Adversarial Bandits Against Arbitrary Strategies

Anonymous authors
Paper under double-blind review

Abstract

We study the adversarial bandit problem against arbitrary strategies, where the difficulty is captured by an unknown parameter S, which is the number of switches in the best arm in hindsight. To handle this problem, we adopt the master-base framework using the online mirror descent method (OMD). We first provide a master-base algorithm with simple OMD, achieving $\tilde{O}(S^{1/2}K^{1/3}T^{2/3})$, in which $T^{2/3}$ comes from the variance of loss estimators. To mitigate the impact of the variance, we propose using adaptive learning rates for OMD and achieve $\tilde{O}(\min\{\sqrt{SKT\rho}, S\sqrt{KT}\})$, where ρ is a variance term for loss estimators.

1 Introduction

The bandit problem is a fundamental framework in sequential decision-making that addresses the exploration-exploitation trade-off. At each time step, an agent selects an action (referred to as an arm) and observes a corresponding reward or loss. In applications such as recommendation systems, arms may represent items presented to users, and user preferences can evolve over time. This dynamic nature can be modeled by allowing the identity of the best arm to change, leading to the notion of switching best arms.

To capture such evolving environments, it is natural to consider performance against a changing sequence of optimal arms rather than a single fixed one. The problem of competing against switching arms has been extensively studied. In the expert setting with full-information feedback (Cesa-Bianchi et al., 1997), several algorithms (Daniely et al., 2015; Jun et al., 2017) have been developed that achieve a near-optimal regret bound of $\tilde{\mathcal{O}}(\sqrt{ST})$ for S-switch regret (formally defined later), without requiring prior knowledge of the number of switches S. In contrast, the bandit setting presents a greater challenge, as the agent only observes the feedback for the selected arm rather than the full loss vector, making the problem significantly harder than in the full-information case.

In the stochastic bandit setting where each arm's reward distribution may change over time, commonly referred to as the non-stationary bandit problem, several studies have addressed the challenge of switching environments (Garivier & Moulines, 2008; Auer et al., 2019; Russac et al., 2019; Suk & Kpotufe, 2022). Notably, Auer et al. (2019) and Suk & Kpotufe (2022) achieved near-optimal regret bounds of $\tilde{O}(\sqrt{SKT})$ without requiring prior knowledge of the number of switches S. However, these methods rely on stochastic assumptions and are not applicable in the adversarial setting, where losses can be chosen arbitrarily. For the adversarial bandit setting, as summarized in Table 1, the `EXP3.S` algorithm (Auer et al., 2002) achieves a regret bound of $\tilde{O}(\sqrt{SKT})$, but assumes that S is known in advance. When S is unknown, the Bandit-over-Bandit (`BOB`) approach has been proposed, achieving a regret bound of $\tilde{O}(\sqrt{SKT} + T^{3/4})$ (Cheung et al., 2019; Foster et al., 2020). More recently, Luo et al. (2022) studied the problem of switching adversarial linear bandits and achieved a regret of $\tilde{O}(\sqrt{dST})$ under the assumption that S is known.

In this paper, we study adversarial bandit problems against arbitrarily switching arms. Crucially, we allow the number of switches S to be unknown to the agent, thereby targeting arbitrary strategies without prior knowledge of their complexity. To address this setting, we adopt the master-base framework combined with the online mirror descent (OMD) method, inspired by Agarwal et al. (2017); Pacchiano et al. (2020); Luo et al. (2022). We begin by analyzing a master-base algorithm that employs OMD with a negative entropy regularizer, and show that it achieves a regret bound of $\tilde{O}(S^{1/2}K^{1/3}T^{2/3})$. However, this approach relies on a fixed learning rate, which limits its ability to adapt to the variance of the loss estimators, leading to a suboptimal regret term proportional to $T^{2/3}$.

Algorithm	**Regret Bound**	**S Known?**
`EXP3.S` (Auer et al., 2002)	$\tilde{O}(\sqrt{SKT})$	Yes
`BOB` (Cheung et al., 2019)	$\tilde{O}(\sqrt{SKT} + T^{3/4})$	No
Algorithm 1 (Ours)	$\tilde{O}(S^{1/2}K^{1/3}T^{2/3})$	No
Algorithm 2 (Ours)	$\tilde{O}(\min\{\sqrt{SKT\rho},\, S\sqrt{KT}\})$	No

Table 1: Comparison of regret bounds for adversarial bandit algorithms with switches S.

Building on this analysis, we propose using adaptive learning rates within the OMD framework to better control the variance of loss estimators. This refinement leads to an improved regret bound of $\tilde{O}(\min\{\sqrt{SKT\rho}, S\sqrt{KT}\})$ with respect to T, where ρ captures the variance associated with a comparator strategy. Crucially, rather than employing the standard negative entropy regularizer, we adopt a log-barrier regularizer, which enables tighter control over worst-case scenarios in terms of the variance term ρ.

2 Problem Statement

We now formalize the problem setting. Let $\mathcal{A} = [K]$ denote the set of K arms, and let $\boldsymbol{l}_t \in [0,1]^K$ be the loss vector at time t, where $l_t(a)$ denotes the loss incurred by arm $a \in \mathcal{A}$ at time t. The environment is adversarial, generating an arbitrary sequence of loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [0,1]^K$ over a time horizon of T rounds. At each round $t \in [T]$, the agent selects an arm $a_t \in \mathcal{A}$ and observes only the loss $l_t(a_t)$ of the chosen arm. Our goal is to minimize the S-switch regret, which measures the performance gap between the agent and the best sequence of actions that switches arms at most S times.

Formally, let $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_T\} \in \mathcal{A}^T$ denote a sequence of comparator actions. For a positive integer $S < T$, define the set of sequences with at most S switches as

$$B_S = \left\{ \sigma \in [K]^T : \sum_{t=1}^{T-1} \mathbb{1}\{\sigma_t \neq \sigma_{t+1}\} \leq S \right\}.$$

Then, the S-switch regret is defined by

$$R_S(T) = \max_{\sigma \in B_S} \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{l}_t(a_t)] - \boldsymbol{l}_t(\sigma_t).$$

We consider the setting where the number of switches S is unknown to the agent (i.e., not provided in advance). Our goal is to design algorithms that perform well against any sequence of actions, without relying on prior knowledge of S. This requires the development of universal algorithms that achieve tight regret bounds uniformly over all values of $S \in [T-1]$, where S characterizes the hardness of the problem (Auer et al., 2002). Notably, this setting generalizes non-stationary stochastic bandit problems in which the switching parameter is unknown (Auer et al., 2019; Chen et al., 2019).

3 Algorithms and Regret Analysis

To address the problem of adversarial bandits with an unknown switching budget S, we propose algorithms based on the master-base framework combined with the online mirror descent (OMD) method.

3.1 Master-Base Framework

In the master-base framework, a master algorithm selects one among several base algorithms at each round, and the selected base then chooses an arm to play. Since the true switch parameter S is not known in advance, we instantiate each base algorithm with a different candidate value of S from a predefined set.

Let $\mathcal{H}$ denote the set of candidate values for S, defined as:

$$\mathcal{H} = \{T^0, T^{\frac{1}{\lceil \log T \rceil}}, T^{\frac{2}{\lceil \log T \rceil}}, \dots, T\}.$$

Each base is associated with a candidate $h \in \mathcal{H}$ and tunes its learning rate accordingly. Let $H = |\mathcal{H}| = O(\log T)$ be the number of base algorithms. When there is no ambiguity, we refer to a base instantiated with parameter $h \in \mathcal{H}$ simply as base h. Define $h^\dagger \in \mathcal{H}$ as the largest candidate not exceeding the true (unknown) value of S: $h^\dagger = \max\{h \in \mathcal{H} : h \le S\}$. By construction, this ensures

$$e^{-1}S \le h^\dagger \le S,$$

which guarantees that $h^\dagger$ provides a near-optimal approximation of S.

