Scale-free Adversarial Reinforcement Learning

Mingyu Chen MINGYUC@BU.EDU
Boston University

Xuezhou Zhang XUEZHOUZ@BU.EDU
Boston University

Editors: Shipra Agrawal and Aaron Roth

Abstract

This paper initiates the study of scale-free learning in Markov Decision Processes (MDPs), where the scale of rewards/losses is unknown to the learner. We design a generic algorithmic framework, Scale Clipping Bound (SCB), and instantiate this framework in both the adversarial Multi-armed Bandit (MAB) setting and the adversarial MDP setting. Through this framework, we achieve the first minimax optimal expected regret bound and the first high-probability regret bound in scale-free adversarial MABs, resolving an open problem raised in Hadiji and Stoltz (2023). On adversarial MDPs, our framework also give birth to the first scale-free RL algorithm with a $\tilde{\mathcal{O}}(\sqrt{T})$ high-probability regret guarantee.

Keywords: bandit, MDP, scale-free learning, high-probability regret.

1. Introduction

Reinforcement learning (RL) refers to the problem of an agent interacting with an unknown environment with the goal of improving its policy and minimizing cumulative loss through time. The environment is commonly modeled as a Markov Decision Process (MDP) with an unknown transition function. In this paper we focus on the adversarial MDP setting, where the losses are allowed to be generated adversarially (Even-Dar et al., 2009). Curiously, virtually *all* prior works on RL assume that the rewards/losses are uniformly bounded, e.g. the mean reward for any state-action pair is within $[0, 1]$. This regularity condition is crucial in allowing existing algorithms to set their hyperparameters such as learning rate properly to achieve low regrets. In many real-world applications, however, such natural loss bound does not always exist. For instance, in quantitative trading, stock prices can vary significantly over time and across different stocks. More importantly, the scale of such variance is often not known to the algorithm a priori. In such settings, most existing algorithms no longer work.

Motivated by the above limitations, in this paper, we initiate the study of scale-free RL algorithms in MDPs, i.e. algorithms that require no prior knowledge on the scale of the losses. Scale-free algorithm have previously been studied in the online learning literature (Freund and Schapire, 1997; De Rooij et al., 2014; Cesa-Bianchi et al., 2007; Mayo et al., 2022; Jacobsen and Cutkosky, 2023; Cutkosky, 2019). However, for decision-making under uncertainty, the only relevant studies are limited to Multi-armed Bandits (MAB) (Hadiji and Stoltz, 2023; Putta and Agrawal, 2022; Chen and Zhang, 2023; Huang et al., 2023), which can be considered as a 1-layer MDP with a single state. Existing algorithms for scale-free MAB essentially adapts algorithms designed for online learning to the bandit feedback setting. This leads to some fundamental problems. For example, due to the limitations of regularizers, no existing scale-free adversarial MAB algorithms can achieve minimax

Setting	Loss	Algorithm	Regret	Type
Adversarial MABs	Bound	Audibert et al. (2009)	$\Theta(\ell_\infty\sqrt{nT})$	Exp.
		Neu (2015)	$\Theta(\ell_\infty\sqrt{nT\log(n/\delta)})$	High prop.
	Unbound	Hadiji and Stoltz (2023)	$\Theta(\ell_\infty\sqrt{nT}\log n)$	Exp.
		Chen and Zhang (2023)	$\Theta(\ell_\infty\sqrt{nT}\log T)$	Exp.
		SCB (Theorem 1)	$\Theta(\ell_\infty\sqrt{nT})$	Exp.
		SCB-IX (Theorem 4)	$\Theta(\ell_\infty\sqrt{nT\log(n/\delta)})$	High prop.
Adversarial MDPs	Bound	Jin et al. (2019)	$\tilde{O}(\sum_{h\in[H]}\ell_{\infty,h}S\sqrt{AT})$	High prop.
	Unbound	**SCB-RL (Theorem 5)**	$\tilde{O}(\sum_{h\in[H]}\ell_{\infty,h}S^{3/2}\sqrt{AT})$	High prop.

Table 1: An overview of the proposed algorithms/results and comparisons with related works.

optimality (i.e., optimal to logarithm terms). Secondly, existing works only bound the regret incurred by the important weighted estimators. As a result, they can only bound the expected regret and cannot be generalized to high probability regret. Third, considering the existing algorithms' dependence on important weighted estimators, their results cannot be generalized to the setting of adversarial MDP with unknown transition function.

In this paper, we propose the first scale-free algorithm for adversarial MDPs with unknown transition function. We design a unified framework called Scale Clipping Bound (SCB). This framework can be applied to both MAB and MDP and significantly improves previous results across the board. Our technical contributions can be summarized below, and an overview that compare our results with those in prior works can be found in Table 1.

1. We propose SCB, a scale-free adversarial MAB algorithm that achieves **minimax optimal** expected regret bounds without the knowledge of the loss magnitude. Our result eliminates the $\log(n)$ and $\log(T)$ factors in prior works, and matches the minimax lower-bound Auer et al. (2002a) upto constant factors. This result gives a positive answer to the open problem raised in Hadiji and Stoltz (2023).

2. Based on the idea of SCB, we build SCB-IX, the first scale-free adversarial MAB algorithm that achieves a **high probability** regret bound.

3. Finally, we extend the above ideas to the setting of adversarial MDPs and present SCB-RL, the **first** scale-free algorithm that achieve $\tilde{O}(\sqrt{T})$ high probability regret bound for adversarial MDP with unknown transition function, unbounded losses and bandit feedback.

2. Related Works

Scale-free learning: Scale-free algorithms refer to the algorithms that do not need to know any upper or lower bounds on the loss functions. Scale-free regret bounds were first studied in the full information setting, such as experts problems (Freund and Schapire, 1997; De Rooij et al., 2014; Cesa-Bianchi et al., 2007) and online convex optimization (Mayo et al., 2022; Jacobsen and Cutkosky, 2023; Cutkosky, 2019). For experts problems, the AdaHedge algorithm from De Rooij et al. (2014) achieves the first scale-free regret bound. For online convex optimization, past algorithms can be categorized into two generic algorithmic frameworks: Mirror Descent (MD) and Follow The Regularizer Leader (FTRL). The scale-free regret from the MD family is achieved by `AdaGrad` proposed by Duchi et al. (2011). However, the regret bound of Duchi et al. (2011) is only non-trivial when the Bregman divergence associated with the regularizer can be well bounded. Later, the Orabona and Pál (2018) proposed the `AdaFTRL` algorithm which achieves the first scale-free regret bound in the FTRL family and generalizes Duchi et al. (2011)'s results to cases where the Bregman divergence associated with the regularizer is unbounded. For the adversarial MAB problem, Hadiji and Stoltz (2023) extends the method of Duchi et al. (2011) and provides a scale-free regret bound of $\widetilde{O}\left(\ell_\infty\sqrt{nT}\right)$, which is optimal (up to log terms) in the worst case. Putta and Agrawal (2022) design a bandit FTRL algorithm and presents scale-free bounds that adapt to the individual size of losses across time. Unfortunately, the worst-case guarantee of Putta and Agrawal (2022) is $\widetilde{O}\left(\ell_\infty n\sqrt{T}\right)$, which scales linearly to the number of actions. To close the gap, Chen and Zhang (2023) proposes algorithms that achieves an adaptive regret better than Putta and Agrawal (2022), as well as an optimal worst-case regret that matches with Hadiji and Stoltz (2023).

