t, \widehat{c}_t \rangle + \frac{1}{\varepsilon} \log \left(1 - \varepsilon \langle p_t, \widehat{c}_t \rangle + \varepsilon^2 \sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 \right) \\
&\le \varepsilon \sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2.
\end{aligned}$$

The first inequality is because $\exp(-x) \le 1 - x + x^2$ holds for $x \ge -\frac{1}{2}$. To show that $\varepsilon \max_{i \in [t]} |\widehat{c}_{t,i}| \le \frac{1}{2}$, we have

$$\varepsilon \max_{i \in [t]} |\widehat{c}_{t,i}| = \varepsilon \max_{i \in [t]} \left| \left\langle \widetilde{a}_t^{(i)}, \widetilde{M}_t^{-1} x_t x_t^\top \ell_t \right\rangle - b_i \right| \le \varepsilon \max_{i \in [t]} \left| \widetilde{a}_t^{(i)\top} \widetilde{M}_t^{-1} x_t \right| + \varepsilon \max_{i \in [t]} |b_{t,i}|.$$

We can bound the first term by Hölder's inequality

$$\varepsilon \max_{i \in [t]} \left| \widetilde{a}_t^{(i)\top} \widetilde{M}_t^{-1} x_t \right| \le \varepsilon \max_{i \in [t]} \|\widetilde{a}_t^{(i)}\|_p \|\widetilde{M}_t^{-1} x_t\|_q \le \varepsilon \|\widetilde{M}_t^{-1} x_t\|_2 \le \frac{\varepsilon d}{\beta} \|x_t\|_2 \le \frac{\varepsilon d}{\beta} \|x_t\|_p \le \frac{\varepsilon d}{\beta},$$

where $\frac{1}{p} + \frac{1}{q} = 1$. The second inequality is by $\|\widetilde{a}_t^{(i)}\|_p \le 1$ and $p \le 2 \le q$. The third one is because $\widetilde{M}_t$ has the smallest eigenvalue $\frac{\beta}{d}$. By the definition of $b_{t,i}$, we can bound the second term as

$$\varepsilon \max_{i \in [t]} |b_{t,i}| \le \frac{\varepsilon}{\lambda T(1 - \beta)} \cdot \frac{1}{\gamma} \le \frac{2\varepsilon}{\lambda T \gamma}.$$

Therefore, according to the choice of ε, γ and λ, we have $\varepsilon \max_{i\in[t]} |\widehat{c}_{t,i}| \leq \frac{1}{2}$. Furthermore, by the definition of $v_{t+1,i}$, we have $\sum_{j=1}^{T} p_{t,j} \exp(-\varepsilon\widehat{c}_{t,j}) = p_{t,i}\exp(-\varepsilon\widehat{c}_{t,i})/v_{t+1,i}$. Therefore, we have

$$\frac{1}{\varepsilon}\log\left(\sum_{j=1}^{T} p_{t,j}\exp(-\varepsilon\widehat{c}_{t,j})\right) = -\frac{1}{\varepsilon}\log\left(\frac{v_{t+1,i}}{p_{t,i}}\right) - \widehat{c}_{t,i}.$$

Combining the two equations and taking summation over $t \in \mathcal{I}$, we have for any $e_j \in \Delta_T$, $j \in [T]$,

$$\sum_{t\in\mathcal{I}} \langle p_t, \widehat{c}_t\rangle - \sum_{t\in\mathcal{I}} \langle e_j, \widehat{c}_t\rangle \leq \varepsilon \sum_{t\in\mathcal{I}}\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2 + \frac{1}{\varepsilon}\sum_{t\in\mathcal{I}}\log\left(\frac{v_{t+1,j}}{p_{t,j}}\right).$$

Further note that

$$\begin{aligned}
\sum_{t\in\mathcal{I}}\log\left(\frac{v_{t+1,j}}{p_{t,j}}\right) &= \sum_{t\in\mathcal{I}}\log\left(\frac{p_{t+1,j}}{p_{t,j}}\right) + \sum_{t\in\mathcal{I}}\log\left(\frac{v_{t+1,j}}{\frac{\mu}{T} + (1-\mu)v_{t+1,j}}\right) \\
&\leq \log\left(\frac{p_{q+1,j}}{p_{s,j}}\right) + |\mathcal{I}|\log\left(\frac{1}{1-\mu}\right) \qquad \text{(let } \mathcal{I} = [s,q]) \\
&\leq \log(T^2) + \mathcal{O}\left(\frac{|\mathcal{I}|}{T}\right) \qquad (22)
\end{aligned}$$

where the last step is due to $p_{t,j} \geq \frac{\mu}{T} = \frac{1}{T^2}$ for $j \in [T]$ and $t \in [T]$, and moreover, we have $\log(\frac{1}{1-\mu}) = \log(1 + \frac{\mu}{1-\mu}) = \mathcal{O}(1/T)$ as $\mu = \frac{1}{T} \leq \frac{1}{2}$.

Combining the above two inequalities achieves

$$\sum_{t\in\mathcal{I}}\langle p_t - e_j, \widehat{c}_t\rangle \leq \varepsilon\sum_{t\in\mathcal{I}}\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2 + \frac{2\log T}{\varepsilon} + \mathcal{O}\left(\frac{|\mathcal{I}|}{\varepsilon T}\right).$$

which finishes the proof. ■

Next, we prove the following lemma, which bounds the second term shown in Eq. (21)

Lemma 9 *For any $t \in [T]$, setting $\lambda^2\gamma = \Theta\left(\sqrt{\frac{1}{dST^3}}\right)$ and $\beta \leq \frac{1}{2}$, Algorithm 2 guarantees that*

$$\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2 \leq \sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}^2 \leq 2\sum_{i\in[t]}\widehat{p}_{t,i}c_{t,i}^2 + \mathcal{O}\left(\sqrt{\frac{dS}{T}}\right), \qquad (23)$$

where $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t\rangle$ for $i \in [t]$.

Proof According to the definition of $\widehat{p}_t$, we have

$$\begin{aligned}
\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2 &= \sum_{i\in[t]} p_{t,i}\widehat{c}_{t,i}^2 + \sum_{i>t} p_{t,i}\left(\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}\right)^2 \leq \sum_{i\in[t]} p_{t,i}\widehat{c}_{t,i}^2 + \sum_{i>t} p_{t,i}\left(\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}^2\right) \\
&= \left(\sum_{i\in[t]} p_{t,i}\right)\left(\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}^2\right) + \left(\sum_{i>t} p_{t,i}\right)\left(\sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}^2\right) = \sum_{i\in[t]}\widehat{p}_{t,i}\widehat{c}_{t,i}^2,
\end{aligned}$$

where the inequality is because of Cauchy-Schwarz inequality. Besides, recall that $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle$ and $\widehat{c}_{t,i}^2 = (c_{t,i} - b_{t,i})^2 \leq 2c_{t,i}^2 + 2b_{t,i}^2$. According to the definition of $b_{t,i}$, we know that

$$\sum_{i\in[t]} \widehat{p}_{t,i} b_{t,i}^2 \leq \frac{4}{(\lambda T)^2} \frac{1}{\gamma} \sum_{i\in[t]} \widehat{p}_{t,i} \frac{1 - \|a_t^{(i)}\|_p}{1 - \sum_{j\in[t]} \widehat{p}_{t,j} \|a_t^{(j)}\|_p} = \frac{4}{(\lambda T)^2} \frac{1}{\gamma} = \mathcal{O}\left(\sqrt{\frac{dS}{T}}\right),$$

where the first inequality uses the fact that $b_{t,i} \leq \frac{1}{\lambda T \gamma (1-\beta)} \leq \frac{2}{\lambda T \gamma}$ and the last step holds because we choose $\lambda^2 \gamma = \Theta\left(\sqrt{\frac{1}{dST^3}}\right)$. ■

Combining Lemma 8 and Lemma 9, we obtain the following lemma to bound the meta-regret.

Lemma 10 *Define $C = \sqrt{p-1} \cdot 2^{-\frac{2}{p-1}}$. Set parameters $\varepsilon = \min\left\{\sqrt{\frac{S}{dT}}, \frac{1}{16d}, \frac{C^2}{2}\right\}$, $\beta = 8d\varepsilon$, $\lambda = \frac{C}{\sqrt{dST}}$, $\gamma = 4C\sqrt{\frac{dS}{T}}$ and $\mu = \frac{1}{T}$. Then, Algorithm 2 guarantees that*

$$\mathbb{E}\left[\text{META-REGRET}\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{dST}\right).$$

Proof It is evident to verify that the choice of ε, λ, β and γ satisfies the condition required in Lemma 8 and Lemma 9, then based on the two lemmas, with $\beta = 8d\varepsilon \leq \frac{1}{2}$, for each interval $\mathcal{I}_k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{I}_k} \langle p_t - e_{j_k}, \widehat{c}_t \rangle\right] \\
&\leq \frac{2\log T}{\varepsilon} + 2\varepsilon \mathbb{E}\left[\sum_{t\in\mathcal{I}_k}\sum_{i\in[t]} \widehat{p}_{t,i} c_{t,i}^2\right] + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right) \\
&\leq \frac{2\log T}{\varepsilon} + 2\varepsilon \mathbb{E}\left[\sum_{t\in\mathcal{I}_k}\sum_{i=1}^{t} \widehat{p}_{t,i} \widetilde{a}_t^{(i)\top} \widetilde{M}_t^{-1} x_t x_t^\top \widetilde{M}_t^{-1} \widetilde{a}_t^{(i)}\right] + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right) \\
&\leq \frac{2\log T}{\varepsilon} + 2\varepsilon \sum_{t\in\mathcal{I}_k}\sum_{i=1}^{t} \widehat{p}_{t,i} \widetilde{a}_t^{(i)\top} \widetilde{M}_t^{-1} \widetilde{a}_t^{(i)} + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right) \qquad (\mathbb{E}_t[x_t x_t^\top] = \widetilde{M}_t) \\
&\leq \frac{2\log T}{\varepsilon} + \frac{2\varepsilon}{1-\beta} \sum_{t\in\mathcal{I}_k}\sum_{i=1}^{t} \widehat{p}_{t,i} \widetilde{a}_t^{(i)\top} \left(\sum_{i=1}^{t} \widehat{p}_{t,i} \widetilde{a}_t^{(i)} \widetilde{a}_t^{(i)\top}\right)^{-1} \widetilde{a}_t^{(i)} + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right) \\
&\leq \frac{2\log T}{\varepsilon} + 4\varepsilon d|\mathcal{I}_k| + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right). && (24)
\end{aligned}$$

Summing the regret over all the intervals achieves the following meta-regret upper bound:

$$\begin{aligned}
\mathbb{E}\left[\text{META-REGRET}\right] &= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k} \langle p_t - e_{j_k}, \widehat{c}_t \rangle\right] && (25) \\
&\leq \frac{2S\log T}{\varepsilon} + 4\varepsilon dT + \mathcal{O}\left(\varepsilon\sqrt{dST}\right) + \mathcal{O}\left(1/\varepsilon\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{dST}\right),
\end{aligned}$$

where the last inequality is because we choose $\varepsilon = \min\left\{\sqrt{\frac{S}{dT}}, \frac{C^2}{2}, \frac{1}{16d}\right\}$. ■

B.7. Proof of Theorem 1

Putting everything together, we are now ready to prove our main result (Theorem 1).

Proof Based on the regret decomposition in Eq. (15), upper bound of bias term in Eq. (16), upper bound of positive term Eq. (17), base regret upper bound in Lemma 7 and meta regret upper bound in Eq. (25), we have

$$\begin{aligned}
\mathbb{E}[\text{REG}_S] &= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\langle x_t - \mathring{u}_k, \ell_t\rangle\right] \\
&\le \frac{2}{\lambda} + \sum_{k=1}^{S}\frac{\log(1/\gamma)}{\eta_{j_k}} + \left(\frac{2^{\frac{4}{p-1}}d\eta_{j_k}}{(p-1)(1-\beta)} - \frac{1}{\lambda T(1-\beta)}\right)\sum_{t\in\mathcal{I}_k}\frac{1-\|a_t^{(j_k)}\|_p}{1-\sum_{i=1}^{t}\widehat{p}_{t,i}\|a_t^{(i)}\|_p} \\
&\quad + (\beta+\gamma)T + \widetilde{\mathcal{O}}\left(\sqrt{dST}\right).
\end{aligned}$$

Importantly, note that the coefficient of the third term is actually zero. Indeed, due to the parameter configurations that $\gamma = 4C\sqrt{\frac{dS}{T}}$, $\eta = C\sqrt{\frac{S}{dT}}$, $\lambda = \frac{C}{\sqrt{dST}}$, $\beta = 8d\varepsilon$, $\varepsilon = \min\left\{\frac{1}{16d}, \frac{C^2}{2}, \sqrt{\frac{S}{dT}}\right\}$ and $C = \sqrt{p-1}\cdot 2^{-\frac{2}{p-1}}$, we can verify that

$$\frac{2^{\frac{4}{p-1}}d\eta}{p-1} - \frac{1}{\lambda T} = \frac{2^{\frac{2}{p-1}}\sqrt{dS}}{\sqrt{(p-1)T}} - \frac{\sqrt{dS}}{C\sqrt{T}} = 0.$$

Therefore, we obtain the following switching regret:

$$\mathbb{E}[\text{REG}_S] \le \frac{2}{\lambda} + 8\sqrt{dST} + 4C\sqrt{dST} + \widetilde{\mathcal{O}}(\sqrt{dST}) \le \widetilde{\mathcal{O}}\left(\sqrt{dST}\right),$$

which finishes the proof. ■

In addition, we also provide the following theorem showing the expected interval regret bound, which will be useful in the later analysis, for example, the unconstrained linear bandits in Section 5.