3.2 Online Mirror Descent (OMD)

We now present the *Online Mirror Descent* (OMD) method (Lattimore & Szepesvári, 2020), which serves as the fundamental update rule for the master and each base algorithm in our framework. OMD generalizes classic online gradient descent by incorporating a flexible geometry through a regularizer.

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable regularizer function. This regularizer defines the *Bregman divergence* between two points $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d$:

$$D_F(\boldsymbol{p}, \boldsymbol{q}) = F(\boldsymbol{p}) - F(\boldsymbol{q}) - \langle \nabla F(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle.$$

In our context, the decision variable is a probability distribution $\boldsymbol{p}_t$ over d arms, i.e., $\boldsymbol{p}_t \in \mathcal{P}_d$ where $\mathcal{P}_d$ denotes the d-dimensional probability simplex. At each round t, given a loss vector $\boldsymbol{l} \in \mathbb{R}^d$, OMD computes the next distribution by solving the following optimization problem:

$$\boldsymbol{p}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathcal{P}_d} \Big\{ \langle \boldsymbol{p}, \boldsymbol{l} \rangle + D_F(\boldsymbol{p}, \boldsymbol{p}_t) \Big\}.$$

This formulation reflects the trade-off between exploiting current loss information (via the linear term) and staying close to the previous distribution (via the Bregman divergence).

In practice, the update equation 3.2 is often implemented in two steps for computational convenience:

$$\tilde{\boldsymbol{p}}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathbb{R}^d} \Big\{ \langle \boldsymbol{p}, \boldsymbol{l} \rangle + D_F(\boldsymbol{p}, \boldsymbol{p}_t) \Big\},$$
$$\boldsymbol{p}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathcal{P}_d} D_F(\boldsymbol{p}, \tilde{\boldsymbol{p}}_{t+1}).$$

Here, the first step performs an unconstrained update in the dual space defined by F, while the second step projects the intermediate point back onto the simplex, ensuring that the updated distribution is a valid probability vector. The choice of regularizer F determines the specific algorithm instance; it typically includes a learning rate parameter that controls the step size and will be specified in subsequent sections.

Finally, in the bandit setting, since the learner does not observe the full loss vector $\boldsymbol{l}_t$ but only the loss of the selected arm, the true loss must be replaced with an appropriate unbiased estimator. This ensures that OMD remains applicable even with partial feedback.

3.3 Master-Base OMD

We employ an OMD framework with a hierarchical structure consisting of a master and multiple bases. We first present a simple Master-Base OMD algorithm (Algorithm 1) that employs the negative entropy regularizer:

$$F_\eta(\boldsymbol{p}) = (1/\eta) \sum_{i=1}^{d} (p(i) \log p(i) - p(i)),$$

where $\boldsymbol{p} \in \mathbb{R}^d$, $p(i)$ denotes the i-index entry for $\boldsymbol{p}$, and η is a learning rate. This regularizer is commonly used in adversarial bandit algorithms, including the well-known `EXP3` algorithm (Auer et al., 2002).

In Algorithm 1, at each round t, the master algorithm selects a base $h_t \in \mathcal{H}$ according to a probability distribution $\boldsymbol{p}_t$. The selected base h_t then chooses an arm $a_t \in [K]$ using its internal distribution $\boldsymbol{p}_{t,h_t}$ and observes the incurred loss $l_t(a_t)$. From this partial feedback, we construct unbiased estimators: $l'_t(h)$ estimates the loss for each base $h \in \mathcal{H}$, and $l''_{t,h}(a)$ estimates the loss for each arm $a \in [K]$ under base h. Using these estimators, the algorithm updates both the master distribution $\boldsymbol{p}_{t+1}$ and each base's distribution $\boldsymbol{p}_{t+1,h}$ via OMD.

To control variance in the estimator $l'_t(h) = l_t(a_t)\mathbb{1}(h = h_t)/p_t(h)$, we define the master's update domain as a clipped probability simplex $\mathcal{P}_H^\alpha = \mathcal{P}_H \cap [\alpha, 1]^H$ for some $\alpha > 0$. This clipping ensures that each base is selected with non-negligible probability, preventing high variance in the importance-weighted loss estimates. Within this domain, the update of $\boldsymbol{p}_{t+1}$ is then computed using the negative entropy regularizer with learning rate η.

Each base $h \in \mathcal{H}$ maintains its own arm selection distribution $\boldsymbol{p}_{t,h}$. The update for this distribution is also based on the negative entropy regularizer but with a learning rate that adapts to the candidate switch parameter h:

$$\xi(h) = \frac{h^{1/2}}{K^{1/3}T^{2/3}}.$$

This choice of $\xi(h)$ helps base h adapt to environments with up to h switches in the best arm sequence. The domain for updating $\boldsymbol{p}_{t,h}$ is $\mathcal{P}_K^\beta = \mathcal{P}_K \cap [\beta, 1]^K$ for $\beta > 0$. Unlike α, which controls variance in master-level estimation, β acts as a regularization mechanism to stabilize learning under the switching best arms in hindsight.

Algorithm 1 Master-base OMD

```
Input: $T$, $K$, $\mathcal{H}$.
Initialization: $\alpha = K^{1/3}/(T^{1/3}H^{1/2})$, $\beta = 1/(KT)$, $\eta = 1/\sqrt{TK}$, $\xi(h) = h^{1/2}/(K^{1/3}T^{2/3})$, $p_1(h) = 1/H$, $p_{1,h}(a) = 1/K$ for $h \in \mathcal{H}$ and $a \in [K]$.
for $t = 1, \dots, T$ do
    Select a base and an arm:
    Draw $h_t \sim$ probabilities $p_t(h)$ for $h \in \mathcal{H}$.
    Draw $a_{t,h_t} \sim$ probabilities $p_{t,h_t}(a)$ for $a \in [K]$.
    Pull $a_t = a_{t,h_t}$ and Receive $l_t(a_{t,h_t}) \in [0, 1]$.
    Obtain loss estimators:
    $l'_t(h_t) = \frac{l_t(a_{t,h_t})}{p_t(h_t)}$ and $l'_t(h) = 0$ for $h \in \mathcal{H}/\{h_t\}$.
    $l''_{t,h_t}(a_{t,h_t}) = \frac{l'_t(h_t)}{p_{t,h_t}(a_{t,h_t})}$ and $l''_{t,h}(a) = 0$ for $h \in \mathcal{H}/\{h_t\}$, $a \in [K]/\{a_{t,h_t}\}$.
    Update distributions:
    $\boldsymbol{p}_{t+1} = \arg\min_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \langle \boldsymbol{p}, \boldsymbol{l}'_t \rangle + D_{F_\eta}(\boldsymbol{p}, \boldsymbol{p}_t)$
    $\boldsymbol{p}_{t+1,h} = \arg\min_{\boldsymbol{p} \in \mathcal{P}_K^\beta} \langle \boldsymbol{p}, \boldsymbol{l}''_{t,h} \rangle + D_{F_{\xi(h)}}(\boldsymbol{p}, \boldsymbol{p}_{t,h})$ for all $h \in \mathcal{H}$
end for
```

Now we provide a regret bound for the algorithm in the following theorem.