Notice that all previous studies are unable to attain logarithmic optimality, e.g., the regret bound of Putta and Agrawal (2022); Chen and Zhang (2023); Huang et al. (2023) is $\Theta(\ell_\infty\sqrt{nT\log T})$, and the regret bound of Hadiji and Stoltz (2023) is $\Theta(\ell_\infty\sqrt{nT\log n})$. This is due to some inherent limitations of their algorithms. To be more specific, Hadiji and Stoltz (2023) is an extension of `AdaHedge` De Rooij et al. (2014), with a structure similar to `EXP3`, leading to an additional $\sqrt{\log n}$ regret. On the other hand, the analysis of Putta and Agrawal (2022); Chen and Zhang (2023); Huang et al. (2023) is only applicable to algorithms with a log-barrier regularizer, which also results in an additional $\sqrt{\log T}$ regret. To achieve logarithmic optimality, a promising approach would be to use algorithms with Tsallis-INF regularizer (Audibert et al., 2009), which can achieve $\Theta(\ell_\infty\sqrt{nT})$ regret bound when ℓ_∞ is known. However, when ℓ_∞ is unknown, it is unclear whether this bound can be achieved. This has been posed as an open problem in Hadiji and Stoltz (2023) (Remark 8), which we answer in the positive.

High probability regrets: High-probability regrets for adversarial MAB were first provided by Auer et al. (2002b) and explored in a more generic way by Abernethy and Rakhlin (2009). The idea is to reduce the variance of importance weighted estimators by adding *explicit exploration* on the action distribution. Later, Kocák et al. (2014) and Neu (2015) improve the *explicit exploration* method to *implicit exploration*, and design algorithms for more complex models with potentially large action sets and side information. Notably, all the above algorithms require carefully constructing *biased* loss estimators. In contrast, Lee et al. (2020) develops algorithms based on *unbiased* loss estimators, and enjoy *data-dependent* high probability regret bounds, which could be much smaller than the bounds in the form of $\tilde{O}(\sqrt{T})$ when the data is "good". For the adversarial MDP problem, a rencent line of works develop algorithms with high-probability regret bounds (Jin et al., 2019; Lee et al., 2020; Luo et al., 2021; Dai et al., 2022; Jin et al., 2022). Most of them are based on the idea of reducing an adversarial MDP problem to an adversarial MAB problem through *occupancy measure* and then solve it using bandit algorithms, and achieve the same regret guarantee. To the best of our knowledge, there are no studies on the high probability regret for either adversarial MAB or MDP with considering unbounded losses, a gap that we fill in this work.

Variance dependent regrets: Variance dependent regrets have been studied in both adversarial MAB (Hazan and Kale, 2011; Bubeck et al., 2018; Wei and Luo, 2018; Ito, 2021) and MDP (Talebi and Maillard, 2018; Simchowitz and Jamieson, 2019; Zhang et al., 2023; Zanette and Brunskill, 2019). At first glance, it seems that a variance-dependent regret can automatically adapt to the scale of the losses, thereby directly implying scale-adaptive regret. However, in fact, there is a fundamental gap between the concept of scale-free and scale-adaptive. Firstly, scale-adaptive algorithms require a *strict* assumption that the scale of losses can be bounded by a known constant L. This applies to all the above-mentioned work. If the assumption is violated, their analyses will not hold. Secondly, the above results on variance-dependent regret all include a burn-in term that scales polynomial to L. This leads to the optimality of these results being guaranteed only in a "large-sample" regime. As L goes towards infinity, the burn-in term eventually dominates.

3. Adversarial Multi-armed Bandit

Let us start our discussion with adversarial MAB. The scale-free MAB problem proceeds in rounds between a player and an adversary. In each round $t = 1, \ldots, T$, the player selects one of the n available actions $k_t \in [n]$, while the adversary at the same time picks a loss vector $\ell_t \in \mathbb{R}^n$ with $\ell_{t,k}$ being the loss for action k. We assume the adversary is *adaptive*: the adversary can choose ℓ_t base on the player's previous actions in an arbitrary way. At the end of round t, the learner observes the loss of the chosen action ℓ_{t,k_t} and nothing else. We measure the scale of the losses by $\ell_\infty = \max_{t\in[T],k\in[n]} |\ell_{t,k}|$. We measure the performance of the learner in terms of its *regret*:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \ell_{t,k_t} - \min_{k\in[n]} \sum_{t=1}^{T} \ell_{t,k}.$$

3.1. Minimax Optimal Expected Regret

In this subsection we focus on bounding the *expected regret*, i.e., $\mathbb{E}[\mathcal{R}(T)]$. Compared to existing works, there are several important advantages of our approach: 1). Our algorithm archives the first minimax optimal expected regret $\Theta(\ell_\infty\sqrt{nT})$, which significantly improves upon the $\Theta(\ell_\infty\sqrt{nT \log T})$ results in Putta and Agrawal (2022); Chen and Zhang (2023); Huang et al. (2023) and $\Theta(\ell_\infty\sqrt{nT \log n})$ in Hadiji and Stoltz (2023), and matches the lower bound $\Omega(\ell_\infty\sqrt{nT})$ proposed in Auer et al. (2002b). 2). Our algorithm is *strongly scale-free* (Orabona and Pál, 2018), that is, with the same parameters, the sequence of action distributions of the algorithm does not change if the sequence of loss is multiplied by a positive constant. Such property is previously implemented only in Hadiji and Stoltz (2023).

Our design is illustrated in Algorithm 1. The algorithm follows a standard Follow-the-regularized-Leader (FTRL) framework. At the beginning of round t, the algorithm computes an action distribution $\mathbf{p}_t \in \Delta_n$ such that

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} \left(\sum_{s=1}^{t-1} \langle \hat{\ell}_s, \mathbf{p} \rangle + \frac{1}{\eta_t} \Psi(\mathbf{p}) \right), \tag{1}$$

where $\hat{\ell}_s$ is an estimator of ℓ_s and Ψ is the regularizer. Then, the algorithm derives $\mathbf{q}_t$ by mixing $\mathbf{p}_t$ with a uniform distribution, samples and plays action $k_t \sim \mathbf{q}_t$, and obtains loss ℓ_{t,k_t}. The key of our