Theorem 11 *Define $C = \sqrt{p-1}\cdot 2^{-\frac{2}{p-1}}$. Set parameters $\varepsilon = \min\left\{\sqrt{\frac{S}{dT}}, \frac{1}{16d}, \frac{C^2}{2}\right\}$, $\beta = 8d\varepsilon$, $\lambda = \frac{C}{\sqrt{dST}}$, $\gamma = 4C\sqrt{\frac{dS}{T}}$, $\mu = \frac{1}{T}$ and $\eta = C\sqrt{\frac{S}{dT}}$. Then, Algorithm 2 guarantees that for any interval $\mathcal{I}$ and comparator $u \in \mathcal{X}$,*

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}}\ell_t^\top x_t - \sum_{t\in\mathcal{I}}\ell_t^\top u\right] \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{dT}{S}} + |\mathcal{I}|\sqrt{\frac{dS}{T}}\right). \tag{26}$$

Proof Based on the regret decomposition Eq. (15), Eq. (16), Eq. (17), Lemma 7 and Eq. (24) within rounds $t \in \mathcal{I}$ starting at round j, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{I}} \ell_t^\top x_t - \sum_{t\in\mathcal{I}} \ell_t^\top u\right] \\
&\le \frac{2|\mathcal{I}|}{\lambda T} + \frac{\log(1/\gamma)}{\eta_j} + \left(\frac{2^{\frac{4}{p-1}} d\eta_j}{(p-1)(1-\beta)} - \frac{1}{\lambda T(1-\beta)}\right)\sum_{t\in\mathcal{I}} \frac{1-\|a_t^{(j)}\|_p}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_p} \\
&\quad + (\beta+\gamma)|\mathcal{I}| + \widetilde{\mathcal{O}}\left(\varepsilon|\mathcal{I}|\sqrt{\frac{dS}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}|}{\varepsilon T}\right).
\end{aligned}$$

Again, note that according to the choice of γ, η, λ, β and ε, we have

$$\frac{2^{\frac{4}{p-1}} d\eta}{p-1} - \frac{1}{\lambda T} = \frac{2^{\frac{2}{p-1}}\sqrt{dS}}{\sqrt{(p-1)T}} - \frac{\sqrt{dS}}{C\sqrt{T}} = 0.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\mathcal{I}} \ell_t^\top x_t - \sum_{t\in\mathcal{I}} \ell_t^\top u\right] \\
&\le \frac{2}{C}|\mathcal{I}|\sqrt{\frac{dS}{T}} + \frac{\log\left(\frac{1}{4C}\cdot\sqrt{\frac{T}{dS}}\right)}{C}\cdot\sqrt{\frac{dT}{S}} + 8d|\mathcal{I}|\sqrt{\frac{S}{dT}} + 4C|\mathcal{I}|\sqrt{\frac{dS}{T}} + \widetilde{\mathcal{O}}\left(\varepsilon|\mathcal{I}|\sqrt{\frac{dS}{T}} + \frac{|\mathcal{I}|}{\varepsilon T}\right) \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{\frac{dT}{S}} + |\mathcal{I}|\sqrt{\frac{dS}{T}}\right),
\end{aligned}$$

which finishes the proof. ■

Appendix C. Extension to Smooth and Strongly Convex Set

In this section, we extend our results for linear bandits with ℓ_p-ball feasible domain in Section 4 to the setting when the feasible domain is a *smooth and strongly convex set*. Kerdreux et al. (2021) studied the static regret for linear bandits in this setting, and we focus on the S-switching regret.

C.1. Main Results

Formally, we investigate adversarial linear bandits with a smooth and strongly convex feasible domain. In the following, we present the definitions of smooth set (Kerdreux et al., 2021, Definition 1) and strongly convex set (Kerdreux et al., 2021, Definition 3).

Definition 12 (smooth set) *A compact convex set $\mathcal{X}$ is smooth if and only if $|N_{\mathcal{X}}(x) \cap \partial\mathcal{X}^\circ| = 1$ for any $x \in \partial\mathcal{X}$, where $N_{\mathcal{X}}(x) \triangleq \{u \in \mathbb{R}^d \mid \langle x-y, u\rangle \ge 0, \forall y \in \mathcal{X}\}$, $\partial\mathcal{X}$ is the boundary of $\mathcal{X}$ and $\mathcal{X}^\circ = \{u \in \mathbb{R}^d \mid \langle u, x\rangle \le 1, \forall x \in \mathcal{X}\}$ is the polar of $\mathcal{X}$.*

Definition 13 (strongly convex set) *Let $\mathcal{X}$ be a centrally symmetric set with non-empty interior. Let $\alpha > 0$ be the curvature coefficient. The set $\mathcal{X}$ is α-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$ if and only if for any $x, y, z \in \mathcal{X}$ and $\gamma \in [0,1]$, we have*

$$\left(\gamma x + (1-\gamma)y + \frac{\alpha}{2}\gamma(1-\gamma)\|x-y\|_{\mathcal{X}}^2 \cdot z\right) \in \mathcal{X},$$

where $\|x\|_{\mathcal{X}} \triangleq \inf\{\lambda > 0 \mid x \in \lambda\mathcal{X}\}$ is the gauge function to $\mathcal{X}$.

Conventionally, we assume that $|\ell_t^\top x| \leq 1$ holds for all $x \in \mathcal{X}$ and $t \in [T]$. We also assume that $\ell_p(1) \subseteq \mathcal{X} \subseteq \ell_q(1)$ with $p \in (1,2]$ and $\frac{1}{p} + \frac{1}{q} = 1$, where $\ell_s(r) \triangleq \{x \in \mathbb{R}^d \mid \|x\|_s \leq r\}$ denotes the ℓ_s-norm ball ($s \geq 1$) with radius $r > 0$. We here stress the connection and difference between the strongly convex set setting and the ℓ_p-ball setting considered in Section 4. Note that $\mathcal{X}$ is a subset of ℓ_q ball and includes ℓ_p ball. Besides, ℓ_p ball is also smooth when $p \in (1,2]$. Therefore, it includes ℓ_p-ball feasible set for $p \in (1,2]$ but can be more general. Nevertheless, the switching regret bound we will prove is $\widetilde{\mathcal{O}}(d^{1/p}\sqrt{ST})$, which recovers the $\widetilde{\mathcal{O}}(\sqrt{dST})$ switching regret of ℓ_p-ball feasible domain in Theorem 1 only when $p = 2$ but leads to a slightly worse dependence on d when $p \in (1,2)$. Note that as $p > 1$, this bound is still better than $\widetilde{\mathcal{O}}(d\sqrt{ST})$.

Our proposed algorithm for smooth and strongly convex set is basically the same as the one proposed for the ℓ_p ball setting, except that we now need to modify the base algorithm based on the algorithm introduced in (Kerdreux et al., 2021) and also need to revise the construction of injected bias $b_{t,i}$ and the loss estimator $\widehat{\ell}_t$ in the meta level. Specifically, in the base algorithm we use online mirror descent with the following regularizer,

$$R(x) = -\log(1 - \|x\|_{\mathcal{X}}) - \|x\|_{\mathcal{X}},$$

whose detailed update procedures are presented in Algorithm 7. For the meta algorithm, the update procedures are in Algorithm 6, notably, the injected bias b_t is constructed according to Eq. (29) and the base loss estimator $\widehat{\ell}_t$ is constructed according to Eq. (27).

We have the following theorem regarding the switching regret of our proposed algorithm for linear bandits on smooth and strongly convex feasible domain.

Theorem 14 *Consider a compact convex set $\mathcal{X}$ that is centrally symmetric with non-empty interior. Suppose that $\mathcal{X}$ is smooth and α-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$ and $\ell_p(1) \subseteq \mathcal{X} \subseteq \ell_q(1)$, $p \in (1,2]$, $\frac{1}{p} + \frac{1}{q} = 1$. Define $C = \sqrt{\frac{\alpha}{10\alpha+8}}$. Set parameters $\gamma = 4Cd^{\frac{1}{q}}\sqrt{\frac{S}{T}}$, $\lambda = \frac{Cd^{-\frac{1}{q}}}{\sqrt{ST}}$, $\beta = 8d^{\frac{2}{p}}\varepsilon$, $\varepsilon = \min\left\{\frac{1}{16d^{\frac{2}{p}}}, \frac{C^2}{2}, d^{-\frac{1}{p}}\sqrt{\frac{S}{T}}\right\}$, $\mu = \frac{1}{T}$ and $\eta = Cd^{-\frac{1}{p}}\sqrt{\frac{S}{T}}$. Then, Algorithm 6 guarantees*

$$\mathbb{E}[\text{REG}_S] = \mathbb{E}\left[\sum_{t=1}^T \ell_t^\top x_t - \sum_{t=1}^T \ell_t^\top u_t\right] \leq \widetilde{\mathcal{O}}\left(d^{1/p}\sqrt{ST}\right),$$

where $u_1, \ldots, u_T \in \mathcal{X}$ is the comparator sequence such that $\sum_{t=2}^T \mathbb{1}\{u_{t-1} \neq u_t\} \leq S - 1$.

In the following, we first introduce some definitions and lemmas useful for the analysis in strongly convex set in Appendix C.2 and then prove Theorem 14 in Appendix C.3–C.8. To prove Theorem 14, similar to the analysis structure in Appendix B, we first prove the unbiasedness of loss estimators in Appendix C.3, and then in Appendix C.4, we decompose the regret into several terms, and subsequently upper bound each term in Appendix C.5, Appendix C.6, and Appendix C.7. We finally put everything together and present the proof in Appendix C.8.

Algorithm 6 Algorithm for adversarial linear bandits over smooth and strongly convex set with switching regret

Input: clipping parameter γ, base learning rate η, meta learning rate ε, mixing rate μ, exploration parameter β, bias coefficient λ, initial uniform distribution $p_1 \in \Delta_T$.

for $t = 1$ ***to*** T **do**

Start a new base algorithm $\mathcal{B}_t$, which is an instance of Algorithm 7 with learning rate η, clipping parameter γ, and initial round t.

Receive local decision $(\widetilde{a}_t^{(i)}, a_t^{(i)}, \xi_t^{(i)})$ from base algorithm $\mathcal{B}_i$ for each $i \le t$.

Compute the renormalized distribution $\widehat{p}_t \in \Delta_t$ such that $\widehat{p}_{t,i} \propto p_{t,i}$ for $i \in [t]$.

Sample a Bernoulli random variable ρ_t with mean β. If $\rho_t = 1$, uniformly sample x_t from $\{\pm e_n\}_{n=1}^d$; otherwise, sample $i_t \in [t]$ according to $\widehat{p}_t$, and set $x_t = \widetilde{a}_t^{(i_t)}$ and $\xi_t = \xi_t^{(i_t)}$.

Make the final decision x_t and receive feedback $\ell_t^\top x_t$.

Construct the base loss estimator $\widehat{\ell}_t \in \mathbb{R}^d$ as follows and send it to all base algorithms $\{\mathcal{B}_i\}_{i=1}^t$:

$$\widehat{\ell}_t = \frac{\mathbb{1}\{\rho_t = 0\}\mathbb{1}\{\xi_t = 0\}}{1-\beta} \cdot \frac{d(\ell_t^\top x_t)}{1 - \sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} \cdot x_t. \tag{27}$$

Construct another loss estimator $\bar{\ell}_t \in \mathbb{R}^d$ as

$$\bar{\ell}_t = \widetilde{M}_t^{-1} x_t x_t^\top \ell_t, \tag{28}$$

where $\widetilde{M}_t = \frac{\beta}{d}\sum_{i=1}^d e_i e_i^\top + (1-\beta)\sum_{i=1}^t \widehat{p}_{t,i}\widetilde{a}_t^{(i)}\widetilde{a}_t^{(i)\top}$.

Construct the meta loss estimator $\widehat{c}_t \in \mathbb{R}^T$ as:

$$\widehat{c}_{t,i} = \begin{cases} \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle - b_{t,i}, & i \le t, \\ \sum_{j=1}^t \widehat{p}_{t,j}\widehat{c}_{t,j}, & i > t, \end{cases} \quad \text{where } \; b_{t,i} = \frac{1}{\lambda T(1-\beta)} \frac{1 - \|a_t^{(i)}\|_{\mathcal{X}}}{1 - \sum_{j=1}^t \widehat{p}_{t,j}\|a_t^{(j)}\|_{\mathcal{X}}}. \tag{29}$$

Meta algorithm updates the weight $p_{t+1} \in \Delta_T$ according to

$$p_{t+1,i} = (1-\mu)\frac{p_{t,i}\exp(-\varepsilon \widehat{c}_{t,i})}{\sum_{j=1}^T p_{t,j}\exp(-\varepsilon \widehat{c}_{t,j})} + \frac{\mu}{T}, \quad \forall i \in [T]. \tag{30}$$

end

C.2. Preliminary

This subsection collects some useful definitions and lemmas for the analysis. We refer the reader to (Kerdreux et al., 2021) for detailed introductions. Define $\|\cdot\|_{\mathcal{X}}$ is the gauge function to $\mathcal{X}$ as

$$\|x\|_{\mathcal{X}} = \inf\{\lambda > 0 \mid x \in \lambda\mathcal{X}\}. \tag{32}$$

The polar of $\mathcal{X}$ is defined as $\mathcal{X}^\circ = \{\ell \in \mathbb{R}^d \mid \langle x, \ell \rangle \le 1, \forall x \in \mathcal{X}\}$. If $\mathcal{X}$ is symmetric, then based on the assumption $|\langle x, \ell_t \rangle| \le 1$, we have $\ell_t \in \mathcal{X}^\circ$. Based on the definition of gauge function, we have $\|x\|_{\mathcal{X}} \le 1$ for all $x \in \mathcal{X}$. In addition, we have the Hölder's inequality $\langle x, \ell \rangle \le \|x\|_{\mathcal{X}} \cdot \|\ell\|_{\mathcal{X}^\circ}$. In this problem, we also assume that $\ell_p(1) \subseteq \mathcal{X} \subseteq \ell_q(1)$, $p \in (1, 2]$, $\frac{1}{p} + \frac{1}{q} = 1$ which implies