Theorem 3.1. *For any switch number $S \in [T-1]$, Algorithm 1 achieves a regret bound of*

$$R_S(T) = \tilde{O}(S^{1/2}K^{1/3}T^{2/3})$$

Proof Sketch. Here, we provide a proof sketch, and the full version is provided in Appendix A.1.

In our proof, we decompose the regret into two parts: one is the regret from the master selecting a base at each time, and the other is the regret from the base selecting an arm at each time. Let t_s be the time when the s-th switch of the best arm happens and $t_{S+1} - 1 = T$, $t_0 = 1$. Also let $t_{s+1} - t_s = T_s$. For any t_s for all $s \in [0, S]$, the S-switch regret can be expressed as

$$\begin{aligned} R_S(T) &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_t)\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \\ &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] + \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s), \end{aligned} \tag{1}$$

in which the first two terms are closely related with the regret from the master algorithm against the near optimal base $h^\dagger$, and the remaining terms are related with the regret from $h^\dagger$ base algorithm against the best arms in hindsight. We note that the algorithm does not need to know $h^\dagger$ in prior and $h^\dagger$ is brought here only for regret analysis.

Regret from the near-optimal base. First we provide a bound for the following regret from base $h^\dagger$:

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s),$$

where the first term is the loss from the base and the second one is the loss from the optimal arm in hindsight. Let $k_s^* = \arg\min_{k\in[K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k)$ and $\boldsymbol{e}_{j,K}$ denote the unit vector with 1 at j-index and 0 at the rest of $K-1$ indexes. Then, we have

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[l_t(a_{t,h^\dagger}) - l_t(k_s^*)\right] \leq \beta T_s(K-1) + \mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_K^\beta} \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle\right], \tag{2}$$

where the first term in the last inequality is obtained from the clipped domain $\mathcal{P}_K^\beta$ and the second term is obtained from the unbiased estimator $\boldsymbol{l}''_{t,h^\dagger}$ such that $\mathbb{E}[\boldsymbol{l}''_{t,h^\dagger}|\mathcal{F}_{t-1}] = \mathbb{E}[\boldsymbol{l}_t|\mathcal{F}_{t-1}]$ where $\mathcal{F}_{t-1}$ is the filtration. We can observe that the clipped domain controls the distance between the initial distribution at t_s and the best arm unit vector for the time steps over $[t_s, t_{s+1}-1]$. Let

$$\tilde{\boldsymbol{p}}_{t+1,h^\dagger} = \arg\min_{\boldsymbol{p}\in\mathbb{R}^K} \langle \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle + D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}, \boldsymbol{p}_{t,h^\dagger}).$$

Then, by solving the optimization problem, for all $k \in [K]$ we can obtain

$$\tilde{p}_{t+1,h^\dagger}(k) = p_{t,h^\dagger}(k) \exp(-\xi(h^\dagger) l''_{t,h^\dagger}(k)).$$

For the second term of the last inequality in equation 2, we have for any $\boldsymbol{p} \in \mathcal{P}_K^\beta$,

$$\sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle \leq D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) + \sum_{t=t_s}^{t_{s+1}-1} D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger}). \tag{3}$$

The first term is for the initial point diameter at time t_s and the second term is for the divergence of the updated policy. Using the definition of the Bregman divergence and the fact that $p_{t_s,h^\dagger}(k) \geq \beta$, the initial point diameter term can be shown to be bounded as follows:

$$D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) \leq \frac{\log(1/\beta)}{\xi(h^\dagger)}. \tag{4}$$

Next, for the updated policy divergence term, using $\tilde{p}_{t+1,h^\dagger}(k) = p_{t,h^\dagger}(k)\exp(-\xi(h^\dagger) l''_{t,h^\dagger}(k))$ for all $k \in [K]$, we have

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})\right] \leq \frac{\xi(h^\dagger) K T_s}{2\alpha}. \tag{5}$$

Then from equation 2, equation 3, equation 4, and equation 5, by summing up over $s \in [S]$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \le \beta T(K-1) + \frac{S \log(1/\beta)}{\xi(h^\dagger)} + \frac{\xi(h^\dagger) K T}{2\alpha}. \tag{6}$$

Next, we provide a bound for the following regret from the master:

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right].$$

Let $\tilde{\boldsymbol{p}}_{t+1} = \arg\min_{\boldsymbol{p} \in \mathbb{R}^H} \langle \boldsymbol{p}, \boldsymbol{l}'_t \rangle + D_{F_\eta}(\boldsymbol{p}, \boldsymbol{p}_t)$ and $\boldsymbol{e}_{h,H}$ denote the unit vector with 1 at base h-index and 0 at the rest of $H-1$ indexes. For ease of presentation, we define $\tilde{l}_t(h) = l_t(a_{t,h})$. Then, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t}) - l_t(a_{t,h^\dagger})\right] \le \alpha T(H-1) + \mathbb{E}\left[\max_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t \rangle\right]. \tag{7}$$

Regret from the master. For bounding the second term in equation 7, which arises from the master, we use the following:

$$\max_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t \rangle \le \max_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \left(F_\eta(\boldsymbol{p}) - F_\eta(\boldsymbol{p}_1) + \sum_{t=1}^{T} D_{F_\eta}(\boldsymbol{p}_t, \tilde{\boldsymbol{p}}_{t+1}) \right). \tag{8}$$

From equation 7 and equation 8, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] \\
&\le \alpha T(H-1) + \mathbb{E}\left[\max_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \left(F_\eta(\boldsymbol{p}) - F_\eta(\boldsymbol{p}_1) + \sum_{t=1}^{T} D_{F_\eta}(\boldsymbol{p}_t, \tilde{\boldsymbol{p}}_{t+1}) \right)\right] \\
&\le \alpha T(H-1) + \frac{\log(H)}{\eta} + \frac{\eta T K}{2}.
\end{aligned} \tag{9}$$

Overall Regret. Therefore, putting equation 1, equation 6, and equation 9 altogether, we have

$$\begin{aligned}
R_S(T) &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_t)\right] - \sum_{s=0}^{S} \min_{1 \le k_s \le K} \sum_{t=T_s}^{T_{s+1}-1} l_t(k_s) \\
&\le \alpha T H + \frac{\log(H)}{\eta} + \frac{\eta T K}{2} + \beta T(K-1) + \frac{S \log(1/\beta)}{\xi(h^\dagger)} + \frac{\xi(h^\dagger) K T}{2\alpha} \\
&= \tilde{O}(S^{1/2} T^{2/3} K^{1/3}),
\end{aligned}$$

where $\alpha = K^{1/3}/(T^{1/3} H^{1/2})$, $\beta = 1/(KT)$, $\eta = 1/\sqrt{TK}$, $\xi(h^\dagger) = {h^\dagger}^{1/2}/(K^{1/3} T^{2/3})$, $h^\dagger = \Theta(S)$, and $H = \log(T)$. This concludes the proof. □

Compared to the prior parameter-free algorithm based on the Bandit-over-Bandit (`BOB`) approach (Cheung et al., 2019), which incurs a suboptimal dependence on the time horizon T, specifically, a regret term of order $T^{3/4}$, our algorithm achieves a tighter regret bound in terms of T. In particular, when $T = \omega(S^6 K^4)$, Algorithm 1 achieves a tighter regret bound than that of `BOB`.

However, the regret bound achieved by Algorithm 1 remains of order $O(T^{2/3})$, rather than the optimal $O(\sqrt{T})$, due to the high variance in the loss estimators. This variance arises from the double sampling process—first selecting a base, then an arm—at each round. To address this issue, we next propose an improved algorithm that leverages *adaptive* learning rates to better control the variance of the estimators.