Algorithm 1: SCB: Scale Clipping Bound

```
Input: 1/2-Tsallis Entropy $\Psi$, $\eta_1 = \infty$, $\beta_1 = n/(2n + \sqrt{n})$, $C_1 = 0$
for $t = 1, \dots, T$ do
    Compute the action distribution $\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} \left(\sum_{s=1}^{t-1}\langle \hat{\ell}_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p})\right)$
    Add extra exploration $\mathbf{q}_t = (1 - \beta_t)\mathbf{p}_t + \beta_t \frac{\mathbf{1}_n}{n}$
    Sample and play action $k_t \sim \mathbf{q}_t$. Receive loss $\ell_{t,k_t}$
    Clip received loss by $[-C_t, C_t]$: $\ell^c_{t,k} = \max(-C_t, \min(C_t, \ell_{t,k}))$
    Construct estimator $\hat{\ell}_t$ such that $\hat{\ell}_{t,k} = \frac{\ell^c_{t,k} + C_t}{q_{t,k}} \mathbb{1}\{k = k_t\}$, $\forall k \in [n]$
    If $|\ell_{t,k_t}| > C_t$, set $C_{t+1} = 2|\ell_{t,k_t}|$, otherwise $C_{t+1} = C_t$
    Update learning rate $\eta_{t+1} = \frac{1}{2C_{t+1}\sqrt{t+1}}$. Update exploration rate $\beta_{t+1} = \frac{n}{2n+\sqrt{n(t+1)}}$
end
```

design lies in the construction of the loss estimator. In round t, the algorithm holds a "scale clipping bound" (i.e., clipping threshold) C_t, which is twice the largest scale among the previously observed losses. After receiving the loss ℓ_{t,k_t}, the algorithm clips the loss within the interval $[-C_t, C_t]$, incorporates an offset C_t to make the loss non-negative, and construct the importance-weighted loss estimator, i.e., $\hat{\ell}_{t,k} = (\max(-C_t, \min(C_t, \ell_{t,k})) + C_t)\mathbb{1}\{k = k_t\}/q_{t,k}$. Specifically, notice that C_t is independent to $\mathbf{p}_t$, thereby $\hat{\ell}_{t,k}$ is an unbiased estimator of $\max(-C_t, \min(C_t, \ell_{t,k})) + C_t$. At the end of round t, the algorithm updates parameters $C_{t+1}, \eta_{t+1}, \beta_{t+1}$, and then move to the next round. The regret guarantee of Algorithm 1 can be summarized below.

Theorem 1 *Algorithm 1 achieves*

$$\mathbb{E}[\mathcal{R}(T)] \leq \Theta\left(\ell_\infty(n + \sqrt{nT})\right).$$

Remark 2 *We emphasize that* `SCB` *is* ***strongly scale-free****. When the sequence of losses is multiplied by a positive constant, the clipping threshold will also be rescaled accordingly, resulting in the distributions of actions not changing. This property is also inherited by the derived algorithms* `SCB-IX` *and* `SCB-RL`*. More details about this property are provided in Appendix A.2.*

Proof Sketch: Denoted by $\ell^c_{t,k} = \max(-C_t, \min(C_t, \ell_{t,k}))+C_t)$, we start with the following regret decomposition.

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}^*\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell^c_t, \mathbf{p}_t - \mathbf{p}^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t - \ell^c_t, \mathbf{p}_t - \mathbf{p}^*\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell^c_t + C_t\mathbf{1}_n, \mathbf{p}_t - \mathbf{p}^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t - \ell^c_t, \mathbf{p}_t - \mathbf{p}^*\rangle\right]
\end{aligned}$$

$$= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^*\rangle\right]}_{\textcircled{1}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}_t\rangle\right]}_{\textcircled{2}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t - \ell_t^c, \mathbf{p}_t - \mathbf{p}^*\rangle\right]}_{\textcircled{3}}$$

Here, $\mathbf{p}^*$ denotes the optimal comparator, which can be dependent on the algorithm's actions $k_1, \ldots, k_T$. The third equality is due to $\langle \mathbf{1}_n, \mathbf{q}_t - \mathbf{p}^*\rangle = 0$, and the last equality is because $\hat{\ell}_t$ is an unbiased estimator of $\ell_t^c + C_t\mathbf{1}_n$. Here, term $\textcircled{1}$ is the regret of the corresponding FTRL algorithm; term $\textcircled{2}$ corresponds to the error incurred by mixing with uniform distribution; term $\textcircled{3}$ measures the error of the clipping.

Bounding $\textcircled{1}$: We first bound the FTRL regret. The proof is founded on an observation that $0 \le \ell_{t,k_t}^c + C_t \le 2C_t$ for every $t \in [T]$, where C_t is a value known to the algorithm at the beginning of round t. In this case, we can tune the learning rate to fit the scale of the loss before observing it, thereby reducing the analysis to the bounded case. The main result is as follows. The detailed proof is delayed to Appendix A.1.

Lemma 3 *Algorithm 1 ensures*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\star\rangle\right] \le \Theta\left(\ell_\infty\sqrt{nT}\right),$$

where $\ell_\infty = \max_{t\in[T],k\in[n]}|\ell_{t,k}|$.

Bounding $\textcircled{2}$: The proof is trivial since

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\beta_t\langle \ell_t, \frac{\mathbf{1}_n}{n} - \mathbf{p}_t\rangle\right] \le 2\ell_\infty\sum_{t=1}^{T}\frac{n}{2n+\sqrt{nt}} \le 4\ell_\infty\sqrt{nT}.$$

Bounding $\textcircled{3}$: Bounding the clipping error is the key to the entire proof. Define $K := \arg\min_{j\in\mathbb{N}}\left\{\ell_\infty \le 2^j\right\}$. Define $\ell_t^i \in \mathbb{R}^n$ such that $\ell_{t,k}^i = \ell_{t,k}\mathbb{1}\{2^{i-1} < |\ell_{t,k}| \le 2^i\}$ for $k \in [n]$. Notice that $\ell_t = \sum_{i=-\infty}^{K}\ell_t^i$. In this case, there is

$$\mathbb{E}[\sum_{t=1}^{T}\langle \ell_t - \ell_t^c, \mathbf{p}_t - \mathbf{p}^\star\rangle] \le 2\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t - \ell_t^c\|_\infty\right] \le 2\mathbb{E}\left[\sum_{i=-\infty}^{K}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t^i - \ell_t^{ci}\|_\infty\right]\right].$$

We focus on the inner terms. We first note

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t^i - \ell_t^{ci}\|_\infty\right] \le \mathbb{E}\left[2^i\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\}\right]$$

since the clipping threshold is non-decreasing and all non-zero entries in ℓ_t^i are within $[2^{i-1}, 2^i]$. Now it suffices to bound $\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\}]$. To this end, an important observation is that for every integer $m \ge 1$, there is

$$\mathbb{P}\left\{\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\} \ge m\right\} \le \left(1 - \frac{\beta_T}{n}\right)^{m-1}. \tag{2}$$

This is because $\sum_{t=1}^{T} \mathbb{1}\{\ell_t^i \neq \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\} \geq m$ implies that the algorithm does not play the non-zero entries of the first $m-1$ non-zeros losses ℓ_t^i. Otherwise, in the m-th round where ℓ_t^i is non-zero, the clipping threshold should be no less than 2^i, which implies that no clipping can happen. Since each action has probability at least $\beta_t/n \geq \beta_T/n$ to be played every round, (2) immediately follows. Thus, there is

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\ell_t^i \neq \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\}\right] &= \sum_{m=1}^{\infty} \mathbb{P}\left\{\sum_{t=1}^{T} \mathbb{1}\{\ell_t^i \neq \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\} \geq m\right\}\\ &\leq \frac{n}{\beta_T} = 2n + \sqrt{nT}.\end{aligned}$$

Now we take the sum of all $i \leq K$.