Algorithm 7 Base algorithm for linear bandits on strongly convex set

Input: learning rate η, clipping parameter γ, initial round t_0.
Define: clipped feasible domain $\mathcal{X}' = \{x \mid \|x\|_{\mathcal{X}} \le 1-\gamma, x \in \mathcal{X}\}$.
Initialize: $a_{t_0}^{(t_0)} = \operatorname{argmin}_{x\in\mathcal{X}'} R(x)$ and $\xi_{t_0}^{(t_0)} = 0$.
Draw $\widetilde{a}_{t_0}^{(t_0)}$ uniformly randomly from $\{\pm e_n\}_{n=1}^d$.
for $t = t_0$ ***to*** T **do**

Send $(\widetilde{a}_t^{(t_0)}, a_t^{(t_0)}, \xi_t^{(t_0)})$ to the meta algorithm.
Receive a loss estimator $\widehat{\ell}_t$.
Update the strategy based on OMD with regularizer $R(x) = -\log(1-\|x\|_{\mathcal{X}}) - \|x\|_{\mathcal{X}}$:

$$a_{t+1}^{(t_0)} = \operatorname*{argmin}_{a\in\mathcal{X}'} \left\{ \left\langle a, \widehat{\ell}_t \right\rangle + \frac{1}{\eta} D_R(a, a_t^{(t_0)}) \right\}. \tag{31}$$

Generate a random variable $\xi_{t+1}^{(t_0)} \sim \mathbf{Ber}(\|a_{t+1}^{(t_0)}\|_{\mathcal{X}})$ and set

$$\widetilde{a}_{t+1}^{(t_0)} = \begin{cases} a_{t+1}^{(t_0)}/\|a_{t+1}^{(t_0)}\|_{\mathcal{X}} & \text{if } \xi_{t+1}^{(t_0)} = 1, \\ \delta e_n & \text{if } \xi_{t+1}^{(t_0)} = 0, \end{cases}$$

where n is uniformly chosen from $\{1, \dots, d\}$ and δ is a uniform random variable over $\{-1, +1\}$.

end

that $\{\pm e_n\}_{n\in[d]} \subseteq \mathcal{X}$. By the definition of $\mathcal{X}^\circ$, we also have $\ell_p(1) \subseteq \mathcal{X}^\circ \subseteq \ell_q(1)$. The following lemmas show some useful identities for the regularizer $R(x)$.

Lemma 15 (Lemma 5 of Kerdreux et al. (2021)) *A gauge function $\|\cdot\|_{\mathcal{X}}$ is differentiable at $x \in \mathbb{R}^d\backslash\{\mathbf{0}\}$ if and only if its support set $S(\mathcal{X}^\circ, x) = \{h \in \mathcal{X}^\circ \mid \langle h, x\rangle = \sup_{h'\in\mathcal{X}^\circ} \langle h', x\rangle\}$ contains a single point h. If this is the case, we have $\nabla\|\cdot\|_{\mathcal{X}}(x) = d$. Besides, the following assertions are true: (1) $\|(\nabla\|\cdot\|_{\mathcal{X}}(x))\|_{\mathcal{X}^\circ} = 1$; (2) $\nabla\|\cdot\|_{\mathcal{X}}(\lambda x) = \nabla\|\cdot\|_{\mathcal{X}}(x)$, for any $\lambda > 0$; (3) if $\mathcal{X}^\circ$ is strictly convex, then $\|\cdot\|_{\mathcal{X}}$ is differentiable in $\mathbb{R}^d\backslash\{\mathbf{0}\}$.*

Lemma 16 (Corollary 8 of Kerdreux et al. (2021)) *Let $\mathcal{X}$ be a centrally symmetric set with non empty interior. Assume that $\mathcal{X}$ is α-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$. Then for any $(u, v) \in \mathbb{R}^n$,*

$$D_{\frac{1}{2}\|\cdot\|^2_{\mathcal{X}^\circ}}(u, v) \le \frac{4(\alpha+1)}{\alpha}\|u-v\|^2_{\mathcal{X}^\circ}.$$

Lemma 17 (Lemma 15 of Kerdreux et al. (2021)) *Assume $\mathcal{X} \subseteq \mathbb{R}^d$ is strictly convex compact and smooth set. Let $x \in \mathcal{X}$ such that $\|x\|_{\mathcal{X}} < 1$ and $h \in \mathbb{R}^d\backslash\{\mathbf{0}\}$. We have $R(x)$ is differentiable on int$(\mathcal{X})$ and*

$$\nabla R(x) = \frac{\|x\|_{\mathcal{X}}}{1-\|x\|_{\mathcal{X}}} \cdot \nabla\|\cdot\|_{\mathcal{X}}(x),$$
$$R^*(h) = \|h\|_{\mathcal{X}^\circ} - \log(1+\|h\|_{\mathcal{X}^\circ}),$$
$$\nabla R^*(h) = \frac{\|h\|_{\mathcal{X}^\circ}}{1+\|h\|_{\mathcal{X}^\circ}}\nabla\|\cdot\|_{\mathcal{X}^\circ}(h).$$

C.3. Unbiasedness of Loss Estimator

We first show that the loss estimator for the meta algorithm $\bar{\ell}_t$ and the one for the base algorithm $\widehat{\ell}_t$ constructed in Algorithm 6 are unbiased.

Lemma 18 *The meta loss estimator $\bar{\ell}_t$ defined in Eq. (28) and the base loss estimator $\widehat{\ell}_t$ defined in Eq. (27) satisfy that $\mathbb{E}_t[\bar{\ell}_t] = \ell_t$ and $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$ for all $t \in [T]$.*

Proof The the unbiasedness of $\bar{\ell}_t$ can be proven in the exact same way as in Eq. (14). For $\widehat{\ell}_t$, according to the sampling scheme of x_t, we have

$$\begin{aligned}
\mathbb{E}_t[\widehat{\ell}_t] &= \mathbb{E}_t\left[\frac{1-\xi_t}{1-\beta} \cdot \frac{d}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} x_t x_t^\top \ell_t \cdot \mathbb{1}\{\rho_t = 0\}\right] \\
&= \mathbb{E}_t\left[(1-\xi_t) \cdot \frac{d}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} x_t x_t^\top \ell_t \,\middle|\, \rho_t = 0\right] \\
&= \mathbb{E}_t\left[\sum_{j=1}^t \widehat{p}_{t,j} \cdot \frac{d(1-\xi_t^{(j)})}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} \widetilde{a}_t^{(j)} \widetilde{a}_t^{(j)\top} \ell_t\right] \\
&= \sum_{j=1}^t \widehat{p}_{t,j} \cdot \frac{d(1-\|a_t^{(j)}\|_{\mathcal{X}})}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} \frac{1}{d} \sum_{n=1}^d e_n e_n^\top \ell_t = \ell_t.
\end{aligned}$$

This ends the proof. ■

C.4. Regret Decomposition

Similar to the analysis in Appendix B, we decompose the expected switching regret into five terms and then bound each term respectively. Again, we split the horizon to $\mathcal{I}_1, \ldots, \mathcal{I}_S$, and let j_k be the start time stamp of $\mathcal{I}_k$. We introduce $u_t' = (1-\gamma)u_t$ and $\mathring{u}_k' = (1-\gamma)\mathring{u}_k$ to ensure that $u_t' \in \mathcal{X}'$ for $t \in [T]$ and $\mathring{u}_k' \in \mathcal{X}'$ for $k \in [S]$, where $\mathcal{X}' = \{x \mid \|x\|_{\mathcal{X}} \le 1-\gamma, x \in \mathcal{X}\}$. Similar to the decomposition method of Eq. (15), the expected regret can be decomposed as

$$\begin{aligned}
\mathbb{E}[\textsc{Reg}_S] &= \mathbb{E}\left[\sum_{t=1}^T \langle x_t, \ell_t\rangle - \sum_{t=1}^T \langle u_t, \ell_t\rangle\right] \\
&= \mathbb{E}\Bigg[\underbrace{\sum_{k=1}^S \sum_{t\in\mathcal{I}_k} \langle p_t - e_{j_k}, \widehat{c}_t\rangle}_{\textsc{Meta-Regret}} + \underbrace{\sum_{k=1}^S \sum_{t\in\mathcal{I}_k} \left\langle a_t^{(j_k)} - \mathring{u}_k', \widehat{\ell}_t\right\rangle}_{\textsc{Base-Regret}} + \underbrace{\sum_{t=1}^T \sum_{i=1}^t \widehat{p}_{t,i} b_{t,i}}_{\textsc{Pos-Bias}} \\
&\qquad \underbrace{- \sum_{k=1}^S \sum_{t\in\mathcal{I}_k} b_{t,j_k}}_{\textsc{Neg-Bias}} + \underbrace{\sum_{t=1}^T \langle u_t' - u_t, \ell_t\rangle - \beta \sum_{t=1}^T \sum_{i=1}^t \widehat{p}_{t,i} \left\langle \widetilde{a}_t^{(i)}, \ell_t\right\rangle}_{\textsc{Deviation}}\Bigg].
\end{aligned} \tag{33}$$

In the following, we will bound each term respectively.

C.5. Bounding DEVIATION and POS-BIAS

DEVIATION. DEVIATION term can still be bounded by $(\beta+\gamma)T$ as

$$\begin{aligned}
&\sum_{t=1}^{T}\left\langle u_t'-u_t,\ell_t\right\rangle-\beta\sum_{t=1}^{T}\sum_{i=1}^{t}\widehat{p}_{t,i}\left\langle\widetilde{a}_t^{(i)},\ell_t\right\rangle\\
&\le\sum_{t=1}^{T}\left((1-\gamma)-1\right)\left\langle u_t,\ell_t\right\rangle+\beta T\\
&\le\sum_{t=1}^{T}\left(1-(1-\gamma)\right)+\beta T=(\beta+\gamma)T. \qquad (34)
\end{aligned}$$

POS-BIAS. According to the definition of $b_{t,i}$, we have

$$\frac{1}{\lambda T(1-\beta)}\sum_{t=1}^{T}\sum_{i=1}^{t}\frac{\widehat{p}_{t,i}(1-\|a_t^{(i)}\|_{\mathcal{X}})}{1-\sum_{j=1}^{t}\widehat{p}_{t,j}\|a_t^{(j)}\|_{\mathcal{X}}}=\frac{1}{\lambda(1-\beta)}\le\frac{2}{\lambda}, \qquad (35)$$

where the last inequality is because $\beta\le\frac{1}{2}$.

In the following two subsections, we bound BASE-REGRET and META-REGRET respectively.

C.6. Bounding BASE-REGRET

Before bounding the term BASE-REGRET, we show the following two lemmas which will be useful in the analysis. The first lemma bounds the scale of the loss estimator used for the base algorithm.

Lemma 19 *For any $x\in(1-\gamma)\mathcal{X}$ and η, define $u=\nabla R(x)-\eta\widehat{\ell}_t$ and $v=\nabla R(x)$ with $\widehat{\ell}_t$ defined in Algorithm 6. We have*

$$\frac{\|u\|_{\mathcal{X}^\circ}-\|v\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}}\ge-\frac{2\eta d}{\gamma}.$$

Proof First, note that using Lemma 15 and Lemma 17, the denominator can be written as

$$\frac{1}{1+\|v\|_{\mathcal{X}^\circ}}=\frac{1}{1+\|\nabla R(x)\|_{\mathcal{X}^\circ}}=(1-\|x\|_{\mathcal{X}})(\|(\nabla\|\cdot\|_{\mathcal{X}}(x))\|_{\mathcal{X}^\circ})^{-1}=1-\|x\|_{\mathcal{X}}. \qquad (36)$$

For the numerator, note that $a_t^{(i)}\in(1-\gamma)\mathcal{X}$ for all $t\in[T]$ and $i\in[t]$ and $\beta\le\frac{1}{2}$, we have

$$\|\widehat{\ell}_t\|_{\mathcal{X}^\circ}\le\frac{d}{(1-\beta)(1-(1-\gamma))}|x_t^\top\ell_t|\cdot\|x_t\|_{\mathcal{X}^\circ}\cdot\mathbb{1}\{x_t\in\{\pm e_n\}_{n\in[d]}\}\le\frac{2d\|x_t\|_{\mathcal{X}^\circ}}{\gamma}\mathbb{1}\{x_t\in\{\pm e_n\}_{n\in[d]}\}.$$

Therefore, according to triangle inequality, we have

$$\|u\|_{\mathcal{X}^\circ}-\|v\|_{\mathcal{X}^\circ}\ge-\eta\|\widehat{\ell}_t\|_{\mathcal{X}^\circ}\ge-\frac{2d\eta}{\gamma}\|x_t\|_{\mathcal{X}^\circ}\cdot\mathbb{1}\{x_t\in\{\pm e_n\}_{n\in[d]}\}.$$

Note that $\mathcal{X}\subseteq\ell_q(1)$, we have $\ell_q(1)^\circ=\ell_p(1)\subseteq\mathcal{X}^\circ$, which means that $e_n\in\mathcal{X}^\circ$. This means that $\|e_n\|_{\mathcal{X}^\circ}\le 1$ and we have

$$\|u\|_{\mathcal{X}^\circ}-\|v\|_{\mathcal{X}^\circ}\ge-\frac{2\eta d}{\gamma},$$

which finishes the proof. ■

The second lemma helps to bound the stability of the base algorithm, which is originally introduced in (Kerdreux et al., 2021, Lemma 17). For completeness, we include the proof here.