3.4 Master-Base OMD with Adaptive Learning Rates

We now propose Algorithm 2, which incorporates adaptive learning rates to better control the variance of the loss estimators. We begin by describing the base algorithm. Each base employs the negative entropy regularizer, but with a time-varying adaptive learning rate $\xi_t(h)$, defined as:

$$F_{\xi_t(h)}(\boldsymbol{p}) = \frac{1}{\xi_t(h)} \sum_{i=1}^{d} (p(i) \log p(i) - p(i)),$$

where $\boldsymbol{p} \in \mathbb{R}^d$ is a probability distribution over arms. The learning rate $\xi_t(h)$ is dynamically adjusted at each round t based on the variance of the loss estimators, and is given by:

$$\xi_t(h) = \sqrt{h/(KT\rho_t(h))},$$

where $\rho_t(h)$ is a variance threshold term that will be defined later. This formulation ensures that when the variance of the estimators is small, a larger learning rate is used—allowing for more aggressive updates—while high variance naturally leads to more conservative updates. Notably, this adaptive base algorithm is effectively combined with the master employing log-barrier regularization to control the regret due to variance, resulting in equation 14—an integration that, to the best of our knowledge, has not been explored before.

For the master algorithm, inspired by the approach in Agarwal et al. (2017), we employ a log-barrier regularizer with increasing learning rates. This design introduces a negative bias term that effectively offsets the variance arising from the base algorithms, particularly by addressing the worst-case scenario in terms of the variance threshold $\rho_t(h^\dagger)$. The log-barrier regularizer is defined as:

$$F_{\boldsymbol{\eta}_t}(\boldsymbol{p}) = -\sum_{i=1}^{d} \frac{\log p(i)}{\eta_t(i)},$$

where $\boldsymbol{p} \in \mathcal{P}_d$ is the master distribution, and $\boldsymbol{\eta}_t = (\eta_t(1), \ldots, \eta_t(d))$ denotes the vector of learning rates at time t. We describe the learning rate update procedures for both the master and the base algorithms in Algorithm 2; all other components remain identical to those in Algorithm 1. The variance of the loss estimator $l'_{t+1}(h) = l_t(a_{t+1,h})\mathbf{1}(h_{t+1} = h)/p_{t+1}(h)$ for base h is given by $1/p_{t+1}(h)$. If this variance exceeds the threshold $\rho_t(h)$, i.e., $1/p_{t+1}(h) > \rho_t(h)$, the master increases the learning rate as: $\eta_{t+1}(h) = \gamma\eta_t(h)$ for some fixed $\gamma > 1$. Simultaneously, the variance threshold is updated to: $\rho_{t+1}(h) = 2/p_{t+1}(h)$ which is also used to adaptively tune the base learning rate $\xi_t(h)$ as described earlier. If the variance does not exceed the threshold, both $\eta_t(h)$ and $\rho_t(h)$ remain unchanged from the previous time step.

In the following theorem, we provide a regret bound of Algorithm 2.

Theorem 3.2. *For any switch number $S \in [T-1]$ and any $\rho \geq \rho_T(h^\dagger)$, Algorithm 2 achives a regret bound of*

$$R_S(T) = \tilde{O}\left(\min\left\{\sqrt{SKT\rho}, S\sqrt{KT}\right\}\right).$$

Proof Sketch. Here, we provide a proof sketch, and the full version is provided in Appendix A.2. As in the Theorem 3.1, we decompose the regret into two parts: one is the regret from the master selecting a base at each time, and the other is the regret from the base selecting an arm at each time. Let t_s be the time when the s-th switch of the best arm happens and $t_{S+1} - 1 = T$, $t_0 = 1$. Also let $t_{s+1} - t_s = T_s$. For any t_s for all $s \in [0, S]$, the S-switch regret can be expressed as

$$\begin{aligned} R_S(T) &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_t)\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \\ &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] + \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s). \end{aligned} \tag{10}$$

Algorithm 2 Master-base OMD with adaptive learning rates

```
Input: $T$, $K$, $\mathcal{H}$
Initialization: $\alpha = 1/(TH)$, $\beta = 1/(TK)$, $\gamma = e^{\frac{1}{\log T}}$, $\eta = \sqrt{H/T}$, $\rho_1(h) = 2H$, $\eta_1(h) = \eta$, $p_1(h) = 1/H$, $p_{1,h}(a) = 1/K$ for $h \in \mathcal{H}$ and $a \in [K]$.
for $t = 1, \dots, T$ do
    Select a base and an arm:
    Draw $h_t \sim$ probabilities $p_t(h)$ for $h \in \mathcal{H}$.
    Draw $a_{t,h_t} \sim$ probabilities $p_{t,h_t}(a)$ for $a \in [K]$.
    Pull $a_t = a_{t,h_t}$ and Receive $l_t(a_{t,h_t}) \in [0,1]$.
    Update loss estimators:
    $l'_t(h_t) = \frac{l_t(a_{t,h_t})}{p_t(h_t)}$ and $l'_t(h) = 0$ for $h \in \mathcal{H}/\{h_t\}$.
    $l''_{t,h_t}(a_{t,h_t}) = \frac{l'_t(h_t)}{p_{t,h_t}(a_{t,h_t})}$ and $l''_{t,h}(a) = 0$ for $h \in \mathcal{H}/\{h_t\}$, $a \in [K]/\{a_{t,h_t}\}$.
    Update distributions:
    $\boldsymbol{p}_{t+1} = \arg\min_{\boldsymbol{p} \in \mathcal{P}_H^\alpha} \langle \boldsymbol{p}, \boldsymbol{l}'_t \rangle + D_{F_{\boldsymbol{\eta}_t}}(\boldsymbol{p}, \boldsymbol{p}_t)$
    $\boldsymbol{p}_{t+1,h} = \arg\min_{\boldsymbol{p} \in \mathcal{P}_K^\beta} \langle \boldsymbol{p}, \boldsymbol{l}''_{t,h} \rangle + D_{F_{\xi_t(h)}}(\boldsymbol{p}, \boldsymbol{p}_{t,h})$ for $h \in \mathcal{H}$
    Update learning rates:
    For $h \in \mathcal{H}$
        If $\frac{1}{p_{t+1}(h)} > \rho_t(h)$, then
            $\rho_{t+1}(h) = \frac{2}{p_{t+1}(h)}, \eta_{t+1}(h) = \gamma \eta_t(h)$.
        Else, $\rho_{t+1}(h) = \rho_t(h), \eta_{t+1}(h) = \eta_t(h)$.
end for
```

Regret from the near-optimal base. First we provide a bound for the following regret from base $h^\dagger$. From equation 2, we can obtain

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \leq \beta T_s K + \mathbb{E}\left[\max_{p \in \mathcal{P}_K^\beta} \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle\right]. \tag{11}$$

Then for the second term of the last inequality in equation 11, we provide the following lemma.