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t - \ell_t', \mathbf{p}_t - \mathbf{p}^\star\rangle\right] &\leq 2\mathbb{E}\left[\sum_{i=-\infty}^{K} \mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t^i - \ell_t^{ci}\|_\infty\right]\right]\\ &\leq 2\mathbb{E}\left[\sum_{i=-\infty}^{K} 2^i(2n + \sqrt{nT})\right]\\ &\leq 2^{K+2}(2n + \sqrt{nT}) \leq 8\ell_\infty(2n + \sqrt{nT}).\end{aligned}$$

The last inequality is due to $2^{K-1} \leq \ell_\infty$. Combining ①,② and ③, we have

$$\mathbb{E}\left[\mathcal{R}(T)\right] \leq \Theta\left(\ell_\infty(n + \sqrt{nT})\right),$$

which is optimal upto constant factors.

3.2. High probability regret

Next, we study the more challenging problem of high-probability regret. The goal is to design algorithms for which $\mathcal{R}(T)$ can be bounded with high probability. We propose the first scale-free adversarial MAB algorithm with a high-probability regret guarantee.

The algorithm SCB-IX is provided in Algorithm 2. Conceptually, the algorithm is a variant of EXP3-IX in Neu (2015) combined with the clipping idea in Algorithm 1. By Hoeffding's inequality, it suffices to focus on bounding $\sum_{t=1}^{T}\langle \ell_t, \mathbf{q}_t - \mathbf{p}^\star\rangle$. Similar to the proof of Theorem 1, we can decompose the regret into $\sum_{t=1}^{T}\langle \ell_t^c + C_t\mathbf{1}_n, \mathbf{q}_t - \mathbf{p}^\star\rangle$ and $\sum_{t=1}^{T}\langle \ell_t - \ell_t^c, \mathbf{q}_t - \mathbf{p}^\star\rangle$. For the first term, due to $0 \leq \ell_{t,k}^c + C_t \leq 2C_t$, where C_t is known at the beginning of round t, it suffices to show that the regret can be well bounded with high probability based on the proof of EXP3-IX. For the second term, as shown in inequality (2), we have $\sum_{t=1}^{T} \mathbb{1}\{\ell_t^i \neq \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\} \geq \log(1/\delta)n/\beta_T$ with probability at least $1-\delta$, which immediately imply a high probability bound for the clipping error. Our results can be summarized in the following theorem.

Theorem 4 *With probability at least $1-\delta$, Algorithm 2 ensures*

$$\mathcal{R}(T) \leq \Theta\left(\ell_\infty\sqrt{\frac{n^2 + nT}{\log n}}\log(1/\delta) + \ell_\infty\sqrt{nT\log(n)}\right),$$

Algorithm 2: SCB-IX: Scale Clipping Bound with Implicit Exploration

```
Input: Shannon Entropy $\Psi$, $\eta_1 = \infty$, $\gamma_1 = 0$, $\beta_1 = \sqrt{n\log(n)/(n\log(n)+1)}$, $C_1 = 0$
for $t = 1, \ldots, T$ do
    Compute the action distribution $\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n}\left(\sum_{s=1}^{t-1}\langle\hat{\ell}_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p})\right)$
    Add extra exploration $\mathbf{q}_t = (1-\beta_t)\mathbf{p}_t + \beta_t\frac{\mathbf{1}_n}{n}$
    Sample and play action $k_t \sim \mathbf{q}_t$. Receive loss $\ell_{t,k_t}$
    Clip received loss by $[-C_t, C_t]$: $\ell^c_{t,k} = \max(-C_t, \min(C_t, \ell_{t,k}))$
    Construct estimator $\hat{\ell}_t$ such that $\hat{\ell}_{t,k} = \frac{\ell^c_{t,k}+C_t}{q_{t,k}+\gamma_t}\mathbb{1}\{k = k_t\}$, $\forall k \in [n]$
    If $|\ell_{t,k_t}| > C_t$, set $C_{t+1} = 2|\ell_{t,k_t}|$, otherwise $C_{t+1} = C_t$
    Update learning rate $\eta_{t+1} = \frac{1}{C_{t+1}}\sqrt{\frac{\log n}{n(t+1)}}$. Update exploration rate $\beta_{t+1} = \sqrt{\frac{n\log n}{n\log n+t+1}}$,
     $\gamma_{t+1} = \eta_{t+1}C_{t+1}/2$
end
```

Due to the space limit, the detailed proof is delayed to Appendix A.3. Specifically, when $T \geq n$, the regret reduces to $\Theta(\ell_\infty\sqrt{nT/\log n}\log(1/\delta) + \ell_\infty\sqrt{nT\log(n)})$[1]. This matches the results in Neu (2015) for the bounded loss setting.

4. Adversarial Markov Decision Process

With our preparation in the MAB setting, we now turn our attention to adversarial MDPs. We consider the episodic MDP setting with finite horizon, unknown transition matrix, bandit feedback, and adversarial losses, same as the setting in Jin et al. (2019). However, unlike Jin et al. (2019), where the losses are assumed to be in $[0, 1]$, we allow the losses to be *unbounded*. To the best of our knowledge, this is the first study of RL with unbounded losses.

An adversarial MDP is defined by a tuple $(S, A, P, \{\ell_t\}_{t=1}^T)$. S is the finite state space and A is the finite action space. $P : S\times A\times S \to [0, 1]$ is an unknown transition function where $P(s'|s, a)$ is the probability of reaching state s' after taking action a at state s. $\ell_t : S \times A \to \mathbb{R}$ is a loss function determined by the adversary, which can depend on the player's actions before t. Learning proceeds in T episodes. In each episode t, the learner starts from state x_1 and deploys a stochastic policy $\pi_t \in \Pi : S \times A \to [0, 1]$ with $\pi_t(a|s)$ being the probability of taking action a at state s. The learner observes a state-action-loss trajectory $(s_1, a_1, \ell_t(s_1, a_1), \ldots, s_H, a_H, \ell_t(s_H, a_H))$ before reaching the ending state s_{H+1}. With a slight abuse of notation, we assume $\ell_t(\pi) = \mathbb{E}[\sum_{h\in[H]}\ell_t(s_h, a_h)|P, \pi]$. The performance is measured by the regret, which is defined by

$$\mathcal{R}(T) = \sum_{t=1}^T \ell_t(\pi_t) - \min_{\pi\in\Pi}\sum_{t=1}^T \ell_t(\pi).$$

Without loss of generality, we consider a layered structure MDP: the state space is partitioned into $H + 2$ horizons $S_0, \ldots, S_{H+1}$ such that $S = \cup_{h=1}^H S_h$, $\emptyset = S_i \cap S_j$ for every $i \neq j$, $S_0 = \{s_0\}$ and $S_{H+1} = \{s_{H+1}\}$. We further assume that the number of states in each horizon is the same,

[1] Note that the bound scales linearly with $\log(1/\delta)$ for all levels δ. The dependence can be improved to $\sqrt{\log(1/\delta)}$ if the algorithm use δ to tune its parameter (Neu, 2015). This is the way to derive the results presented in Table 1.

i.e., $S_h = S/H$ for all $h = [H]$. Given the structure, with the help of "occupancy meansure" concept, this problem can be restructured in a way that makes it highly similar to adversarial MAB: denoted the probability that policy π visits the state-action pair (s, a) with transition function P by $q^{P,\pi}(s, a)$, the loss can be expressed as $\ell_t(\pi) = \sum_{s\in[S]} \sum_{a\in[A]} q^{P,\pi}(s, a)\ell_t(s, a) = \langle q^{P,\pi}, \ell_t\rangle$.