Lemma 20 *Suppose $\mathcal{X}$ to be a α-strongly convex and centrally symmetric set with non-empty interior. Let $x \in \mathcal{X}$ such that $\|x\|_{\mathcal{X}} \leq 1-\gamma$ and if $\eta\|\widehat{\ell}_t\|_{\mathcal{X}^\circ} \leq \frac{1}{2}$,*

$$D_{R^*}(\nabla R(x) - \eta\widehat{\ell}_t, \nabla R(x)) \leq (1-\|x\|_{\mathcal{X}})\left(1+\frac{4(\alpha+1)}{\alpha}\right)\eta^2\|\widehat{\ell}_t\|^2_{\mathcal{X}^\circ}.$$

Proof Define $u = \nabla R(x) - \eta\widehat{\ell}_t$, $v = \nabla R(x)$ and $z = \frac{\|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}}$. By the definition of Bregman divergence and using Lemma 17, we have

$$\begin{aligned}
D_{R^*}(u,v) &= R^*(u) - R^*(v) - \langle \nabla R^*(v), u-v\rangle \\
&= \|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ} - \log\left(\frac{1+\|u\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}}\right) - \frac{\|v\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}}\langle \nabla\|\cdot\|_{\mathcal{X}^\circ}(v), u-v\rangle \\
&= z - \log(1+z) + \frac{1}{1+\|\nu\|_{\mathcal{X}^\circ}}\left[\|v\|_{\mathcal{X}^\circ}(\|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ}) - \|v\|_{\mathcal{X}^\circ}\langle \nabla\|\cdot\|_{\mathcal{X}^\circ}, u-v\rangle\right] \\
&= z - \log(1+z) - \frac{1}{2}\frac{(\|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ})^2}{1+\|v\|_{\mathcal{X}^\circ}} + \frac{D_{\frac{1}{2}\|\cdot\|^2_{\mathcal{X}^\circ}}(u,v)}{1+\|v\|_{\mathcal{X}^\circ}} \\
&\leq z - \log(1+z) + \frac{D_{\frac{1}{2}\|\cdot\|^2_{\mathcal{X}^\circ}}(u,v)}{1+\|v\|_{\mathcal{X}^\circ}}
\end{aligned}$$

Note that $z \geq -\frac{1}{2}$ as $\frac{\|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}} \geq -\eta\|\widehat{\ell}_t\|_{\mathcal{X}^\circ} \geq -\frac{1}{2}$, we have $z - \log(1+z) \leq z^2$. Therefore, we have

$$D_{R^*}(u,v) \leq \left(\frac{\|u\|_{\mathcal{X}^\circ} - \|v\|_{\mathcal{X}^\circ}}{1+\|v\|_{\mathcal{X}^\circ}}\right)^2 + \frac{1}{1+\|v\|_{\mathcal{X}^\circ}} D_{\frac{1}{2}\|\cdot\|^2_{\mathcal{X}^\circ}}(u,v).$$

Note that according to Lemma 15, we have $\frac{1}{1+\|v\|_{\mathcal{X}^\circ}} = 1 - \|x\|_{\mathcal{X}}$. Therefore, using triangle inequality leads to

$$D_{R^*}(u,v) \leq (1-\|x\|_{\mathcal{X}})^2\|u-v\|^2_{\mathcal{X}^\circ} + (1-\|x\|_{\mathcal{X}}) D_{\frac{1}{2}\|\cdot\|^2_{\mathcal{X}^\circ}}(u,v).$$

Finally, using Lemma 16, we have

$$\begin{aligned}
D_{R^*}(u,v) &\leq (1-\|x\|_{\mathcal{X}})^2\|u-v\|^2_{\mathcal{X}^\circ} + (1-\|x\|_{\mathcal{X}})\cdot\frac{4(\alpha+1)}{\alpha}\|u-v\|^2_{\mathcal{X}^\circ} \\
&\leq (1-\|x\|_{\mathcal{X}})\left(1+\frac{4(\alpha+1)}{\alpha}\right)\eta^2\|\widehat{\ell}_t\|^2_{\mathcal{X}^\circ}.
\end{aligned}$$

■

With the help of Lemma 19 and Lemma 20, we are able to bound BASE-REGRET.

Lemma 21 *For an arbitrary interval $\mathcal{I}$ started at round j, setting $\gamma = 4d\eta'$ for all $j' \in [T]$, Algorithm 6 ensures that the base regret of $\mathcal{B}_j$ with learning rate η (starting from round j) for any comparator $u \in \mathcal{X}'$ is at most*

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}}\left\langle a_t^{(j)} - u, \widehat{\ell}_t\right\rangle\right] \le \frac{\log(1/\gamma)}{\eta} + \frac{2d^{\frac{2}{p}}\eta}{1-\beta}\cdot\left(1+\frac{4(\alpha+1)}{\alpha}\right)\sum_{t\in\mathcal{I}}\frac{1-\|a_t^{(j)}\|_{\mathcal{X}}}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}}.$$

Proof Again, according to the standard analysis of OMD (see Lemma 27) we have

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}}\left\langle a_t^{(j)} - u, \widehat{\ell}_t\right\rangle\right] \le \frac{R(u)-R(a_j^{(j)})}{\eta} + \frac{1}{\eta}\sum_{t\in\mathcal{I}}\mathbb{E}\left[D_{R^*}\left(\nabla R(a_t^{(j)}) - \eta\widehat{\ell}_t, \nabla R(a_t^{(j)})\right)\right].$$

The first term can still be upper bounded by $\frac{\log(1/\gamma)}{\eta}$ as $a_j^{(j)} = \operatorname{argmin}_{x\in\mathcal{X}'} R(x)$ and $u \in \mathcal{X}' = \{x \mid \|x\|_{\mathcal{X}} \le 1-\gamma\}$, we have

$$R(u) - R(a_j^{(j)}) \le -\log(1-(1-\gamma)) - 0 = -\log\gamma.$$

For the second term, we will show that

$$\mathbb{E}_t\left[D_{R^*}\left(\nabla R(a_t^{(j)}) - \eta\widehat{\ell}_t, \nabla R(a_t^{(j)})\right)\right] \le \frac{2d^{\frac{2}{p}}\eta^2}{1-\beta}\cdot\left(1+\frac{4(\alpha+1)}{\alpha}\right)\sum_{t\in\mathcal{I}}\frac{1-\|a_t^{(j)}\|_{\mathcal{X}}}{1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}}.$$

According to Eq. (36) and the choice of η and γ, we have $\eta\|\widehat{\ell}_t\|_{\mathcal{X}^\circ} \le \frac{2d\eta}{\gamma} = \frac{1}{2}$. Based on Lemma 20, we only need to show that

$$\mathbb{E}_t\left[\|\widehat{\ell}_t\|_{\mathcal{X}^\circ}^2\right] \le \frac{2d^{\frac{2}{p}}}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})}.$$

In fact, according to the definition of $\widehat{\ell}_t$, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\|\widehat{\ell}_t\|_{\mathcal{X}^\circ}^2\right] \\
&\le \frac{d^2}{(1-\beta)^2(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})^2}\mathbb{E}_t\left[(1-\xi_t)^2\|x_t\|_{\mathcal{X}^\circ}^2\cdot|x_t^\top\ell_t|^2\cdot\mathbb{1}\{\rho_t=0\}\right] \\
&\le \frac{d^2}{(1-\beta)^2(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})^2}\mathbb{E}_t\left[(1-\beta)\sum_{j=1}^t\widehat{p}_{t,j}(1-\xi_t^{(j)})^2\|\widetilde{a}_t^{(j)}\|_{\mathcal{X}^\circ}^2\cdot|\widetilde{a}_t^{(j)\top}\ell_t|^2\right] \\
&\le \frac{d^2}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})^2}\sum_{j=1}^t\widehat{p}_{t,j}\mathbb{E}_t\left[(1-\|a_t^{(j)}\|_{\mathcal{X}})\cdot\frac{1}{d}\sum_{n=1}^d\|e_n\|_{\mathcal{X}^\circ}^2\cdot|\ell_{t,n}|^2\right].
\end{aligned}$$

Note that $\mathcal{X} \subseteq \ell_q(1)$, we have $\ell_p(1) \subseteq \mathcal{X}^\circ$, which means that $e_n \in \mathcal{X}^\circ$ and $\|e_n\|_{\mathcal{X}^\circ} \le 1$. Also using the fact that $\ell_p(1) \subseteq \mathcal{X}$, we have $\ell_t \in \mathcal{X}^\circ \subseteq \ell_q(1)$ and $\|\ell_t\|_2^2 \le d^{1-\frac{2}{q}}\|\ell_t\|_q^2 \le d^{1-\frac{2}{q}}$. Therefore, we have

$$\mathbb{E}_j\left[\|\widehat{\ell}_t\|_{\mathcal{X}^\circ}^2\right] \le \frac{2d^{\frac{2}{p}}\sum_{j=1}^t\widehat{p}_{t,j}(1-\|a_t^{(j)}\|_{\mathcal{X}})}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})^2} = \frac{2d^{\frac{2}{p}}}{(1-\beta)(1-\sum_{i=1}^t \widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}})},$$

which finishes the proof. ■

C.7. Bounding META-REGRET

In this section, we first prove several useful lemmas and then bound the term META-REGRET. We prove the following lemma, which is a counterpart of Lemma 8.

Lemma 22 *For an arbitrary interval $\mathcal{I} \subseteq [T]$ started at round j, setting $\frac{\varepsilon}{\lambda\gamma T} \leq \frac{1}{8}$, $\beta = 8d^{\frac{2}{p}}\varepsilon \leq \frac{1}{2}$ and $\mu = \frac{1}{T}$, Algorithm 6 guarantees that*

$$\sum_{t\in\mathcal{I}} \langle p_t - e_j, \widehat{c}_t\rangle \leq \frac{2\log T}{\varepsilon} + \varepsilon \sum_{t\in\mathcal{I}}\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2 + \mathcal{O}\left(\frac{|\mathcal{I}|}{\varepsilon T}\right). \tag{37}$$

Proof Define $v_{t+1,i} \triangleq \frac{p_{t,i}\exp(-\varepsilon\widehat{c}_{t,i})}{\sum_{t=1}^{T} p_{t,i}\exp(-\varepsilon\widehat{c}_{t,i})}$ for all $i \in [T]$. Then $p_{t+1,i} = \frac{\mu}{T} + (1-\mu)v_{t+1,i}$. Note that

$$\begin{aligned}
&\langle p_t, \widehat{c}_t\rangle + \frac{1}{\varepsilon}\log\left(\sum_{i=1}^{T} p_{t,i}\exp(-\varepsilon\widehat{c}_{t,i})\right)\\
&\leq \langle p_t, \widehat{c}_t\rangle + \frac{1}{\varepsilon}\log\left(\sum_{i=1}^{T} p_{t,i}(1-\varepsilon\widehat{c}_{t,i}+\varepsilon^2\widehat{c}_{t,i}^2)\right)\\
&= \langle p_t, \widehat{c}_t\rangle + \frac{1}{\varepsilon}\log\left(1-\varepsilon\langle p_t, \widehat{c}_t\rangle + \varepsilon^2\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2\right)\\
&\leq \varepsilon\sum_{i=1}^{T} p_{t,i}\widehat{c}_{t,i}^2.
\end{aligned}$$

The first inequality is because $\exp(-x) \leq 1 - x + x^2$ for $x \geq -\frac{1}{2}$ and according to the choice of ε, γ and λ, we have

$$\varepsilon\max_{i\in[t]}|\widehat{c}_{t,i}| \leq \varepsilon\max_{i\in[t]}\left|\widetilde{a}_t^{(i)\top}\widetilde{M}_t^{-1}x_t - b_{t,i}\right| \leq \varepsilon\max_{i\in[t]}\left|\widetilde{a}_t^{(i)\top}\widetilde{M}_t^{-1}x_t\right| + \varepsilon\max_{i\in[t]}|b_{t,i}|.$$

For the first term, by using Hölder's inequality, we have

$$\begin{aligned}
\varepsilon\max_{i\in[t]}\left|\widetilde{a}_t^{(i)\top}\widetilde{M}_t^{-1}x_t\right| &\leq \varepsilon\max_{i\in[t]}\|\widetilde{a}_t^{(i)}\|_{\mathcal{X}}\cdot\|\widetilde{M}_t^{-1}x_t\|_{\mathcal{X}^\circ}\\
&\leq \varepsilon\|\widetilde{M}_t^{-1}x_t\|_p && (\widetilde{a}_t^{(i)} \in \mathcal{X} \text{ and } \ell_p(1) \subseteq \mathcal{X}^\circ)\\
&\leq \varepsilon d^{\frac{1}{p}-\frac{1}{2}}\|\widetilde{M}_t^{-1}x_t\|_2\\
&\leq \frac{\varepsilon d^{\frac{1}{2}+\frac{1}{p}}}{\beta}\cdot\|x_t\|_2 && (\widetilde{M}_t \succeq \tfrac{\beta}{d}I)\\
&\leq \frac{\varepsilon d^{\frac{2}{p}}}{\beta}. && (\|x\|_2 \leq d^{\frac{1}{2}-\frac{1}{q}}\|x\|_q \leq d^{\frac{1}{p}-\frac{1}{2}})
\end{aligned}$$

In above argument, we use the fact that for vector $x \in \mathbb{R}^d$ and $0 < s < r$, we have $\|x\|_r \leq \|x\|_s \leq d^{\frac{1}{s}-\frac{1}{r}}\|x\|_r$. Moreover, note that $p \in (1,2]$ and $\ell_p(1) \subseteq \mathcal{X} \subseteq \ell_q(1)$.