Lemma 3.3. *For any $\boldsymbol{p} \in \mathcal{P}_K^\beta$ we can show that*

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[\langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle\right] \leq \mathbb{E}\left[2\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \frac{T_s}{2}\sqrt{\frac{SK\rho_T(h^\dagger)}{T}}\right].$$

Then from equation 7 and Lemma 3.3, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{1 \leq k_s \leq K} \sum_{t=T_s}^{T_{s+1}-1} l_t(k_s) \leq \beta T(K-1) + \mathbb{E}\left[2S\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \frac{1}{2}\sqrt{TSK\rho_T(h^\dagger)}\right]. \tag{12}$$

Regret from the master. Next, we provide a bound for the regret from the master in the following:

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] \leq O\left(\frac{H\log(T)}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho_T(h^\dagger)}{40\eta\log T}\right] + \alpha T(H-1). \tag{13}$$

The negative bias term in equation 13 is derived from the log-barrier regularizer and increasing learning rates $\eta_t(h)$. This term is critical to bound the worst case regret which will be shown soon. Also, $H\log(T)/\eta$ is obtained from $H\log(1/(H\alpha))/\eta$ considering the clipped domain.

Overall Regret. Then, putting equation 10, equation 12 and equation 13 altogether, we have

$$R_S(T) = \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_t)\right] - \sum_{s=0}^{S} \min_{1\le k_s\le K} \sum_{t=T_s}^{T_{s+1}-1} l_t(k_s) = \tilde{O}\left(\mathbb{E}\left[\sqrt{SKT\rho_T(h^\dagger)}\right]\right) - \mathbb{E}\left[\frac{\rho_T(h^\dagger)\sqrt{TK}}{40\sqrt{H}\log(T)}\right], \tag{14}$$

Then, we can obtain

$$R_S(T) = \tilde{O}\left(\min\left\{\sqrt{SKT\rho}, S\sqrt{KT}\right\}\right),$$

where $\tilde{O}(S\sqrt{KT})$ is obtained from the worst case of $\rho_T(h^\dagger)$. The worst case can be found by considering a maximum value of the concave bound of the last equality in equation 14 with variable $\rho_T(h^\dagger) > 0$ such that $\rho_T(h^\dagger) = \tilde{\Theta}(S)$. This concludes the proof. □

We now provide a comparison of regret bounds with an existing approach. our regret bound depends on ρ, which captures the variance of the loss estimators $l'_t(h^\dagger)$ over the time horizon $t \in [T]$. While the bound is variance-dependent, it is noteworthy that in the worst case, it is always upper bounded by $\tilde{O}(S\sqrt{KT})$. Importantly, Algorithm 2 achieves a tight dependence of $O(\sqrt{T})$ in the regret bound, in contrast to Algorithm 1 and `BOB` (Cheung et al., 2019; Foster et al., 2020), which incur suboptimal $T^{2/3}$ and $T^{3/4}$ term, respectively. Therefore, when T is sufficiently large, Algorithm 2 yields a strictly better regret guarantee than Algorithm 1 and `BOB`. We also note that the variance term ρ is commonly observed in Luo et al. (2022), but we control the worst case without information of S using an adaptive learning rate.

Remark 3.4. *Since we incorporate $\rho_t(h)$ into the adaptive learning rate for each base h as $\xi_t(h) = \sqrt{h/KT\rho_t(h)}$, we can optimize the regret bound to depend on $\sqrt{\rho_T(h^\dagger)}$, as demonstrated in Lemma 3.3. Notably, this adaptive base algorithm is effectively combined with the master employing log-barrier regularization Agarwal et al. (2017) to control the regret due to variance, resulting $\sqrt{S\rho_T(h^\dagger)} - \rho_T(h^\dagger)$. This integration is the main reason why Algorithm 2 can achieve a order of $\sqrt{T}$, even in the worst case, without the knowledge of S.*

Remark 3.5. *Our algorithms are designed to perform well across different regimes, particularly with respect to T and S. To recall the regret bounds, Algorithm 1 achieves a regret of $\tilde{O}(S^{1/2}K^{1/3}T^{2/3})$, while Algorithm 2 achieves $\tilde{O}(\min\{\sqrt{SKT\rho}, S\sqrt{KT}\})$. This indicates that Algorithm 1 is advantageous when S is large, whereas Algorithm 2 is preferable for larger T due to its use of an adaptive learning rate that accounts for the variance of the loss estimator.*

Remark 3.6. *For implementation of our algorithms, we describe how to update policy $\boldsymbol{p}_t$ using OMD in general. Let $\widehat{l}_t(a)$ be a loss estimator for action $a \in [d]$. For the negative entropy regularizer, by solving the optimization in equation 3.2, from Lattimore & Szepesvári (2020), we have*

$$p_{t+1}(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t} \widehat{l}_s(a)\right)}{\sum_{b\in[d]} \exp\left(-\eta \sum_{s=1}^{t} \widehat{l}_s(b)\right)}.$$

In the case of the log-barrier regularizer, we have $p_{t+1}(a) = (\eta \sum_{s=1}^{t} \widehat{l}_s(a) + Z)^{-1}$, where Z is a normalization factor for a probability distribution Luo et al. (2022). Also, a clipped domain with $0 < \epsilon < 1$ in the distribution can be implemented by adding a uniform probability to the policy such that $p_{t+1}(a) \leftarrow (1-\epsilon)p_{t+1}(a) + \epsilon/d$ for all $a \in [d]$.

Remark 3.7. *The regret bounds in Theorems 3.1 and 3.2 also extend to non-stationary stochastic bandit problems with unknown switching parameters, where the reward distributions may change over time. This generalization is possible because the adversarial bandit setting encompasses the stochastic setting as a special case.*

4 Conclusion

In this paper, we studied the adversarial bandit problem with S-switch regret, where the agent competes against any sequence of arms that switches at most S times, without prior knowledge of S. To address this challenge, we proposed two algorithms based on the master-base framework integrated with the Online Mirror Descent (OMD) method.

First, we introduced Algorithm 1, which employs a simple OMD update with a fixed learning rate and achieves a regret bound of $\tilde{O}(S^{1/2}K^{1/3}T^{2/3})$. To further improve performance with respect to T, we proposed Algorithm 2, which incorporates adaptive learning rates to control the variance of the loss estimators. This leads to an improved regret bound of $\tilde{O}(\min\{\sqrt{SKT\rho}, S\sqrt{KT}\})$ where ρ captures the variance of the estimators associated with the near-optimal base.

References

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

Peter Auer, Nicolo 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 *Conference on Learning Theory*, pp. 138–158, 2019.

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087, 2019.

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

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Open problem: Model selection for contextual bandits. In *Conference on Learning Theory*, pp. 3842–3846. PMLR, 2020.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems, 2008.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *Artificial Intelligence and Statistics*, pp. 943–951. PMLR, 2017.

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

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. *arXiv preprint arXiv:2202.06151*, 2022.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv preprint arXiv:2003.01704*, 2020.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pp. 12017–12026, 2019.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Conference on Learning Theory*, pp. 2160–2182. PMLR, 2022.

A Appendix

A.1 Proof of Theorem 3.1

Let t_s be the time when the s-th switch of the best arm happens and $t_{S+1}-1=T$, $t_0=1$. Also let $t_{s+1}-t_s=T_s$. For any t_s for all $s\in[0,S]$, the S-switch regret can be expressed as

$$\begin{aligned}R_S(T)&=\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_t)\right]-\sum_{s=0}^{S}\min_{k_s\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k_s)\\&=\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h_t})\right]-\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]+\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]-\sum_{s=0}^{S}\min_{k_s\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k_s),\end{aligned}\tag{15}$$

in which the first two terms are closely related with the regret from the master algorithm against the near optimal base $h^\dagger$, and the remaining terms are related with the regret from $h^\dagger$ base algorithm against the best arms in hindsight. We note that the algorithm does not need to know $h^\dagger$ in prior and $h^\dagger$ is brought here only for regret analysis.