While we have formulated the loss function in a form similar to the ones in adversarial MAB with the help of occupancy measure, a significant distinction still exists, which also constitutes the main challenge: for adversarial MDP, there is no explicit "exploration policy" guaranteeing that every state can be visited. In particular, some states may be hardly accessible by any policy. In such cases, directly implementing the proposed scale-free MAB algorithms would result in unbounded clipping errors, as the algorithm is unable to detect the scale changes in states that are not accessible. In order to design scale-free algorithms for adversarial MDP, two critical questions need to be addressed: 1). How to find a good exploration policy for every state within $o(T)$ episodes? 2). How to handle the states that are hardly accessible for all policies?

To address these two questions, we design an exploration algorithm RF-ELP, as shown in Algorithm 3. Conceptually, for each state s that is accessible by some policy with a probability exceeding $\tilde{\mathcal{O}}(H\sqrt{SA/T})$, i.e., $\max_{\pi\in\Pi} q^{P,\pi}(s) \geq \tilde{\mathcal{O}}(H\sqrt{SA/T})$, RF-ELP is capable of producing a policy $\pi^{s,N}$ that successfully visit the state s at least once every $\mathcal{O}(\sqrt{ST/A \max_{\pi\in\Pi} q^{P,\pi}(s)})$ episodes. Additionally, for those states that are inaccessible by RF-ELP, we demonstrate that the maximum regret incurred by such a state can be bounded by $\tilde{\mathcal{O}}(\sum_{h\in[H]} \ell_{\infty,h}\sqrt{SAT})$. More details are provided in the next paragraph and the appendix. RF-ELP allows us to effectively reduces the problem of scale-free adversarial MDP to that of scale-free adversarial bandits. Building upon RF-ELP and UOB-REPS in Jin et al. (2019), we develop the main algorithm SCB-RL and subalgorithm UOB-REPS-EX. The pseudocode of SCB-RL is presented in Algorithm 4, and the pseudocode of UOB-REPS-EX is delayed to Appendix B.2.

Specifically, SCB-RL starts by calling RF-ELP for ξST episodes and obtains an exploration policy for each of the states. Then, in every episode t, it calls the subalgorithm UOB-REPS-EX to learn policy π_t, plays π_t and receives a trajectory, clips and adds offset on the loss, updates the clipping threshold, and sends the information back to UOB-REPS-EX. The subroutine UOB-REPS-EX is a variant of UOB-REPS, incorporating multiple designs for dealing with the unbounded losses, such as mixing with exploration policies and tuning the learning rate with the clipping threshold. More details about UOB-REPS-EX are provided in Appendix B.2. The main theorem for SCB-RL is below.

Theorem 5 *With probability at least $1 - \delta$, SCB-RL guarantees*

$$\mathcal{R}(T) \leq \tilde{\mathcal{O}}\left(\sum_{h\in[H]} \ell_{\infty,h} S^{3/2}\sqrt{AT}\right).$$

where we denote $\ell_{\infty,h} = \max_{t\in[T],s\in[S_h],a\in[A]} \ell_t(s, a)$.

Remark 6 *Compared to the best existing results with bounded losses $\tilde{\mathcal{O}}(\sum_{h\in[H]} \ell_{\infty,h} S\sqrt{AT})$, Theorem 5 achieves the same optimality in terms A, T but worse by a factor of $\sqrt{S}$. Nevertheless, this result is quite surprising, considering that we spend additional episodes to learn the exploration policy for every state in RF-ELP. Furthermore, if all states are visitable, for sufficiently large T, we can further reduce the regret of SCB-RL to $\tilde{\mathcal{O}}(\sum_{h\in[H]} \ell_{\infty,h} S\sqrt{AT})$ by designing an early*

Algorithm 3: Reward free exploration in RL (RF−ELP)

```
Input: State $s$; Exploration episodes number $N$
Output: Policy $\pi \in \Pi$
Initialize reward: $r^s(s', a') \leftarrow \mathbb{1}\{s' = s\}$ for all $(s', a') \in [S] \times A$
Run MVP (Zhang et al., 2023) $N$ episodes, get policies: $\{\pi_1^s, \dots, \pi_N^s\} \leftarrow$ MVP$(r^s, N)$, set
  $\pi^{s,N} \leftarrow \text{Uniform}(\pi_1^s, \dots, \pi_N^s)$
Set policy $\pi^{s,N}(\cdot|s) \leftarrow \text{Uniform}(A)$
Return $\pi^{s,N}$
```

Algorithm 4: SCB−RL: Scale Clipping Bound for RL

```
Input: state space $S$, action space $A$, episode number $T$, state exploration parameter $\xi$
Initialize: Clipping threshold $C_{1,h} = 0$ for $h \in [H]$
for $s \in [S]$ do
    Run RF−ELP and update exploration policy: $\pi^s \leftarrow$ RL−ELP$(s, \xi T)$
end
Send extra exploration policies $\{\pi^s\}_{s\in[S]}$ to UOB−REPS−EX
for $t = \xi ST + 1$ to $T$ do
    Receive policy $\pi_t \leftarrow$ UOB−REPS−EX
    Execute policy $\pi_t$ for $H$ horizons and obtain trajectory $\{s_h, a_h, \ell_t(s_h, a_h)\}_{h\in[H]}$
    Clip received loss: $\ell_t^c(s_h, a_h) = \max(-C_{t,h}, \min(C_{t,h}, \ell_t(s_h, a_h)))$, $\forall h \in [H]$
    Send trajectory $\{s_h, a_h, \ell_t^c(s_h, a_h) + C_{t,h}\}_{h\in[H]}$ and clipping threshold $\{C_{t,h}\}_{h\in[H]}$ to
      UOB−REPS−EX.
    If $|\ell_t(s_h, a_h)| > C_{t,h}$, set $C_{t+1,h} = 2|\ell_t(s_h, a_h)|$, otherwise $C_{t+1,h} = C_{t,h}$, $\forall h \in [H]$
end
```

stopping strategy on RF−ELP, matching the best known regret in the bounded loss setting. We delay the details of this extension to Appendix B.11.