For the second term, according to the definition of $b_{t,i}$, $|b_{t,i}| \le \frac{1}{\lambda T(1-\beta)\gamma} \le \frac{2}{\lambda T\gamma}$. Therefore, combining the above two bounds shows that $\varepsilon \max_{i\in[t]} |\widehat{c}_{t,i}| \le \frac{1}{8} + \frac{1}{4} \le \frac{1}{2}$ according to the choice of ε, γ, and λ. Furthermore, by the definition of $v_{t+1,i}$, we have $\sum_{j=1}^{T} p_{t,j} \exp(-\varepsilon \widehat{c}_{t,j}) = p_{t,i} \exp(-\varepsilon \widehat{c}_{t,i})/v_{t+1,i}$. Therefore, we have

$$\frac{1}{\varepsilon} \log \left(\sum_{j=1}^{T} p_{t,j} \exp(-\varepsilon \widehat{c}_{t,j}) \right) = -\frac{1}{\varepsilon} \log \left(\frac{v_{t+1,i}}{p_{t,i}} \right) - \widehat{c}_{t,i}.$$

Combining the two equations and taking summation over $t \in \mathcal{I}$, we have for any $e_j \in \Delta_T$, $j \in [T]$,

$$\sum_{t\in\mathcal{I}} \langle p_t, \widehat{c}_t \rangle - \sum_{t\in\mathcal{I}} \langle e_j, \widehat{c}_t \rangle \le \varepsilon \sum_{t\in\mathcal{I}} \sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 + \frac{1}{\varepsilon} \sum_{t\in\mathcal{I}} \log \left(\frac{v_{t+1,j}}{p_{t,j}} \right).$$

The second term can be dealt with according to Eq. (22) and we then have

$$\sum_{t\in\mathcal{I}} \langle p_t - e_j, \widehat{c}_t \rangle \le \frac{2 \log T}{\varepsilon} + \varepsilon \sum_{t\in\mathcal{I}} \sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 + \mathcal{O}\left(\frac{|\mathcal{I}|}{\varepsilon T} \right),$$

which finishes the proof. ■

Next, we prove the following lemma, which bounds the second term shown in Eq. (37)

Lemma 23 *For any $t \in [T]$, setting $\lambda^2\gamma = \Theta\left(d^{-\frac{1}{q}} \sqrt{\frac{1}{ST^3}}\right)$, Algorithm 6 guarantees that*

$$\sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 \le \sum_{i\in[t]} \widehat{p}_{t,i} \widehat{c}_{t,i}^2 \le 2 \sum_{i\in[t]} \widehat{p}_{t,i} c_{t,i}^2 + \mathcal{O}\left(d^{\frac{1}{q}} \sqrt{\frac{S}{T}} \right), \tag{38}$$

where $c_{t,i} = \langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \rangle$.

Proof According to the definition of $\widehat{p}_t$ and $\widehat{c}_t$, we have

$$\begin{aligned}
\sum_{i=1}^{T} p_{t,i} \widehat{c}_{t,i}^2 &= \sum_{i\in[t]} p_{t,i} \widehat{c}_{t,i}^2 + \sum_{i>t} p_{t,i} \left(\sum_{j=1}^{t} \widehat{p}_{t,j} \widehat{c}_{t,j} \right)^2 \le \sum_{i\in[t]} p_{t,i} \widehat{c}_{t,i}^2 + \sum_{i\notin[t]} p_{t,i} \left(\sum_{j\in[t]} \widehat{p}_{t,j} \widehat{c}_{t,j}^2 \right) \\
&= \left(\sum_{i\in[t]} p_{t,i} \right) \left(\sum_{i\in[t]} \widehat{p}_{t,i} \widehat{c}_{t,i}^2 \right) + \left(\sum_{i\notin[t]} p_{t,i} \right) \left(\sum_{i\in[t]} \widehat{p}_{t,i} \widehat{c}_{t,i}^2 \right) = \sum_{i\in[t]} \widehat{p}_{t,i} \widehat{c}_{t,i}^2,
\end{aligned}$$

where the inequality is because of Cauchy-Schwarz inequality. Besides, recall that $c_{t,i} = \left\langle \widetilde{a}_t^{(i)}, \bar{\ell}_t \right\rangle$ and $\widehat{c}_{t,i}^2 = (c_{t,i} - b_{t,i})^2 \le 2c_{t,i}^2 + 2b_{t,i}^2$. According to the definition of $b_{t,i}$, we know that

$$\sum_{i\in[t]} \widehat{p}_{t,i} b_{t,i}^2 \le \frac{4}{(\lambda T)^2} \frac{1}{\gamma} \sum_{i\in[t]} \widehat{p}_{t,i} \frac{1 - \|a_t^{(i)}\|_{\mathcal{X}}}{1 - \sum_{i\in[t]} \widehat{p}_{t,i} \|a_t^{(i)}\|_{\mathcal{X}}} = \frac{4}{(\lambda T)^2} \frac{1}{\gamma} = \mathcal{O}\left(d^{\frac{1}{q}} \sqrt{\frac{S}{T}} \right),$$

where the last step holds because we choose $\lambda^2\gamma = \Theta\left(d^{-\frac{1}{q}} \sqrt{\frac{1}{ST^3}}\right)$. ■

Combining Lemma 22 and Lemma 23, we obtain the following lemma showing the upper bound for META-REGRET, which is exactly the same as Lemma 10 except for the choice of parameters.

Lemma 24 *Define* $C = \sqrt{\frac{\alpha}{10\alpha+8}}$. *Set* $\varepsilon = \min\left\{d^{-\frac{1}{p}}\sqrt{\frac{S}{T}}, \frac{1}{16d^{\frac{2}{p}}}, \frac{C^2}{2}\right\}$, $\beta = 8d^{\frac{2}{p}}\varepsilon$, $\lambda = \frac{Cd^{-\frac{1}{q}}}{\sqrt{ST}}$, $\gamma = 4Cd^{\frac{1}{q}}\sqrt{\frac{S}{T}}$ *and* $\mu = \frac{1}{T}$. *Algorithm 6 guarantees that*

$$\mathbb{E}\left[\textsc{Meta-Regret}\right] \leq \widetilde{\mathcal{O}}\left(d^{\frac{1}{p}}\sqrt{ST}\right).$$

Proof First, it is direct to check that the choice of λ, γ and ε satisfies the condition required in Lemma 22 and Lemma 23. Based on the two lemmas, for each interval $\mathcal{I}_k$, let j_k be the start time stamp for $\mathcal{I}_k$. As $\beta = 8d^{\frac{2}{p}}\varepsilon \leq \frac{1}{2}$, we follow the derivation of Eq. (24) and obtain that

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}_k}\langle p_t - e_{j_k}, \widehat{c}_t\rangle\right] \leq \frac{2\log T}{\varepsilon} + 4\varepsilon d|\mathcal{I}_k| + \mathcal{O}\left(\varepsilon|\mathcal{I}_k|d^{\frac{1}{q}}\sqrt{\frac{S}{T}}\right) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{\varepsilon T}\right).$$

Summing the regret over all the intervals achieves the bound for META-REGRET:

$$\begin{aligned}
\mathbb{E}\left[\textsc{Meta-Regret}\right] &= \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\langle \widehat{p}_t - e_{j_k}, \widehat{c}_t\rangle\right] \\
&\leq \frac{2S\log T}{\varepsilon} + 4\varepsilon dT + \mathcal{O}\left(d^{\frac{1}{q}}\sqrt{ST}\right) + \mathcal{O}(1/\varepsilon) \leq \widetilde{\mathcal{O}}\left(d^{\frac{1}{p}}\sqrt{ST}\right),
\end{aligned}$$

where the last inequality is because we choose $\varepsilon = \min\left\{d^{-\frac{1}{p}}\sqrt{\frac{S}{T}}, \frac{C^2}{2}, \frac{1}{16d^{\frac{2}{p}}}\right\}$. ■

C.8. Proof of Theorem 14

Putting everything together, we are now ready to prove our main result (Theorem 14) in the setting when the feasible domain is α-strongly convex.

Proof First, it is evident to check that the parameter choice satisfies the condition required in Lemma 21 and Lemma 24. Therefore, based on the regret decomposition in Eq. (33), upper bound of bias term in Eq. (34), upper bound of positive term Eq. (35), base regret upper bound in Lemma 21 and meta regret upper bound in Lemma 24, we have

$$\begin{aligned}
&\mathbb{E}[\textsc{Reg}_S] = \mathbb{E}\left[\sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\langle x_t - \mathring{u}_k, \ell_t\rangle\right] \\
&\leq \frac{2}{\lambda} + \sum_{k=1}^{S}\frac{\log(1/\gamma)}{\eta_{j_k}} + \left(\frac{2d^{\frac{2}{p}}\eta}{(1-\beta)}\cdot\frac{5\alpha+4}{\alpha} - \frac{1}{\lambda T(1-\beta)}\right)\sum_{t\in\mathcal{I}_k}\frac{1-\|a_t^{(j_k)}\|_{\mathcal{X}}}{1-\sum_{i=1}^{t}\widehat{p}_{t,i}\|a_t^{(i)}\|_{\mathcal{X}}} \\
&\quad + (\beta+\gamma)T + \widetilde{\mathcal{O}}\left(d^{\frac{1}{p}}\sqrt{ST}\right).
\end{aligned}$$

Importantly, note that the coefficient of the third term is actually zero. Indeed, due to the parameter configurations that $\gamma = 4Cd^{\frac{1}{q}}\sqrt{\frac{S}{T}}$, $\eta = Cd^{-\frac{1}{p}}\sqrt{\frac{S}{T}}$, $\lambda = \frac{Cd^{-\frac{1}{q}}}{\sqrt{ST}}$, $\beta = 8d^{\frac{2}{p}}\varepsilon$, $\varepsilon =$

$\min\left\{\frac{1}{16d^{\frac{2}{p}}}, \frac{C^2}{2}, d^{-\frac{1}{p}}\sqrt{\frac{S}{T}}\right\}$ and $C = \sqrt{\frac{\alpha}{10\alpha+8}}$, we can verify that

$$\frac{2d\eta(5\alpha+4)}{\alpha} - \frac{1}{\lambda T} = 0.$$

Then we can achieve $\mathbb{E}[\text{REG}_S] \leq \widetilde{\mathcal{O}}\left(d^{\frac{1}{p}}\sqrt{ST}\right)$ and complete the proof. ■

Appendix D. Omitted Details for Section 5

In this section, we consider the switching regret of unconstrained linear bandits.

D.1. Proof of Lemma 2

Proof Our switching regret decomposition for linear bandits is inspired by the existing black-box reduction designed for the full information online convex optimization (Cutkosky and Orabona, 2018) and static regret of linear bandits (van der Hoeven et al., 2020). Indeed, the switching regret can be decomposed in the following way.

$$\begin{aligned}
\text{REG}(u_1, \ldots, u_T) &= \sum_{t=1}^{T} \ell_t^\top x_t - \sum_{t=1}^{T} \ell_t^\top u_t \\
&= \sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \ell_t^\top x_t - \sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \ell_t^\top \mathring{u}_k \\
&= \sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \ell_t^\top (z_t \cdot v_t - \mathring{u}_k) \qquad (x_t = z_t \cdot v_t) \\
&= \sum_{k=1}^{S} \left(\sum_{t \in \mathcal{I}_k} \langle z_t, \ell_t \rangle (v_t - \|\mathring{u}_k\|_2) + \|\mathring{u}_k\|_2 \sum_{t \in \mathcal{I}_k} \left\langle z_t - \frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}, \ell_t \right\rangle \right) \\
&= \sum_{k=1}^{S} \text{REG}^{\mathcal{V}}_{\mathcal{I}_k}(\|\mathring{u}_k\|_2) + \sum_{k=1}^{S} \|\mathring{u}_k\|_2 \cdot \text{REG}^{\mathcal{Z}}_{\mathcal{I}_k}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right),
\end{aligned}$$

which finishes the proof. ■

D.2. Algorithm for Unconstrained OCO with Switching Regret

In this section, we present the details of our proposed algorithm for unconstrained OCO with switching regret.

Under the unconstrained setup, the diameter of the feasible domain is $D = \infty$. However, as observed in (Chen et al., 2021, Appendix D.5), we can simply assume $\max_{k \in [S]} \|\mathring{u}_k\|_2 \leq 2^T$. Otherwise, we will have $T \leq \log_2(\max_{k \in [S]} \|\mathring{u}_k\|_2)$, and by constraining the learning algorithm such that $\|v_t\|_2 \leq 2^T$, we can obtain the following trivial upper bound for switching regret: $\text{REG} \leq \sum_{t=1}^{T} \|\nabla f_t(v_t)\|_2 \|v_t - u_t\|_2 \leq T(2^T + \max_{k \in [S]} \|\mathring{u}_k\|_2) = \widetilde{\mathcal{O}}(\max_{k \in [S]} \|\mathring{u}_k\|_2)$, which is already adaptive to the comparators. Therefore, we can simply focus on the *constrained* online learning with a maximum diameter $D = 2^T$. In addition, as mentioned earlier, we do not assume the knowledge of the number of switch S in advance in this part. To this end, we propose a two-layer approach to simultaneously adapt to the unknown scales of the comparators and the unknown number of switch, which consists of a meta algorithm learning over a set of base learners. Below we specify the details.

Base algorithm. The base algorithm tackles OCO problem with a given scale of feasible domain. The only requirement is as follows: given a constrained domain $\mathcal{X} \subseteq \mathbb{R}^d$ with diameter $D = \sup_{x\in\mathcal{X}}\|x\|_2$, base algorithm running over $\mathcal{X}$ ensures an $\widetilde{\mathcal{O}}(D\sqrt{|\mathcal{I}|})$ static regret over any interval $\mathcal{I} \subseteq [T]$. Formally, we assume the base algorithm to satisfy the following requirement.

Requirement 1 *Consider the online convex optimization problem consisting a convex feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$ and a sequence of convex loss functions $f_1, \ldots, f_T$, where $f_t : \mathcal{X} \mapsto \mathbb{R}$ and we assume $\mathbf{0} \in \mathcal{X}$ and $\|\nabla f_t(v)\|_2 \leq 1$ for any $v \in \mathcal{X}$ and $t \in [T]$. An online algorithm $\mathcal{A}$ running over this problem returns the decision sequence $v_1, \ldots, v_T \in \mathcal{X}$. We require the algorithm $\mathcal{A}$ to ensure the following regret guarantee*

$$\sum_{t\in\mathcal{I}} f_t(v_t) - \min_{u\in\mathcal{X}} \sum_{t\in\mathcal{I}} f_t(u) \leq \widetilde{\mathcal{O}}\left(D\sqrt{|\mathcal{I}|}\right) \tag{39}$$

for any interval $\mathcal{I} \subseteq [T]$, where $D = \sup_{x\in\mathcal{X}}\|x\|_2$ is the diameter of the feasible domain.