First we provide a bound for the following regret from base $h^\dagger$:

$$\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]-\sum_{s=0}^{S}\min_{k_s\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k_s).$$

Let $k_s^*=\arg\min_{k\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k)$ and $\boldsymbol{e}_{j,K}$ denote the unit vector with 1 at j-index and 0 at the rest of $K-1$ indexes. Then, we have

$$\begin{aligned}&\sum_{t=t_s}^{t_{s+1}-1}\mathbb{E}\left[l_t(a_{t,h^\dagger})-l_t(k_s^*)\right]\\&=\sum_{t=t_s}^{t_{s+1}-1}\mathbb{E}\left[\langle\boldsymbol{p}_{t,h^\dagger}-\boldsymbol{e}_{k_s^*,K},\boldsymbol{l}_t\rangle\right]\\&\le\mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_K^\beta}\sum_{t=t_s}^{t_{s+1}-1}\langle\boldsymbol{p}-\boldsymbol{e}_{k_s^*,K},\boldsymbol{l}_t\rangle+\max_{\boldsymbol{p}\in\mathcal{P}_K^\beta}\sum_{t=t_s}^{t_{s+1}-1}\langle\boldsymbol{p}_{t,h^\dagger}-\boldsymbol{p},\boldsymbol{l}_t\rangle\right]\\&\le\beta T_s(K-1)+\mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_K^\beta}\sum_{t=t_s}^{t_{s+1}-1}\langle\boldsymbol{p}_{t,h^\dagger}-\boldsymbol{p},\boldsymbol{l}''_{t,h^\dagger}\rangle\right],\end{aligned}\tag{16}$$

where the first term in the last inequality is obtained from the clipped domain $\mathcal{P}_K^\beta$ and the second term is obtained from the unbiased estimator $\boldsymbol{l}''_{t,h^\dagger}$ such that $\mathbb{E}[\boldsymbol{l}''_{t,h^\dagger}|\mathcal{F}_{t-1}]=\mathbb{E}[\boldsymbol{l}_t|\mathcal{F}_{t-1}]$ where $\mathcal{F}_{t-1}$ is the filtration. We can observe that the clipped domain controls the distance between the initial distribution at t_s and the best arm unit vector for the time steps over $[t_s,t_{s+1}-1]$. Let

$$\tilde{\boldsymbol{p}}_{t+1,h^\dagger}=\arg\min_{\boldsymbol{p}\in\mathbb{R}^K}\langle\boldsymbol{p},\boldsymbol{l}''_{t,h^\dagger}\rangle+D_{F_{\xi(h^\dagger)}}(\boldsymbol{p},\boldsymbol{p}_{t,h^\dagger}).$$

Then, by solving the optimization problem, for all $k\in[K]$ we can obtain

$$\tilde{p}_{t+1,h^\dagger}(k)=p_{t,h^\dagger}(k)\exp(-\xi(h^\dagger)l''_{t,h^\dagger}(k)).$$

For the second term of the last inequality in equation 16, we provide a lemma in the following.

Lemma A.1 (Theorem 28.4 in Lattimore & Szepesvári (2020)). *For any $\boldsymbol{p}\in\mathcal{P}_K^\beta$ we have*

$$\sum_{t=t_s}^{t_{s+1}-1}\langle\boldsymbol{p}_{t,h^\dagger}-\boldsymbol{p},\boldsymbol{l}''_{t,h^\dagger}\rangle\le D_{F_{\xi(h^\dagger)}}(\boldsymbol{p},\boldsymbol{p}_{t_s,h^\dagger})+\sum_{t=t_s}^{t_{s+1}-1}D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}_{t,h^\dagger},\tilde{\boldsymbol{p}}_{t+1,h^\dagger}).$$

In Lemma A.1, the first term is for the initial point diameter at time t_s and the second term is for the divergence of the updated policy. Using the definition of the Bregman divergence and the fact that $p_{t_s,h^\dagger}(k) \geq \beta$, the initial point diameter term can be shown to be bounded as follows:

$$\begin{aligned} D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) &\leq \frac{1}{\xi(h^\dagger)} \sum_{k\in[K]} p(k) \log(1/p_{t_s,h^\dagger}(k)) \\ &\leq \frac{\log(1/\beta)}{\xi(h^\dagger)}. \end{aligned} \tag{17}$$

Next, for the updated policy divergence term, using $\tilde{p}_{t+1,h^\dagger}(k) = p_{t,h^\dagger}(k)\exp(-\xi(h^\dagger)l''_{t,h^\dagger}(k))$ for all $k \in [K]$, we have

$$\begin{aligned} &\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[D_{F_{\xi(h^\dagger)}}(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})\right] \\ &= \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{1}{\xi(h^\dagger)} p_{t,h^\dagger}(k) \left(\exp(-\xi(h^\dagger)l''_{t,h^\dagger}(k)) - 1 + \xi(h^\dagger)l''_{t,h^\dagger}(k)\right)\right] \\ &\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{\xi(h^\dagger)}{2} p_{t,h^\dagger}(k) l''_{t,h^\dagger}(k)^2\right] \\ &\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{\xi(h^\dagger)}{2p_t(h^\dagger)}\right] \leq \frac{\xi(h^\dagger)KT_s}{2\alpha}, \end{aligned} \tag{18}$$

where the first inequality comes from $\exp(-x) \leq 1 - x + x^2/2$ for all $x \geq 0$, the second inequality comes from $\mathbb{E}[l''_{t,h^\dagger}(k)^2 \mid p_{t,h^\dagger}(k), p_t(h^\dagger)] \leq 1/(p_t(h^\dagger)p_{t,h^\dagger}(k))$, and the last inequality is obtained from $p_t(h^\dagger) \geq \alpha$ from the clipped domain. We can observe that the clipped domain controls the variance of estimators. Then from equation 16, Lemma A.1, equation 17, and equation 18, by summing up over $s \in [S]$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s\in[K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \leq \beta T(K-1) + \frac{S\log(1/\beta)}{\xi(h^\dagger)} + \frac{\xi(h^\dagger)KT}{2\alpha}. \tag{19}$$

Next, we provide a bound for the following regret from the master:

$$\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right].$$

Let $\tilde{\boldsymbol{p}}_{t+1} = \arg\min_{\boldsymbol{p}\in\mathbb{R}^H} \langle \boldsymbol{p}, \boldsymbol{l}'_t\rangle + D_{F_\eta}(\boldsymbol{p}, \boldsymbol{p}_t)$ and $\boldsymbol{e}_{h,H}$ denote the unit vector with 1 at base h-index and 0 at the rest of $H-1$ indexes. For ease of presentation, we define $\tilde{l}_t(h_t) = l_t(a_{t,h_t})$ and $\tilde{l}_t(h^\dagger) = l_t(a_{t,h^\dagger})$. Then, we have

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t}) - l_t(a_{t,h^\dagger})\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{p}_t - \boldsymbol{e}_{h^\dagger,H}, \tilde{\boldsymbol{l}}_t\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p} - \boldsymbol{e}_{h^\dagger,H}, \tilde{\boldsymbol{l}}_t\rangle + \max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle\right] \\ &\leq \alpha T(H-1) + \mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle\right]. \end{aligned} \tag{20}$$

For bounding the second term in equation 20, we use the following lemma.