Proof Sketch: We start with the exploration algorithm RF−ELP. As illustrated in Algorithm 3, the goal of RF−ELP is to find a set of policies each capable of visiting a particular state s within N episodes. RF−ELP is essentially a reward-free exploration algorithm with a similar structure to that in Jin et al. (2020), while we replace the RL algorithm used for exploration from EULER (Zanette and Brunskill, 2019) to MVP (Zhang et al., 2023). Specifically, RF−ELP starts by defining the reward r^s as $r^s(s', a') = 1$ if and only if $s' = s$, and then run MVP for N episodes and get policy $\pi^{s,N}$. Following this, RF−ELP resets the action distribution for state s to ensure accessibility for every action. We first present the theoretical guarantee of MVP as follows.

Lemma 7 (Theorem 3 of Zhang et al. (2023)) [2] *For any $N \geq 1$ and $s \in [S]$, with probability at least $1 - \delta$, MVP obeys*

$$\max_{\pi\in\Pi} \mathbb{E}\left[\sum_{h\in[H]} r^s(s_h, a_h)|P, \pi\right] - \mathbb{E}\left[\sum_{h\in[H]} r^s(s_h, a_h)|P, \pi^{s,N}\right] \leq \tilde{O}\left(\sqrt{\frac{SA\text{Var}^s}{N}} + \frac{SAH}{N}\right)$$

[2] Notice that in this work S represents the collection of states in all horizons, corresponding to SH in Zhang et al. (2023).

where $\mathrm{Var}^s = \max_{\pi\in\Pi} \mathrm{Var}\left[\sum_{h\in[H]} r^s(s_h,a_h)|P,\pi\right]$.

Based on Lemma 7, we can derive the theoretical guarantee of RF-ELP. The key observation is that Var^s can be bounded by $\max_{\pi\in\Pi}\mathbb{E}[\sum_{h\in[H]} r^s(s_h,a_h)|P,\pi]$ due to $\sum_{h\in[H]} r^s(s_h,a_h)\le 1$. By the setting of r^s, it suffices to note that $\mathbb{E}[\sum_{h\in[H]} r^s(s_h,a_h)|P,\pi] = q^{P,\pi}(s)$ for every $\pi\in\Pi$. In this case, when $\max_{\pi\in\Pi} q^{P,\pi}(s)\ge \tilde{\mathcal{O}}(SAH/N)$, $\max_{\pi\in\Pi} q^{P,\pi}(s)$ and $q^{P,\pi^{s,N}}(s)$ should be on the same order. We summarize the result in the following lemma.

Lemma 8 (RF-ELP guarantee) *For any $N\ge 1$ and $s\in[S]$, with probability at least $1-\delta$, if $\max_{\pi\in\Pi} q^{P,\pi}(s) > 9C^2\left(\frac{SAH}{N}\right)$, then we have $q^{P,\pi^{s,N}}(s)\ge \frac{1}{2}\max_{\pi\in\Pi} q^{P,\pi}(s)$, where $C = \mathcal{O}(\log^3(T)\log(SA)\log(\frac{SAHT}{\delta}))$ is a poly-log factor w.r.t. S, A, H, N and $1/\delta$.*

The proof of Lemma 8 is proposed in Appendix B.3. Now we begin to prove Theorem 5. As illustrated in Algorithm 4, SCB-RL calls the exploration algorithm RF-ELP in the first ξST episodes. For simplicity of the proof, we let t start from $-\xi ST+1$ instead of 1. Denoted by $q_t = q^{P,\pi_t}$, as in Jin et al. (2019), the total regret can be written as $\sum_{t=-\xi ST+1}^{T}\langle \ell_t, q_t - q^*\rangle$. We first decompose the regret into

$$\begin{aligned}\sum_{t=-\xi ST+1}^{T}\langle \ell_t, q_t-q^*\rangle &= \sum_{t=-\xi ST+1}^{0}\langle \ell_t, q_t-q^*\rangle + \sum_{t=1}^{T}\langle \ell_t^c, q_t-q^*\rangle + \sum_{t=1}^{T}\langle \ell_t-\ell_t^c, q_t-q^*\rangle\\ &\le \sum_{h\in[H]}\ell_{\infty,h}\xi ST + \underbrace{\sum_{t=1}^{T}\langle \ell_t^+, q_t-q^*\rangle}_{\textcircled{1}} + \underbrace{\sum_{t=1}^{T}\langle \ell_t-\ell_t^c, q_t-q^*\rangle}_{\textcircled{2}},\end{aligned}$$

where $\ell_t^c(s,a)\in\mathbb{R}^{SA}$ and $\ell_t^+(s,a)\in\mathbb{R}_+^{SA}$ satisfy $\forall (s,a)\in[S]\times[A]$ [3],

$$\begin{aligned}\ell_t^c(s,a) &= \max\left(-C_{t,h(s)}, \min(C_{t,h(s)}, \ell_t(s,a))\right),\\ \ell_t^+(s,a) &= \ell_t^c(s,a) + C_{t,h(s)}.\end{aligned}$$

Bounding $\textcircled{1}$: The regret of $\textcircled{1}$ is incurred by UOB-REPS-EX. Similarly to SCB-IX, our algorithm differs to UOB-REPS in two key aspects: the mixing of explicit exploration and the presence of loss within the range of $[0, 2C_t]$ rather than $[0,1]$. The result is stated below and the detailed proof is delayed to Appendix B.4.

Lemma 9 *With probability at least $1-\delta$, there is ·*

$$\sum_{t=1}^{T}\langle \ell_t^+, q_t-q^*\rangle \le \mathcal{O}\left(\sum_{h\in[H]}\ell_{\infty,h} S\sqrt{AT\ln\left(\frac{SAT}{\delta}\right)} + \beta T\sum_{h\in[H]}\ell_{\infty,h}\right).$$

[3] $h(s)$ is the index of the layer to which s belongs.

Bounding $\textcircled{2}$: For simplicity, we denote by $q^s = \max_{\pi \in \Pi} q^{P,\pi}(s)$ and $\ell'_t = |\ell_t - \ell^c_t|$. Notice that

$$\begin{aligned}\sum_{t=1}^{T} \langle \ell_t - \ell^c_t, q_t - q^* \rangle &\leq \sum_{s \in [S], a \in [A]} \sum_{t=1}^{T} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \\ &= \sum_{s \in [S], a \in [A]} \sum_{t=1}^{T} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \mathbb{1} \left\{ q^s \leq \tilde{\mathcal{O}}(SAH/\xi T) \right\} \\ &\quad + \sum_{s \in [S], a \in [A]} \sum_{t=1}^{T} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \mathbb{1} \left\{ q^s > \tilde{\mathcal{O}}(SAH/\xi T) \right\}\end{aligned}$$

For the first term, we can bound it directly by

$$\begin{aligned}&\sum_{s \in [S], a \in [A]} \sum_{t=1}^{T} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \mathbb{1} \left\{ q^s \leq \tilde{\mathcal{O}}(SAH/\xi T) \right\} \\ &\leq \sum_{h \in [H]} \ell_{\infty,h} \sum_{s \in [S_h]} 2q^s T = \tilde{\mathcal{O}} \left(\frac{\sum_{h \in [H]} \ell_{\infty,h} S_h SAH}{\xi} \right).\end{aligned}$$

This first inequality is due to $\sum_{a \in [A]} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \leq 2\ell_{\infty,h(s)} q^s$ for all $s \in [S]$.