This requirement can be satisfied by recent OCO algorithms with *interval regret* (or called *strongly adaptive regret*) guarantee, such as Algorithm 1 of Daniely et al. (2015), Algorithm 2 of Jun et al. (2017), Theorem 6 of Cutkosky (2020). We denote by $\mathfrak{B}$ any suitable base algorithm.

Since both the scale of comparators and the number of switch are unknown in advance, we maintain a set of base algorithm instances, defined as

$$\mathcal{S} = \left\{\mathcal{B}_{i,r}, \forall (i,r) \in [H]\times[R] \;\middle|\; \mathcal{B}_{i,r} \leftarrow \mathfrak{B}(\mathcal{X}_i), \text{ with } \mathcal{X}_i = \{x \mid \|x\|_2 \leq D_i = T^{-1}\cdot 2^{i-1}\}\right\}. \tag{40}$$

In above, $H = \lceil\log_2 T\rceil + T + 1$ and the index $i \in [H]$ maintain a grid to deal with uncertainty of unknown comparators' scale; $R = \lceil\log_2 T\rceil$ and the index $r \in [R]$ maintains a grid to handle uncertainty of unknown number of switch S. There are in total $N = H \cdot R$ base learners. For $i \in [H]$ and $r \in [R]$, the base learner $\mathcal{B}_{i,r}$ is an instantiation of the base algorithm whose feasible domain is $\mathcal{X}_i \subseteq \mathbb{R}^d$ with diameter D_i, and $v_{t,(i,r)}$ denotes her returned decision at round t. We stress that even if S is known, the two-layer structure remains necessary due to the unknown comparators' scale.

Meta algorithm. Then, a meta algorithm is used to combine all those base learners, and more importantly, the regret of meta algorithm should be adaptive to the individual loss scale of each base learner, such that the overall algorithm can achieve a comparator-adaptive switching regret. We achieve so by building upon the recent progress in the classic expert problem (Chen et al., 2021). Our proposed algorithm is OMD with a multi-scale entropy regularizer and certain important correction terms. Specifically, let the weight vector produced by the meta algorithm be $w_t \in \Delta_N$, then the overall decision is $v_t = \sum_{i=1}^H \sum_{r=1}^R w_{t,(i,r)} v_{t,(i,r)}$, and the weight is updated by

$$w_{t+1} = \operatorname*{argmin}_{w\in\Omega} \langle w, \ell_t + a_t\rangle + D_\psi(w, w_t), \tag{41}$$

where $\Omega = \{w \mid w \in \Delta_N \text{ and } w_{t,(i,r)} \geq \frac{1}{T^2 \cdot 2^{2i}}, \forall i \in [H], r \in [R]\}$ is the clipped domain. Besides, the meta loss ℓ_t, the correction term a_t, and a certain regularizer ψ are set as follows:

- The regularizer $\psi : \Delta_N \mapsto \mathbb{R}$ is set as a *weighted* negative-entropy regularizer defined as

$$\psi(w) \triangleq \sum_{(i,r)\in[H]\times[R]} \frac{c_i}{\eta_r} w_{(i,r)} \log w_{(i,r)} \text{ with } c_i = T^{-1} \cdot 2^{i-1} \text{ and } \eta_r = \frac{1}{32 \cdot 2^r}. \tag{42}$$

- The feedback loss of meta algorithm $\ell_t \in \mathbb{R}^N$ is set as such to measure the quality of each base learner: $\ell_{t,(i,r)} \triangleq \langle \nabla f_t(v_t), v_{t,(i,r)} \rangle$ for any $(i,r) \in [H] \times [R]$.

- The correction term $a_t \in \mathbb{R}^N$ is set as: $a_{t,(i,r)} \triangleq 32 \frac{\eta_r}{c_i} \ell^2_{t,(i,r)}$ for any $(i,r) \in [H] \times [R]$, which is essential to ensure the meta regret compatible to the final comparator-adaptive bound.

The entire algorithm consists of meta algorithm specified above and base algorithm satisfying Requirement 1. We show the pseudocode in Algorithm 4.

D.3. Proof of Theorem 3

Proof Consider the k-th interval $\mathcal{I}_k$. The regret within this interval can be decomposed as follows.

$$\begin{aligned}\sum_{t\in\mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}_k)\Big) &= \sum_{t\in\mathcal{I}_k} \Big(f_t(v_t) - f_t(v_{t,j})\Big) + \sum_{t\in\mathcal{I}_k} \Big(f_t(v_{t,j}) - f_t(\mathring{u}_k)\Big) \\ &\leq \sum_{t\in\mathcal{I}_k} \langle \nabla f_t(v_t), v_t - v_{t,j} \rangle + \sum_{t\in\mathcal{I}_k} \Big(f_t(v_{t,j}) - f_t(\mathring{u}_k)\Big) \\ &= \underbrace{\sum_{t\in\mathcal{I}_k} \langle w_t - e_j, \ell_t \rangle}_{\text{META-REGRET}} + \underbrace{\sum_{t\in\mathcal{I}_k} \Big(f_t(v_{t,j}) - f_t(\mathring{u}_k)\Big)}_{\text{BASE-REGRET}}, \end{aligned} \tag{43}$$

where the final equality is because $\ell_{t,j} = \langle \nabla f_t(v_t), v_{t,j} \rangle$ and $v_t = \sum_{j'\in[H]\times[R]} w_{t,j'} v_{t,j'}$. Note that the decomposition holds for any index $j = (i,r) \in [H] \times [R]$.

We first consider the case when $\|\mathring{u}_k\|_2 \geq \frac{1}{T}$ and will deal with the other case (when $\|\mathring{u}_k\|_2 < \frac{1}{T}$) at the end of the proof. Under such a circumstance, we can choose $(i,r) = (i_k^*, r_k^*)$ such that

$$\begin{aligned} c_{i_k^*} &= T^{-1} \cdot 2^{i_k^*-1} \leq \|\mathring{u}_k\|_2 \leq T^{-1} \cdot 2^{i_k^*} = c_{i_k^*+1}, \text{ and} \\ \eta_{r_k^*} &= \frac{1}{32 \cdot 2^{r_k^*}} \leq \frac{1}{32\sqrt{|\mathcal{I}_k|}} \leq \frac{1}{32 \cdot 2^{r_k^*-1}} = \eta_{r_k^*-1}, \end{aligned} \tag{44}$$

which is valid as $i \in [H] = [\lceil \log_2 T \rceil + T + 1]$ and $r \in [R] = [\lceil \log_2 T \rceil]$. We now give the upper bounds for META-REGRET and BASE-REGRET respectively.

BASE-REGRET. Based on the assumption of base algorithm, we have base learner $\mathcal{B}_{j_k^*}$ satisfying

$$\sum_{t\in\mathcal{I}_k} \Big(f_t(v_{t,j_k^*}) - f_t(\mathring{u}_k)\Big) \leq \widetilde{\mathcal{O}}\left(c_{i_k^*}\sqrt{|\mathcal{I}_k|}\right) \leq \widetilde{\mathcal{O}}\left(\|\mathring{u}_k\|_2\sqrt{|\mathcal{I}_k|}\right), \tag{45}$$

where we use the interval regret guarantee of base algorithm (see Requirement 1) and also use the fact that the diameter of the feasible domain for base learner $\mathcal{B}_{j_k^*}$ is $2^{i_k^*}$ as $\mathcal{X}_{i_k^*} = \{x \mid \|x\|_2 \leq D_{i_k^*}\}$ and $D_{i_k^*} = c_{i_k^*}$. The last inequality holds by the choice of i_k^* shown in Eq. (44).

META-REGRET. The meta algorithm is essentially online mirror descent with a weighted entropy regularizer. Based on Lemma 1 in (Chen et al., 2021), if for all $i \in [H]$ and $r \in [R]$, $32\frac{\eta_r}{c_i}|\ell_{t,(i,r)}| \leq 1$, then we have for any $q \in \Omega$,

$$\sum_{t\in\mathcal{I}_k} \langle w_t - q, \ell_t \rangle \leq \sum_{t\in\mathcal{I}_k} \Big(D_\psi(q, w_t) - D_\psi(q, w_{t+1}) \Big) + 32 \sum_{t\in\mathcal{I}_k} \sum_{i\in[H]} \sum_{r\in[R]} \frac{\eta_r}{c_i} q_{(i,r)} \ell_{t,(i,r)}^2. \tag{46}$$

Note that this is a simplified version of Lemma 1 in (Chen et al., 2021) for the interval regret, which employs a fixed learning rate for each action and does not include the optimism in the algorithm. We present the simplified lemma in Lemma 28 in Appendix E for completeness.

To this end, we first verify the condition of $32\frac{\eta_r}{c_i}|\ell_{t,(i,r)}| \leq 1$ for all $i \in [H]$, $r \in [R]$. In fact,

$$\frac{32\eta_r}{c_i} \left|\ell_{t,(i,r)}\right| \leq \frac{1}{c_i \cdot 2^r} \|\nabla f_t(v_t)\|_2 \cdot \|v_{t,i}\|_2 \leq \frac{1}{2^r} \leq 1,$$

where the first inequality is by the definition of $\eta_r = \frac{1}{32 \cdot 2^r}$ and the construction of meta loss $\ell_{t,(i,r)} = \langle \nabla f_t(v_t), v_{t,(i,r)} \rangle$, the second inequality is because $\|v_{t,i}\|_2 \leq c_i$ and $\|\nabla f_t(v)\|_2 \leq 1$ for all $v \in \mathbb{R}^d$, and the third inequality holds as $r \geq 1$.

Then we define $\bar{e}_{j_k^*} \triangleq \bar{e}_{(i_k^*, r_k^*)} = \left(1 - \frac{R \cdot a_0}{T^2}\right) e_{(i_k^*, r_k^*)} + \sum_{(i,r)\in[H]\times[R]} \frac{1}{T^2 \cdot 2^{2i}} e_{(i,r)}$, where $a_0 = \sum_{i=1}^H \frac{1}{2^{2i}} = \frac{1}{3}(1 - \frac{1}{4^H})$ is a constant which guarantees $\bar{e}_{j_k^*} \in \Omega$. Using Eq. (46) with $q = \bar{e}_{j_k^*}$, we have

$$\begin{aligned}
\sum_{t\in\mathcal{I}_k} \left\langle w_t - \bar{e}_{j_k^*}, \ell_t \right\rangle &\leq \sum_{t\in\mathcal{I}_k} \Big(D_\psi(\bar{e}_{j_k^*}, w_t) - D_\psi(\bar{e}_{j_k^*}, w_{t+1}) \Big) + 32 \sum_{t\in\mathcal{I}_k} \sum_{i\in[H]} \sum_{r\in[R]} \frac{\eta_r}{c_i} \bar{e}_{j_k^*,(i,r)} \ell_{t,(i,r)}^2 \\
&= \Big(D_\psi(\bar{e}_{j_k^*}, w_{s_k}) - D_\psi(\bar{e}_{j_k^*}, w_{s_{k+1}}) \Big) + 32 \sum_{t\in\mathcal{I}_k} \sum_{i\in[H]} \sum_{r\in[R]} \frac{\eta_r}{c_i} \bar{e}_{j_k^*,(i,r)} \ell_{t,(i,r)}^2,
\end{aligned}$$

where s_k denotes the starting index of the interval $\mathcal{I}_k$ and s_{k+1} is defined as $T + 1$ if $\mathcal{I}_k$ is the last interval. The two terms on the right-hand side are called bias term and stability term respectively. In the following, we will give their upper bound individually.