Lemma A.2 (Theorem 28.4 in Lattimore & Szepesvári (2020))**.**

$$\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha}\sum_{t=1}^{T}\langle\boldsymbol{p}_t-\boldsymbol{p},\tilde{\boldsymbol{l}}_t\rangle\le\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha}\left(F_\eta(\boldsymbol{p})-F_\eta(\boldsymbol{p}_1)+\sum_{t=1}^{T}D_{F_\eta}(\boldsymbol{p}_t,\tilde{\boldsymbol{p}}_{t+1})\right).$$

From equation 20 and Lemma A.2, we have

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h_t})\right]-\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]\\&\le\alpha T(H-1)+\mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha}\left(F_\eta(\boldsymbol{p})-F_\eta(\boldsymbol{p}_1)+\sum_{t=1}^{T}D_{F_\eta}(\boldsymbol{p}_t,\tilde{\boldsymbol{p}}_{t+1})\right)\right]\\&\le\alpha T(H-1)+\frac{\log(H)}{\eta}+\frac{\eta TK}{2},\end{aligned}\tag{21}$$

where the last inequality is obtained from the fact that

$$F_\eta(\boldsymbol{p})-F_\eta(\boldsymbol{p}_1)\le-F_\eta(\boldsymbol{p}_1)\le\frac{\log(H)}{\eta}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}D_{F_\eta}(\boldsymbol{p}_t,\tilde{\boldsymbol{p}}_{t+1})\right]\le\frac{\eta TK}{2}.$$

Therefore, putting equation 15, equation 19, and equation 21 altogether, we have

$$\begin{aligned}R_S(T)&=\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_t)\right]-\sum_{s=0}^{S}\min_{1\le k_s\le K}\sum_{t=T_s}^{T_{s+1}-1}l_t(k_s)\\&\le\alpha TH+\frac{\log(H)}{\eta}+\frac{\eta TK}{2}+\beta T(K-1)+\frac{S\log(1/\beta)}{\xi(h^\dagger)}+\frac{\xi(h^\dagger)KT}{2\alpha}\\&=\tilde{O}(S^{1/2}T^{2/3}K^{1/3}),\end{aligned}$$

where $\alpha=K^{1/3}/(T^{1/3}H^{1/2})$, $\beta=1/(KT)$, $\eta=1/\sqrt{TK}$, $\xi(h^\dagger)={h^\dagger}^{1/2}/(K^{1/3}T^{2/3})$, $h^\dagger=\Theta(S)$, and $H=\log(T)$. This concludes the proof.

A.2 Proof of Theorem 3.2

Let t_s be the time when the s-th switch of the best arm happens and $t_{S+1}-1=T$, $t_0=1$. Also let $t_{s+1}-t_s=T_s$. For any t_s for all $s\in[0,S]$, the S-switch regret can be expressed as

$$\begin{aligned}R_S(T)&=\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_t)\right]-\sum_{s=0}^{S}\min_{k_s\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k_s)\\&=\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h_t})\right]-\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]+\sum_{t=1}^{T}\mathbb{E}\left[l_t(a_{t,h^\dagger})\right]-\sum_{s=0}^{S}\min_{k_s\in[K]}\sum_{t=t_s}^{t_{s+1}-1}l_t(k_s),\end{aligned}\tag{22}$$

in which the first two terms are closely related with the regret from the master algorithm against the near optimal base $h^\dagger$, and the remaining terms are related with the regret from $h^\dagger$ base algorithm against the best arms in hindsight.

First we provide a bound for the following regret from base $h^\dagger$. From equation 16, we can obtain

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{k_s \in [K]} \sum_{t=t_s}^{t_{s+1}-1} l_t(k_s) \leq \beta T_s K + \mathbb{E}\left[\max_{p \in \mathcal{P}_K^\beta} \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle \right]. \tag{23}$$

Then for the second term of the last inequality in equation 23, we provide a following lemma.

Lemma A.3 (Restatement of Lemma 3.3)**.** *For any $\boldsymbol{p} \in \mathcal{P}_K^\beta$ we can show that*

$$\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[\langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_{t,h^\dagger} \rangle\right] \leq \mathbb{E}\left[2\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \frac{T_s}{2}\sqrt{\frac{SK\rho_T(h^\dagger)}{T}}\right].$$

Proof. For ease of presentation, we define the negative entropy regularizer without a learning rate as

$$F(\boldsymbol{p}) = \sum_{i=1}^{K} (p(i)\log p(i) - p(i))$$

and define learning rate $\xi_0(h^\dagger) = \infty$. From the first-order optimality condition for $p_{t+1,h^\dagger}$ and using the definition of the Bregman divergence,

$$\begin{aligned} \langle \boldsymbol{p}_{t+1,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_t \rangle &\leq \frac{1}{\xi_t(h^\dagger)} \langle \boldsymbol{p} - \boldsymbol{p}_{t+1,h^\dagger}, \nabla F(\boldsymbol{p}_{t+1,h^\dagger}) - \nabla F(\boldsymbol{p}_{t,h^\dagger}) \rangle \\ &= \frac{1}{\xi_t(h^\dagger)} \left(D_F(\boldsymbol{p}, \boldsymbol{p}_{t,h^\dagger}) - D_F(\boldsymbol{p}, \boldsymbol{p}_{t+1,h^\dagger}) - D_F(\boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{p}_{t,h^\dagger}) \right). \end{aligned} \tag{24}$$

Also, we have

$$\begin{aligned} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{l}''_t \rangle &= \frac{1}{\xi_t(h^\dagger)} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}_{t+1,h^\dagger}, \nabla F(\boldsymbol{p}_{t,h^\dagger}) - \nabla F(\tilde{\boldsymbol{p}}_{t+1,h^\dagger}) \rangle \\ &= \frac{1}{\xi_t(h^\dagger)} (D_F(\boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{p}_{t,h^\dagger}) + D_F(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger}) - D_F(\boldsymbol{p}_{t+1,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})) \\ &\leq \frac{1}{\xi_t(h^\dagger)} (D_F(\boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{p}_{t,h^\dagger}) + D_F(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})). \end{aligned} \tag{25}$$

Then, we can obtain

$$\begin{aligned} \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}, \boldsymbol{l}''_t \rangle &\leq \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}_{t+1,h^\dagger}, l''_t \rangle \\ &\quad + \sum_{t=t_s}^{t_{s+1}-1} \frac{1}{\xi_t(h^\dagger)} \left(D_F(\boldsymbol{p}, \boldsymbol{p}_{t,h^\dagger}) - D_F(\boldsymbol{p}, \boldsymbol{p}_{t+1,h^\dagger}) - D_F(\boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{p}_{t,h^\dagger}) \right) \\ &= \sum_{t=t_s}^{t_{s+1}-1} \langle \boldsymbol{p}_{t,h^\dagger} - \boldsymbol{p}_{t+1,h^\dagger}, l''_t \rangle + \sum_{t=t_s+1}^{t_{s+1}-1} D_F(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) \left(\frac{1}{\xi_t(h^\dagger)} - \frac{1}{\xi_{t-1}(h^\dagger)} \right) \\ &\quad \frac{1}{\xi_{t_s}(h^\dagger)} D_F(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) - \frac{1}{\xi_{t_{s+1}-1}(h)} D_F(\boldsymbol{p}, \boldsymbol{p}_{t_{s+1},h^\dagger}) - \frac{1}{\xi_t(h^\dagger)} \sum_{t=t_s}^{t_{s+1}-1} D_F(\boldsymbol{p}_{t+1,h^\dagger}, \boldsymbol{p}_{t,h^\dagger}) \\ &\leq 2\frac{\log(1/\beta)}{\xi_t(h^\dagger)} + \sum_{t=t_s}^{t_{s+1}-1} \frac{D_F(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})}{\xi_t(h^\dagger)} \\ &= 2\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \sum_{t=t_s}^{t_{s+1}-1} \frac{D_F(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})}{\xi_t(h^\dagger)}, \end{aligned} \tag{26}$$

where the first inequality is obtained from equation 24 and the last inequality is obtained from equation 25, $D_F(\boldsymbol{p}, \boldsymbol{p}_{t_s,h^\dagger}) \leq \log(1/\beta)$, and $\xi_t(h^\dagger) \geq \xi_t(h^\dagger)$ from non-decreasing $\rho_t(h^\dagger)$.