It suffices to focus on the second term. By Lemma 8, for every $s \in [S]$, if $q^s > \tilde{\mathcal{O}}(SAH/\xi T)$, q^s and $q^{P,\pi^{s,\xi T}}$ will be on the same order. Thus, when

$$\sum_{a \in [A]} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \mathbb{1} \left\{ q^s > \tilde{\mathcal{O}}(SAH/\xi T) \right\} \neq 0,$$

the extra exploration policy ensures the outlier loss of (s,a) has probability at least $\tilde{\mathcal{O}}(\beta q^s / SA)$ to be visited. Moreover, we can note that $\sum_{a \in [A]} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \leq \mathcal{O}(q^s \ell_{\infty,h(s)})$, and thus the terms dependent on q^s can be eliminated. The result is presented below.

Lemma 10 *With probability at least* $1 - \delta$*,*

$$\sum_{s \in [S], a \in [A]} \sum_{t=1}^{T} \ell'_t(s,a) |q_t(s,a) - q^*(s,a)| \mathbb{1} \left\{ q^s > \tilde{\mathcal{O}}(SAH/\xi T) \right\} \leq \tilde{\mathcal{O}} \left(\frac{\sum_{h \in [H]} \ell_{\infty,h} SA}{\beta} \right),$$

where the proof is in Appendix B.5. Summing up all the terms lead to regret

$$\mathcal{R}(T) \leq \tilde{\mathcal{O}} \left(\sum_{h \in [H]} \ell_{\infty,h} \left[S\sqrt{AT} + \beta T + \frac{S_h SHA}{\xi} + \frac{SA}{\beta} + \xi ST \right] \right)$$

Setting $\xi = \mathcal{O}(\sqrt{SA/T})$, $\beta = \mathcal{O}(\sqrt{SA/T})$ and $S_h = S/H$ concludes the proof.

5. Conclusion

This paper initiates the study of scale-free learning in adversarial MDPs. Our framework `SCB` allows us to achieve the minimax optimal expected regret for scale-free adversarial MABs and the first known high-probability regret in both scale-free adversarial MAB and scale-free adversarial MDPs. Future work includes closing the gap in the S-dependency in the adversarial MDPs regret.

References

Jacob Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. In *2009 Information Theory and Applications Workshop*, pages 280–289. IEEE, 2009.

Jean-Yves Audibert, Sébastien Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

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

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127. PMLR, 2018.

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

Mingyu Chen and Xuezhou Zhang. Improved algorithms for adversarial bandits with unbounded losses. *arXiv preprint arXiv:2310.01756*, 2023.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pages 874–894. PMLR, 2019.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. *Advances in Neural Information Processing Systems*, 35:11437–11449, 2022.

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k–armed bandits. *Journal of Machine Learning Research*, 24(13):1–33, 2023.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Jiatai Huang, Yan Dai, and Longbo Huang. Banker online mirror descent: A universal approach for delayed online bandit learning. *arXiv preprint arXiv:2301.10500*, 2023.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583. PMLR, 2021.

Andrew Jacobsen and Ashok Cutkosky. Unconstrained online learning with unbounded losses. *arXiv preprint arXiv:2306.04923*, 2023.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems*, 27, 2014.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533, 2020.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34: 22931–22942, 2021.

Jack J Mayo, Hédi Hadiji, and Tim van Erven. Scale-free unconstrained online learning for curved losses. In *Conference on Learning Theory*, pages 4464–4497. PMLR, 2022.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Sudeep Raja Putta and Shipra Agrawal. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 910–930. PMLR, 2022.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR, 2018.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023.

Appendix A. Omitted details for Section 3

A.1. Proof of Lemma 3

We start by a technical lemma inspired by Chen and Zhang (2023).

Lemma 11 *For any $\hat{\ell}_1, \dots, \hat{\ell}_T \geq 0$, using the update rule of (1), consider any convex regularizer $\Psi \geq 0$ that satisfies $\nabla_{k,k}\Psi(\mathbf{p}) \leq \nabla_{k,k}\Psi(\mathbf{q})$ iff. $p_k \geq q_k$ and $\mathbf{p}, \mathbf{q} \in \Delta_n$. With non-increasing sequence of learning rates $\eta_1, \dots, \eta_{T+1}$, there is*

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \leq \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}$$

for every comparator $\mathbf{p}^\dagger \in \Delta_n$.

Recall the definition of $1/2$-Tsallis Entropy

$$\Psi(\mathbf{p}_t) = 4\sqrt{n} - 4\sum_{k=1}^{n} \sqrt{p_{t,k}}.$$

Notice that $\Psi(\mathbf{p}) \geq 0$ for every $\mathbf{p} \in \Delta_n$ and $\nabla_{k,k}\Psi(\mathbf{p}) = p_k^{-3/2} \leq q_k^{-3/2} = \nabla_{k,k}\Psi(\mathbf{q})$ when $p_k \geq q_k$. Using Lemma 11, there is

$$\begin{aligned}
\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\star \rangle &\leq \frac{\Psi(\mathbf{p}^\star)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}} \\
&\leq \frac{\Psi(\mathbf{p}^\star)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \frac{(\ell^c_{t,k_t} + C_t)^2}{q^2_{t,k_t}} p^{3/2}_{t,k_t} \\
&\leq \frac{\Psi(\mathbf{p}^\star)}{\eta_{T+1}} + 8\sum_{t=1}^{T} \eta_t \frac{(\ell^c_{t,k_t} + C_t)^2}{\sqrt{q_{t,k_t}}}
\end{aligned}$$

The last inequality is due to $p_{t,k_t} \leq q_{t,k_t}/(1-\beta_t) \leq q_{t,k_t}/(1 - n/(2n + \sqrt{nt})) \leq 2q_{t,k_t}$. We further note that, by the clipping rule, there is $|\ell^c_{t,k_t} + C_t| \leq 2C_t$. Moreover, it suffices to say that $C_{t+1} \leq 2\ell_\infty$ for all $t \in [T]$. Remind $\Psi(\mathbf{p}) \leq \sqrt{n}$ for every $\mathbf{p} \in \Delta_n$. Thus, by choosing learning rate $\eta_t = 1/2C_t\sqrt{t}$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\star \rangle\right] &\leq \frac{\Psi(\mathbf{p}^\star)}{\eta_{T+1}} + 8\sum_{t=1}^{T} \mathbb{E}\left[\eta_t \frac{(\ell^c_{t,k_t} + C_t)^2}{\sqrt{q_{t,k_t}}}\right] \\
&\leq \frac{\Psi(\mathbf{p}^\star)}{\eta_{T+1}} + 8\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{E}\left[\frac{|\ell^c_{t,k_t} + C_t|}{\sqrt{q_{t,k_t}}}\right] \\
&\leq 4\ell_\infty\sqrt{n(T+1)} + 16\ell_\infty\sqrt{n}(2\sqrt{T} + 1) \\
&= \Theta(\ell_\infty\sqrt{nT}),
\end{aligned}$$

where the third inequality is due to $\mathbb{E}[1/\sqrt{q_{t,k_t}}] = \sum_{k=1}^{n} \sqrt{q_{t,k_t}} \leq \sqrt{n}$.