For the bias term, we have

$$\begin{aligned}
&D_\psi(\bar{e}_{j_k^*}, w_{s_k}) - D_\psi(\bar{e}_{j_k^*}, w_{s_{k+1}}) \\
&= \sum_{i\in[H]} \sum_{r\in[R]} \frac{c_i}{\eta_r} \left(\bar{e}_{j_k^*,(i,r)} \log \frac{w_{s_k,(i,r)}}{w_{s_{k+1},(i,r)}} + w_{s_k,(i,r)} - w_{s_{k+1},(i,r)} \right) \qquad \text{(by definition in Eq. (42))} \\
&= \sum_{i\in[H]} \sum_{r\in[R]} \frac{c_i}{\eta_r} \left(\bar{e}_{j_k^*,(i,r)} \log \frac{w_{s_k,(i,r)}}{w_{s_{k+1},(i,r)}} \right) + \sum_{i\in[H]} \sum_{r\in[R]} \frac{c_i}{\eta_r} \Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)} \Big) \\
&\leq \frac{c_{i_k^*}}{\eta_{r_k^*}} \log \left(T^2 \cdot 2^{2i_k^*} \right) + \sum_{(i,r)\neq(i_k^*,r_k^*)} \frac{1}{T^2 \cdot 2^{2i}} \frac{c_i}{\eta_r} \log \left(T^2 \cdot 2^{2i} \right) + \sum_{i\in[H]} \sum_{r\in[R]} \frac{c_i}{\eta_r} \Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)} \Big) \\
&\hspace{30em} (w_{s_k}, w_{s_{k+1}} \in \Omega) \\
&\leq \frac{c_{i_k^*}}{\eta_{r_k^*}} \log \left(4T^4 \cdot c_{i_k^*}^2 \right) + \sum_{(i,r)\neq(i_k^*,r_k^*)} \frac{2 \log T + (4 \log 2) \cdot T}{32 \cdot T^3 \cdot 2^{i+r+1}} + \sum_{i\in[H]} \sum_{r\in[R]} \frac{c_i}{\eta_r} \Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)} \Big)
\end{aligned}$$

$$= \widetilde{\mathcal{O}}\left(\frac{c_{i_k^*}}{\eta_{r_k^*}}\log c_{i_k^*}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i\in[H]}\sum_{r\in[R]}\frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big). \tag{47}$$

Moreover, for the stability term, we have

$$\begin{aligned}
&32\sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{\eta_r}{c_i}\bar{e}_{j_k^*,(i,r)}\ell_{t,(i,r)}^2\\
&= 32\sum_{t\in\mathcal{I}_k}\frac{\eta_{r_k^*}}{c_{i_k^*}}\left(1 - \frac{R\cdot a_0}{T^2} + \frac{1}{T^2\cdot 2^{2i_k^*}}\right)\ell_{t,(i_k^*,r_k^*)}^2 + 32\sum_{t\in\mathcal{I}_k}\sum_{(i,r)\neq(i_k^*,r_k^*)}\frac{\eta_r}{c_i}\bar{e}_{j_k^*,(i,r)}\ell_{t,(i,r)}^2\\
&\le 32\sum_{t\in\mathcal{I}_k}\frac{\eta_{r_k^*}}{c_{i_k^*}}\ell_{t,(i_k^*,r_k^*)}^2 + 32\sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{\eta_r c_i}{T^2\cdot 2^{2i}}\\
&\le \mathcal{O}\left(\eta_{r_k^*}c_{i_k^*}|\mathcal{I}_k|\right) + \sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{1}{T^3\cdot 2^{i+r+1}}\\
&= \mathcal{O}\left(\eta_{r_k^*}c_{i_k^*}|\mathcal{I}_k|\right) + \mathcal{O}\left(\frac{1}{T^2}\right)
\end{aligned} \tag{48}$$

where the two inequalities hold as $\ell_{t,(i,r)}^2 = \langle\nabla f_t(v_t), v_{t,(i,r)}\rangle^2 \le \|\nabla f_t(v_t)\|_2^2\|v_{t,(i,r)}\|_2^2 \le c_i^2$. Combining the upper bounds of bias term in Eq. (47) and stability term in Eq. (48), we get

$$\sum_{t\in\mathcal{I}_k}\left\langle w_t - \bar{e}_{j_k^*}, \ell_t\right\rangle \le \widetilde{\mathcal{O}}\left(\eta_{r_k^*}c_{i_k^*}|\mathcal{I}_k| + \frac{c_{i_k^*}}{\eta_{r_k^*}}\log c_{i_k^*}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i\in[H]}\sum_{r\in[R]}\frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big).$$

Further, notice that

$$\sum_{t\in\mathcal{I}_k}\left\langle \bar{e}_{j_k^*} - e_{j_k^*}, \ell_t\right\rangle \le \sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{1}{T^2\cdot 2^{2i}}\cdot\ell_{t,(i,r)} \le \sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{1}{T^3\cdot 2^{i+1}} \le \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right), \tag{49}$$

and we thus obtain the overall meta regret upper bound in the interval $\mathcal{I}_k$:

$$\begin{aligned}
&\sum_{t\in\mathcal{I}_k}\left\langle w_t - e_{j_k^*}, \ell_t\right\rangle = \sum_{t\in\mathcal{I}_k}\left\langle w_t - \bar{e}_{j_k^*}, \ell_t\right\rangle + \sum_{t\in\mathcal{I}_k}\left\langle \bar{e}_{j_k^*} - e_{j_k^*}, \ell_t\right\rangle\\
&\le \widetilde{\mathcal{O}}\left(\eta_{r_k^*}c_{i_k^*}|\mathcal{I}_k| + \frac{c_{i_k^*}}{\eta_{r_k^*}}\log c_{i_k^*}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i\in[H]}\sum_{r\in[R]}\frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big)\\
&= \widetilde{\mathcal{O}}\left(\|\mathring{u}_k\|_2\sqrt{|\mathcal{I}_k|}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i\in[H]}\sum_{r\in[R]}\frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big),
\end{aligned} \tag{50}$$

where the last inequality is because of the choice of i_k^* and r_k^* defined in Eq. (44). The $\widetilde{\mathcal{O}}(\cdot)$-notation omits logarithmic dependence on T and comparator norm $\|\mathring{u}_k\|_2$.

OVERALL REGRET. The overall regret is obtained by combining the base regret and meta regret and further summing over all the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_S$. Indeed, we have the following total meta-regret by taking summation over intervals on Eq. (50),

$$
\begin{aligned}
&\sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \left\langle w_t - e_{j_k^*}, \ell_t \right\rangle \\
&\le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{|\mathcal{I}_k|} \right) + \widetilde{\mathcal{O}}\left(\frac{S}{T^2} \right) + \sum_{k=1}^{S} \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r} \Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)} \Big) \\
&\le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{|\mathcal{I}_k|} \right) + \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r} w_{1,(i,r)} \\
&= \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{|\mathcal{I}_k|} \right),
\end{aligned} \tag{51}
$$

where the final equality is because we choose $w_{1,(i,r)} \propto \frac{\eta_r^2}{c_i^2}$ for all $(i,r) \in [H] \times [R]$. Indeed, such a setting of prior distribution ensures that

$$
\begin{aligned}
\sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r} \cdot w_{1,(i,r)} &= \frac{\sum_{i \in [H]} \sum_{r \in [R]} \frac{\eta_r}{c_i}}{\sum_{i \in [H]} \sum_{r \in [R]} \frac{\eta_r^2}{c_i^2}} = \frac{16}{T} \cdot \frac{\sum_{i \in [H]} \sum_{r \in [R]} \frac{1}{2^{i+r}}}{\sum_{i \in [H]} \sum_{r \in [R]} \frac{1}{2^{2i+2r}}} \\
&= \frac{144}{T} \cdot \frac{\left(1 - \left(\frac{1}{2}\right)^R\right)\left(1 - \left(\frac{1}{2}\right)^H\right)}{\left(1 - \left(\frac{1}{4}\right)^R\right)\left(1 - \left(\frac{1}{4}\right)^H\right)} = \frac{144}{T} \cdot \frac{1}{\left(1 + \left(\frac{1}{2}\right)^R\right)\left(1 + \left(\frac{1}{2}\right)^H\right)} \le \mathcal{O}\left(\frac{1}{T}\right),
\end{aligned}
$$

and also guarantees that $w_1 \in \Omega$ since for any $(i,r) \in [H] \times [R]$,

$$
\begin{aligned}
w_{1,(i,r)} &= \frac{\frac{\eta_r^2}{c_i^2}}{\sum_{i' \in [H]} \sum_{r' \in [R]} \frac{\eta_{r'}^2}{c_{i'}^2}} = \frac{\frac{1}{2^{2i+2r}}}{\sum_{i' \in [H]} \sum_{r' \in [R]} \frac{1}{2^{2i'+2r'}}} \\
&\ge \frac{1}{T^2 \cdot 2^{2i}} \cdot \frac{1}{\frac{1}{9}\left(1 - \left(\frac{1}{4}\right)^R\right)\left(1 - \left(\frac{1}{4}\right)^H\right)} \ge \frac{1}{T^2 \cdot 2^{2i}},
\end{aligned}
$$

where the first inequality holds in that we have $2^r \le T$ for $r \in [R]$.

Substituting the meta regret upper bound Eq. (51) and the base regret upper bound Eq. (45) into the regret decomposition Eq. (43) obtains that

$$
\sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}_k) \Big) \le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{|\mathcal{I}_k|} \right) \le \widetilde{\mathcal{O}}\left(\max_{k \in [S]} \|\mathring{u}_k\|_2 \cdot \sqrt{ST} \right), \tag{52}
$$

which finishes the proof for the case when $\|\mathring{u}_k\|_2 \ge \frac{1}{T}$ holds for every $k \in [S]$.

We now consider the case when the condition is violated. Suppose for some $k \in [S]$, it holds that $\|\mathring{u}_k\|_2 < \frac{1}{T}$. Then, we pick any $\mathring{u}_k' \in \mathbb{R}^d$ such that $\|\mathring{u}_k'\|_2 = \frac{1}{T}$, and obtain that

$$
\sum_{t \in \mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}_k) \Big) = \sum_{t \in \mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}_k') \Big) + \sum_{t \in \mathcal{I}_k} \Big(f_t(\mathring{u}_k') - f_t(\mathring{u}_k) \Big)
$$

$$
\begin{aligned}
&= \sum_{t \in \mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}'_k) \Big) + \sum_{t \in \mathcal{I}_k} \|\nabla f_t(\mathring{u}'_k)\|_2 \|\mathring{u}'_k - \mathring{u}_k\|_2 \\
&\leq \sum_{t \in \mathcal{I}_k} \Big(f_t(v_t) - f_t(\mathring{u}'_k) \Big) + \mathcal{O}\left(\frac{|\mathcal{I}_k|}{T} \right).
\end{aligned}
$$

Clearly, the last additional term will not be the issue even after summation over S intervals. Moreover, notice that now the comparator $\mathring{u}'_k$ satisfies the condition of $\|\mathring{u}'_k\|_2 \geq \frac{1}{T}$, we can still use the earlier results including the base regret bound in Eq. (45) and meta regret bound in Eq. (50). Thus, we can guarantee the same regret bound as Eq. (52) under this scenario.

Hence, we finish the proof for the overall theorem. We finally remark that our algorithm for unconstrained OCO actually does not require the knowledge of S ahead of time. ■

D.4. Data-dependent Switching Regret of Unconstrained Online Convex Optimization

In this subsection, we further consider achieving data-dependent switching regret bound for unconstrained online convex optimization.

In Appendix D.2, we require the base algorithm to achieve an $\widetilde{\mathcal{O}}(D\sqrt{|\mathcal{I}|})$ interval regret for any interval $\mathcal{I} \subseteq [T]$, where D is the diameter of the feasible domain. See Requirement 1 for more details. To achieve a data-dependent switching regret for unconstrained OCO, we require a stronger regret for the base algorithm.

Requirement 2 *Consider the online convex optimization problem consisting a convex feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$ and a sequence of convex loss functions $f_1, \ldots, f_T$, where $f_t : \mathcal{X} \mapsto \mathbb{R}$ and we assume $\mathbf{0} \in \mathcal{X}$ and $\|\nabla f_t(v)\|_2 \leq 1$ for any $v \in \mathcal{X}$ and $t \in [T]$. An online algorithm $\mathcal{A}$ running over this problem returns the decision sequence $v_1, \ldots, v_T \in \mathcal{X}$. We require the algorithm $\mathcal{A}$ to ensure the following regret guarantee*

$$
\sum_{t \in \mathcal{I}} f_t(v_t) - \min_{u \in \mathcal{X}} \sum_{t \in \mathcal{I}} f_t(u) \leq \widetilde{\mathcal{O}}\left(D \sqrt{\sum_{t \in \mathcal{I}} \|\nabla f_t(v_t)\|_2^2} \right) \tag{53}
$$

for any interval $\mathcal{I} \subseteq [T]$, where $D = \sup_{x \in \mathcal{X}} \|x\|_2$ is the diameter of the feasible domain.

This requirement can be satisfied by recent OCO algorithm with *data-dependent interval regret* guarantee, such as Algorithm 2 of Zhang et al. (2019) and the algorithm specified by Theorem 6 of Cutkosky (2020).

Using the new base algorithm and the same meta algorithm as Appendix D.2, the overall algorithm can ensure a data-dependent comparator-adaptive switching regret.

Theorem 25 *Algorithm 4 with a base algorithm satisfying Requirement 2 guarantees that for any S, any partition $\mathcal{I}_1, \ldots, \mathcal{I}_S$ of $[T]$, and any comparator sequence $\mathring{u}_1, \ldots, \mathring{u}_S \in \mathbb{R}^d$, we have*

$$
\begin{aligned}
\sum_{k=1}^{S} \left(\sum_{t \in \mathcal{I}_k} f_t(v_t) - \sum_{t \in \mathcal{I}_k} f_t(\mathring{u}_k) \right) &\leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{\sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2} \right) \\
&\leq \widetilde{\mathcal{O}}\left(\max_{k \in [S]} \|\mathring{u}_k\|_2 \cdot \sqrt{S \sum_{t=1}^{T} \|\nabla f_t(v_t)\|_2^2} \right).
\end{aligned} \tag{54}
$$

Notably, the algorithm does not require the prior knowledge of the number of switch S as the input.

Proof The argument follows the proof of Appendix D.3. Similar to Eq. (43), the regret within the interval can be decomposed into meta-regret and base-regret:

$$\sum_{t\in\mathcal{I}_k}\Big(f_t(v_t)-f_t(\mathring{u}_k)\Big)\leq\underbrace{\sum_{t\in\mathcal{I}_k}\langle w_t-e_j,\ell_t\rangle}_{\textsc{Meta-Regret}}+\underbrace{\sum_{t\in\mathcal{I}_k}\Big(f_t(v_{t,j})-f_t(\mathring{u}_k)\Big)}_{\textsc{Base-Regret}},\tag{55}$$

which holds for any index $j=(i,r)\in[H]\times[R]$.

We first the case when $\|\mathring{u}_k\|_2\geq\frac{1}{T}$ and will deal with the other case (when $\|\mathring{u}_k\|_2<\frac{1}{T}$) at the end of the proof. Under such a circumstance, we can choose $(i,r)=(i_k^*,r_k^*)$ such that

$$\begin{aligned}
&c_{i_k^*}=T^{-1}\cdot 2^{i_k^*-1}\leq\|\mathring{u}_k\|_2\leq T^{-1}\cdot 2^{i_k^*}=c_{i_k^*+1},\text{ and}\\
&\eta_{r_k^*}=\frac{1}{32\cdot 2^{r_k^*}}\leq\frac{1}{32\sqrt{\sum_{t\in\mathcal{I}_k}\|\nabla f_t(v_t)\|_2^2}}\leq\frac{1}{32\cdot 2^{r_k^*-1}}=\eta_{r_k^*-1},
\end{aligned}\tag{56}$$

which is valid as $i\in[H]=[\lceil\log_2 T\rceil+T+1]$ and $r\in[R]=[\lceil\log_2 T\rceil]$. We now give the upper bounds for META-REGRET and BASE-REGRET respectively.