For the second term in the last inequality in equation 26, using $\tilde{p}_{t+1,h^\dagger}(k) = p_{t,h^\dagger}(k)\exp(-\xi(h^\dagger)l''_{t,h^\dagger}(k))$ for all $k \in [K]$, we have

$$
\begin{aligned}
\sum_{t=t_s}^{t_{s+1}-1} \mathbb{E}\left[\frac{D_F(\boldsymbol{p}_{t,h^\dagger}, \tilde{\boldsymbol{p}}_{t+1,h^\dagger})}{\xi_t(h^\dagger)}\right] &= \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\Bigg[\frac{1}{\xi_t(h^\dagger)} p_{t,h^\dagger}(k) \Big(\exp(-\xi_t(h^\dagger)l''_{t,h^\dagger}(k)) \\
&\qquad\qquad -1 + \xi_t(h^\dagger)l''_{t,h^\dagger}(k)\Big)\Bigg] \\
&\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{\xi_t(h^\dagger)}{2} p_{t,h^\dagger}(k) l''_{t,h^\dagger}(k)^2\right] \\
&\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{\xi_t(h^\dagger)}{2p_t(h^\dagger)}\right] \\
&\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{\xi_t(h^\dagger)\rho_t(h^\dagger)}{2}\right] \\
&\leq \sum_{t=t_s}^{t_{s+1}-1} \sum_{k=1}^{K} \mathbb{E}\left[\frac{1}{2}\sqrt{\frac{h^\dagger \rho_t(h^\dagger)}{KT}}\right] \\
&\leq T_s \sqrt{\frac{h^\dagger K}{T}} \frac{\mathbb{E}\left[\rho_T(h^\dagger)^{1/2}\right]}{2}, \qquad (27)
\end{aligned}
$$

where the first inequality comes from $\exp(-x) \leq 1 - x + x^2/2$ for all $x \geq 0$, the second inequality comes from $\mathbb{E}[l''_{t,h^\dagger}(k)^2 \mid p_{t,h^\dagger}(k), p_t(h^\dagger)] \leq 1/(p_t(h^\dagger)p_{t,h^\dagger}(k))$, and the third inequality is obtained from $1/p_t(h^\dagger) \leq \rho_t(h^\dagger)$.

□

Then from equation 20 and Lemma A.3, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] - \sum_{s=0}^{S} \min_{1 \leq k_s \leq K} \sum_{t=T_s}^{T_{s+1}-1} l_t(k_s) \\
&\leq \beta T(K-1) + \mathbb{E}\left[2S\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \frac{1}{2}\sqrt{TSK\rho_T(h^\dagger)}\right]. \qquad (28)
\end{aligned}
$$

Next, we provide a bound for the regret from the master in the following lemma.

Lemma A.4 (Lemma 13 in Agarwal et al. (2017))**.**

$$
\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t})\right] - \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h^\dagger})\right] \leq O\left(\frac{H\log(T)}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho_T(h^\dagger)}{40\eta\log T}\right] + \alpha T(H-1).
$$

Proof. For ease of presentation, we define $\tilde{l}_t(h_t) = l_t(a_{t,h_t})$ and $\tilde{l}_t(h^\dagger) = l_t(a_{t,h^\dagger})$. First, from the clipped probability simplex, we have

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[l_t(a_{t,h_t}) - l_t(a_{t,h^\dagger})\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\langle \boldsymbol{p}_t - \boldsymbol{e}_{h^\dagger,H}, \tilde{\boldsymbol{l}}_t\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p} - \boldsymbol{e}_{h^\dagger,H}, \tilde{\boldsymbol{l}}_t\rangle + \max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle\right] \\ &\leq \alpha T(H-1) + \mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle\right]. \qquad (29)\end{aligned}$$

Now we provide a bound for $\mathbb{E}\left[\max_{\boldsymbol{p}\in\mathcal{P}_H^\alpha} \sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle\right]$ by following the proof steps in Agarwal et al. (2017). We have

$$\begin{aligned}\sum_{t=1}^{T} \langle \boldsymbol{p}_t - \boldsymbol{p}, \tilde{\boldsymbol{l}}_t\rangle &\leq \sum_{t=1}^{T} (D_{F_{\eta_t}}(\boldsymbol{p}, \boldsymbol{p}_t) - D_{F_{\eta_t}}(\boldsymbol{p}, \boldsymbol{p}_{t+1})) + \sum_{t=1}^{T} \sum_{h\in[H]} \eta_t(h) p_t(h)^2 \tilde{l}_t(h) \\ &\leq O\left(\frac{H\log T}{\eta} + T\eta\right) - \frac{\rho_T(h^\dagger)}{40\eta \log T}. \qquad (30)\end{aligned}$$

This concludes the proof from equation 29 and equation 30. ☐

The negative bias term in Lemma A.4 is derived from the log-barrier regularizer and increasing learning rates $\eta_t(h)$. This term is critical to bound the worst case regret which will be shown soon. Also, $H\log(T)/\eta$ is obtained from $H\log(1/(H\alpha))/\eta$ considering the clipped domain. Then, putting equation 22 and Lemmas A.3 and A.4 altogether, we have

$$\begin{aligned}R_S(T) &= \sum_{t=1}^{T} \mathbb{E}\left[l_t(a_t)\right] - \sum_{s=0}^{S} \min_{1\leq k_s \leq K} \sum_{t=T_s}^{T_{s+1}-1} l_t(k_s) \\ &\leq O\left(\frac{H\log T}{\eta} + T\eta\right) - \mathbb{E}\left[\frac{\rho_T(h^\dagger)}{40\eta\log T}\right] \\ &\quad + \alpha T(H-1) + \beta T(K-1) \\ &\quad + \mathbb{E}\left[2S\log(1/\beta)\sqrt{\frac{KT\rho_T(h^\dagger)}{h^\dagger}} + \frac{1}{2}\sqrt{SKT\rho_T(h^\dagger)}\right] \\ &= \tilde{O}\left(\mathbb{E}\left[\sqrt{SKT\rho_T(h^\dagger)}\right]\right) - \mathbb{E}\left[\frac{\rho_T(h^\dagger)\sqrt{TK}}{40\sqrt{H}\log(T)}\right], \qquad (31)\end{aligned}$$

where $\alpha = 1/(TH)$, $\beta = 1/(TK)$, $\eta = \sqrt{H/T}$, $\xi(h^\dagger) = \sqrt{h^\dagger/(KT\rho_T(h))}$, $H = \log(T)$, and $h^\dagger = \Theta(S)$. Then we can obtain

$$R_S(T) = \tilde{O}\left(\min\left\{\sqrt{SKT\rho}, S\sqrt{KT}\right\}\right),$$

where $\tilde{O}(S\sqrt{KT})$ is obtained from the worst case of $\rho_T(h^\dagger)$. The worst case can be found by considering a maximum value of the concave bound of the last equality in equation 31 with variable $\rho_T(h^\dagger) > 0$ such that $\rho_T(h^\dagger) = \tilde{\Theta}(S)$. This concludes the proof.