BASE-REGRET. Based on the assumption of base algorithm, we have base learner $\mathcal{B}_{j_k^*}$ satisfies

$$\sum_{t\in\mathcal{I}_k}\Big(f_t(v_{t,j_k^*})-f_t(\mathring{u}_k)\Big)\leq\widetilde{\mathcal{O}}\left(2^{i_k^*}\sqrt{\sum_{t\in\mathcal{I}_k}\|\nabla f_t(v_t)\|_2^2}\right)\leq\widetilde{\mathcal{O}}\left(\|\mathring{u}_k\|_2\sqrt{\sum_{t\in\mathcal{I}_k}\|\nabla f_t(v_t)\|_2^2}\right),\tag{57}$$

where we use the interval regret guarantee of base algorithm (see Requirement 2) and also use the fact that the diameter of the feasible domain for base learner $\mathcal{B}_{j_k^*}$ is $2^{i_k^*}$ as $\mathcal{X}_{i_k^*}=\{x\mid\|x\|_2\leq D_{i_k^*}\}$ and $D_{i_k^*}=c_{i_k^*}$. The last inequality holds by the choice of i_k^* shown in Eq. (56).

META-REGRET. Note that the meta algorithm remains the same, so we will only improve the analysis to show that the meta algorithm can also enjoy a data-dependent guarantee. The bias term will not be affected, which is the same as the data-independent one presented in Eq. (47), and the main modification will be conducted on the stability term. Indeed, continuing the analysis of the stability term exhibited in Eq. (48), we have

$$\begin{aligned}
&32\sum_{t\in\mathcal{I}_k}\sum_{i\in[H]}\sum_{r\in[R]}\frac{\eta_r}{c_i}\bar{e}_{j_k^*,(i,r)}\ell_{t,(i,r)}^2\\
&\leq 32\sum_{t\in\mathcal{I}_k}\frac{\eta_{r_k^*}}{c_{i_k^*}}\ell_{t,(i_k^*,r_k^*)}^2+\mathcal{O}\left(\frac{1}{T^2}\right)\\
&\leq\mathcal{O}\left(\eta_{r_k^*}c_{i_k^*}\sum_{t\in\mathcal{I}_k}\|\nabla f_t(v_t)\|_2^2\right)+\mathcal{O}\left(\frac{1}{T^2}\right)
\end{aligned}\tag{58}$$

where the last inequality holds as $\ell_{t,(i,r)}^2=\langle\nabla f_t(v_t),v_{t,(i,r)}\rangle^2\leq\|\nabla f_t(v_t)\|_2^2\|v_{t,(i,r)}\|_2^2\leq c_i^2\|\nabla f_t(v_t)\|_2^2$. Then, combining the upper bounds of bias term Eq. (47), above stability term Eq. (58), and additional term Eq. (49) leads to the following result:

$$\sum_{t\in\mathcal{I}_k}\left\langle w_t-e_{j_k^*},\ell_t\right\rangle=\sum_{t\in\mathcal{I}_k}\left\langle w_t-\bar{e}_{j_k^*},\ell_t\right\rangle+\sum_{t\in\mathcal{I}_k}\left\langle\bar{e}_{j_k^*}-e_{j_k^*},\ell_t\right\rangle$$

$$\leq \widetilde{\mathcal{O}}\left(\eta_{r_k^*} c_{i_k^*} \sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2 + \frac{c_{i_k^*}}{\eta_{r_k^*}} \log c_{i_k^*}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big)$$

$$= \widetilde{\mathcal{O}}\left(\|\mathring{u}_k\|_2 \sqrt{\sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right) + \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big), \tag{59}$$

where the last inequality is because of the choice of i_k^* and r_k^* defined in Eq. (56). Summing over all the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_S$ achieves a data-dependent upper bound for the meta-regret:

$$\begin{aligned}
&\sum_{k=1}^{S} \sum_{t \in \mathcal{I}_k} \left\langle w_t - e_{j_k^*}, \ell_t \right\rangle \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{\sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2}\right) + \sum_{k=1}^{S} \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r}\Big(w_{s_k,(i,r)} - w_{s_{k+1},(i,r)}\Big) \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{\sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2}\right) + \sum_{i \in [H]} \sum_{r \in [R]} \frac{c_i}{\eta_r} w_{1,(i,r)} \\
&= \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S} \|\mathring{u}_k\|_2 \sqrt{\sum_{t \in \mathcal{I}_k} \|\nabla f_t(v_t)\|_2^2}\right) \\
&\leq \widetilde{\mathcal{O}}\left(\max_{k \in [S]} \|\mathring{u}_k\|_2 \cdot \sqrt{S \sum_{t=1}^{T} \|\nabla f_t(v_t)\|_2^2}\right).
\end{aligned} \tag{60}$$

The last equality holds by the same argument for Eq. (51) and the final inequality is by Cauchy-Schwarz inequality. Combining the meta-regret and base-regret upper bounds finishes the proof for the case when $\|\mathring{u}_k\|_2 \geq \frac{1}{T}$ holds for every $k \in [S]$.

In addition, when the above condition of the comparators' norm is violated, we can deal with the scenario by the same argument at the end of Appendix D.3 and attain the same regret guarantee. Hence, we finish the proof of the overall theorem. ■

Remark 26 *Note that Theorem 25 is for the unconstrained OCO setting, while from the proof we can see that actually the result holds even if the algorithm is required to make decisions from a bounded domain. Indeed, in the unconstrained setting, we only need to focus on a bounded domain with maximum diameter 2^T as observed in (Chen et al., 2021, Appendix D.5). As a result, when working under constrained OCO with a diameter $D_{\max} > 0$, we can still use our algorithm by simply maintaining the set of base algorithm instances as*

$$\mathcal{S}' = \left\{\mathcal{B}_{i,r}, \forall (i,r) \in [H'] \times [R] \;\middle|\; \mathcal{B}_{i,r} \leftarrow \mathfrak{B}(\mathcal{X}_i), \text{ with } \mathcal{X}_i = \{x \mid \|x\|_2 \leq D_i = T^{-1} \cdot 2^{i-1}\}\right\}.$$

where $H' = \lceil \log_2 T \rceil + \lceil \log_2 D_{\max} \rceil + 1$ and $R = \lceil \log_2 T \rceil$ now. Thus, our result strictly improves the $\widetilde{\mathcal{O}}\Big(D_{\max}\sqrt{S \sum_{t=1}^{T} \|\nabla f_t(v_t)\|_2^2}\Big)$ result of prior works (Cutkosky, 2020; Zhao et al., 2020, 2021b) for the constrained OCO setting.

D.5. Proof of Theorem 4

Proof From Lemma 2, we have

$$\textsc{Reg}(u_1,\ldots,u_T) = \sum_{k=1}^{S} \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{V}}(\|\mathring{u}_k\|_2) + \sum_{k=1}^{S} \|\mathring{u}_k\|_2 \cdot \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{Z}}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right). \tag{61}$$

In the following, we bound the two terms respectively.

The first term on the right-hand side of Eq. (61) is the switching regret of the OCO algorithm $\mathcal{A}_{\mathcal{V}}$, we have

$$\sum_{k=1}^{S} \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{V}}(\|\mathring{u}_k\|_2) = \sum_{k=1}^{S}\sum_{t\in\mathcal{I}_k}\Big(f_t(v_t) - f_t(\|\mathring{u}_k\|_2)\Big) \le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S}\|\mathring{u}_k\|_2\sqrt{|\mathcal{I}_k|}\right),$$

where the first equality is due to the definition of online function $f_t(v) = v\cdot\langle \ell_t, z_t\rangle$ and the second inequality holds by the regret guarantee of $\mathcal{A}_{\mathcal{V}}$ proven in Theorem 3.

The second term on the right-hand side of Eq. (61) requires the switching regret analysis of the online algorithm for constrained linear bandits $\mathcal{A}_{\mathcal{Z}}$. Indeed, since the comparator satisfies that $\|\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\|_2 = 1$, the subroutine $\mathcal{A}_{\mathcal{Z}}$ can be chosen as the proposed algorithm for linear bandits with ℓ_p-ball feasible domain (with $p = 2$), see Algorithm 2. We thus get the following regret bound according to Theorem 11:

$$\mathbb{E}\left[\textsc{Reg}_{\mathcal{I}_k}^{\mathcal{Z}}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right)\right] \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{dT}{S}} + \sqrt{\frac{Sd}{T}}|\mathcal{I}_k|\right).$$

Substituting the above two upper bounds in Eq. (61) gives that

$$\begin{aligned}
\mathbb{E}\left[\textsc{Reg}(u_1,\ldots,u_T)\right] &= \sum_{k=1}^{S}\mathbb{E}\left[\textsc{Reg}_{\mathcal{I}_k}^{\mathcal{V}}(\|\mathring{u}_k\|_2)\right] + \sum_{k=1}^{S}\mathbb{E}\left[\|\mathring{u}_k\|_2\cdot \textsc{Reg}_{\mathcal{I}_k}^{\mathcal{Z}}\left(\frac{\mathring{u}_k}{\|\mathring{u}_k\|_2}\right)\right]\\
&\le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S}\|\mathring{u}_k\|_2\sqrt{|\mathcal{I}_k|}\right) + \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S}\|\mathring{u}_k\|_2\Big(\sqrt{\frac{dT}{S}} + \sqrt{\frac{Sd}{T}}|\mathcal{I}_k|\Big)\right)\\
&\le \widetilde{\mathcal{O}}\left(\sum_{k=1}^{S}\|\mathring{u}_k\|_2\Big(\sqrt{\frac{dT}{S}} + \sqrt{\frac{Sd}{T}}|\mathcal{I}_k|\Big)\right)\\
&\le \widetilde{\mathcal{O}}\left(\max_{k\in[S]}\|\mathring{u}_k\|_2\cdot\sqrt{dST}\right)
\end{aligned}$$

where the second inequality is because $\sqrt{|\mathcal{I}_k|} \le \sqrt{\frac{dT}{S}} + \sqrt{\frac{Sd}{T}}|\mathcal{I}_k|$. Hence, we finish the proof. ■

Appendix E. Lemmas Related to Online Mirror Descent

This section collects several useful lemmas related to online mirror descent (OMD).

We first introduce a general regret guarantee for OMD due to Bubeck and Cesa-Bianchi (2012).

Lemma 27 (Theorem 5.5 of Bubeck and Cesa-Bianchi (2012)) *Let $\mathcal{D} \subset \mathbb{R}^d$ be an open convex set and let $\overline{\mathcal{D}}$ be the closure of $\mathcal{D}$. Let $\mathcal{X}$ be a compact and convex set and let F be a Legendre function defined on $\overline{\mathcal{D}} \supset \mathcal{X}$ such that $\nabla F(x) - \varepsilon \nabla \ell(x) \in \mathcal{D}^*$ holds for any $(x, \ell) \in (\mathcal{X} \cap \mathcal{D}) \times \mathcal{L}$, where $\mathcal{D}^* = \nabla F(\mathcal{D})$ is the dual space of $\mathcal{D}$ under F. Consider the following online mirror descent:*

$$\begin{aligned} x'_{t+1} &= \nabla F^* \left(\nabla F(x_t) - \varepsilon \nabla \ell_t(x_t)\right), \\ x_{t+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} D_F(x, x'_{t+1}), \end{aligned} \tag{62}$$

where F^* is the Legendre–Fenchel transform of F defined by $F^*(u) = \sup_{x \in \mathcal{X}}(x^\top u - F(x))$. Then, we have

$$\sum_{t=1}^T \ell_t(x_t) - \sum_{t=1}^T \ell_t(x) \le \frac{F(x) - F(x_1)}{\varepsilon} + \frac{1}{\varepsilon} \sum_{t=1}^T D_{F^*}\Big(\nabla F(x_t) - \varepsilon \nabla \ell_t(x_t), \nabla F(x_t)\Big). \tag{63}$$

We next introduce an important lemma related to the online mirror descent with weighted entropy regularizer, which is a version of (Chen et al., 2021, Lemma 1) in the fixed learning rate and non-optimistic setting. Note that this is actually an interval version of (Chen et al., 2021, Lemma 1), replacing the summation range from $[T]$ to an interval $\mathcal{I} \subseteq [T]$, which is also used in (Chen et al., 2021, Appendix C.3).

Lemma 28 (Lemma 1 of Chen et al. (2021)) *Consider the following online mirror descent update over a compact convex decision subset $\Omega \subseteq \Delta_d$,*

$$w_{t+1} = \operatorname*{argmin}_{w \in \Omega} \Big\{ \langle w, \ell_t + a_t \rangle + D_\psi(w, w_t) \Big\}$$

where $\psi(w) = \sum_{n=1}^d \frac{1}{\eta_n} w_n \log w_n$ is the weighted entropy regularizer. Suppose that for all $t \in [T]$, $32\eta_n |\ell_{t,n}| \le 1$ holds for all $n \in [d]$ such that $w_{t,n} > 0$. Then the above update ensures for any $u \in \Omega$,

$$\sum_{t \in \mathcal{I}} \langle \ell_t, w_t - u \rangle \le \sum_{t \in \mathcal{I}_k} \Big(D_\psi(u, w_t) - D_\psi(u, w_{t+1})\Big) + 32 \sum_{t \in \mathcal{I}} \sum_{n=1}^d \eta_n u_n \ell_{t,n}^2 - 16 \sum_{t \in \mathcal{I}} \sum_{n=1}^d \eta_n w_{t,n} \ell_{t,n}^2.$$