A.2. Omitted details of Remark 2

Here we prove that `SCB` is invariant to rescaling of losses. Notice that the proof also applies to `SCB-IX` and `SCB-RL`. Starting with losses $\ell_1, \dots, \ell_T$, the rescaled losses are defined by $\ell'_1, \dots, \ell'_T$ such that $\ell_t = c\ell'_t$ for all $t \in [T]$ and $c > 0$. With the use of `SCB`, the action distributions corresponding to these two sequences of losses are represented by $\mathbf{p}_1, \dots, \mathbf{p}_T$ and $\mathbf{p}'_1, \dots, \mathbf{p}'_T$, respectively. Notice that $\mathbf{p}_t$ can be considered as a random vector w.r.t. `SCB`, losses $\ell_1, \dots, \ell_{t-1}$ and past actions $k_1, \dots, k_t$. Our goal is to prove that the distributions of $\mathbf{p}_1, \dots, \mathbf{p}_T$ and $\mathbf{p}'_1, \dots, \mathbf{p}'_T$ are the same. We prove by induction. For $t = 1$, both $\mathbf{p}_1$ and $\mathbf{p}'_1$ are uniform distribution over $[n]$. Assuming at time t, the distributions of $\mathbf{p}_1, \dots, \mathbf{p}_t$ and $\mathbf{p}'_1, \dots, \mathbf{p}'_t$ are the same. Conditioned on $\mathbf{p}_1, \dots, \mathbf{p}_t = \mathbf{p}'_1, \dots, \mathbf{p}'_t$, since $\beta_1, \dots, \beta_t$ are independent to the losses, for `SCB` with these two loss sequences, the probability of taking actions $\{k_1, \dots, k_t\}$ is the same. Conditioned on actions $\{k_1, \dots, k_t\}$, we have $\hat{\ell}_s = c\hat{\ell}'_s$ for all $s \le t$. Then, since the clipping threshold is twice the largest scale among the previously observed losses, we further have $C_{t+1} = cC'_{t+1}$. Thus, by the update rule, it suffices to show that $\mathbf{p}_{t+1} = \mathbf{p}'_{t+1}$. This implies that the distributions of $\mathbf{p}_1, \dots, \mathbf{p}_{t+1}$ and $\mathbf{p}'_1, \dots, \mathbf{p}'_{t+1}$ are the same, which completes the proof.

We emphasize that the use of clipping (or skipping) to deal with unbounded losses has been studied before (Chen and Zhang, 2023; Huang et al., 2023). However, our algorithms fundamentally differ from the previous ones. In previous works, the update of the clipping threshold is accomplished through a double trick, i.e., $C_{t+1} = 2C_t$. This leads to an inevitable logarithm sub-optimality. More importantly, their algorithms must start from a positive clipping threshold $C_1 > 0$, resulting in a failure to achieve strongly scale-free. Relatively, our algorithm starts from $C_1 = 0$. This allows the clipping threshold to be linearly related to the scale of the losses, thereby achieving strongly scale-free.

A.3. Proof of Theorem 4

By Hoeffding's inequality, there is

$$\sum_{t=1}^{T} \ell_{t,k_t} - \sum_{t=1}^{T} \ell_{t,k^\star} \le \ell_\infty \sqrt{2T \log(1/\delta)} + \sum_{t=1}^{T} \langle \ell_t, \mathbf{q}_t - \mathbf{p}^\star \rangle$$

with probability at least $1 - \delta$. It suffices to focus on bounding $\sum_{t=1}^{T} \langle \ell_t, \mathbf{q}_t - \mathbf{p}^\star \rangle$. Similar to the proof of Theorem 1, we decompose the regret into

$$\sum_{t=1}^{T} \langle \ell_t, \mathbf{q}_t - \mathbf{p}^\star \rangle = \underbrace{\sum_{t=1}^{T} \langle \ell_t^c + C_t \mathbf{1}_n, \mathbf{q}_t - \mathbf{p}^\star \rangle}_{\textcircled{1}} + \underbrace{\sum_{t=1}^{T} \langle \ell_t - \ell_t^c, \mathbf{q}_t - \mathbf{p}^\star \rangle}_{\textcircled{2}}$$

and bound these two terms respectively.

Bounding $\textcircled{1}$: The high level idea of bounding $\textcircled{1}$ is inspired by the proof of Theorem 1 in Neu (2015). Compared to `EXP3-IX`, our algorithm adds an additional explicit exploration, i.e., mixes $\mathbf{p}_t$ with uniform distribution, and $\ell_{t,k}^c + C_t$ is within $[0, 2C_t]$ instead of $[0, 1]$. Using a similar idea proposed in the previous section, we can show that $\textcircled{1}$ can be well bounded with high probability. The detailed proof is provided in the appendix.

Lemma 12 *With probability at least* $1-\delta$*, there is*

$$\sum_{t=1}^{T}\langle \ell_t^c + C_t\mathbf{1}_n, \mathbf{q}_t - \mathbf{p}^\star\rangle \le \Theta\left(\ell_\infty\sqrt{\frac{nT}{\log n}}\log(1/\delta) + \ell_\infty\sqrt{nT\log(n)}\right).$$

Bounding ②: Define $K := \arg\min_{j\in\mathbb{N}}\left\{\ell_\infty \le 2^j\right\}$. Define $\ell_t^i \in \mathbb{R}^n$ such that $\ell_{t,k}^i = \ell_{t,k}\mathbb{1}\{2^{i-1} < \ell_{t,k} \le 2^i\}$ for $k \in [n]$. Inspired by the results of the above section, we note that the clipping error can be reduced to the sum of the error incurred by losses within $[2^{i-1}, 2^i]$, i.e.,

$$\sum_{t=1}^{T}\langle \ell_t - \ell_t^c, \mathbf{p}_t - \mathbf{p}^\dagger\rangle \le 2\sum_{i=-\infty}^{K}\sum_{t=1}^{T}\|\ell_t^i - \ell_t^{ci}\|_\infty \le 2\sum_{i=-\infty}^{K}2^i\left(\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\}\right)$$

Apparently, bounding $\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\}$ individually is not difficult due to

$$\mathbb{P}\left\{\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i \ne \mathbf{0}_n\}\mathbb{1}\{C_t < 2^i\} > \frac{n}{\beta_t}\log(1/\delta)\right\} \le \left(1 - \frac{\beta_t}{n}\right)^{\frac{n}{\beta_t}\log(1/\delta)} \le \delta.$$

The challenge is how to achieve a union bound on all $i \le K$ without losing any optimality of the logarithmic terms. This is shown in the following lemma.

Lemma 13 *With probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T}\langle \ell_t - \ell_t^c, \mathbf{p}_t - \mathbf{p}^\star\rangle \le \Theta\left(\ell_\infty\sqrt{\frac{n^2 + nT}{\log n}}\log(1/\delta)\right)$$

Given Lemma 12 and Lemma 13, we can obtain the results of Theorem 4.

A.4. Proof of Lemma 11

The proof refers to Lemma 1 and 2 in Chen and Zhang (2023). Define

$$F_t(\mathbf{p}) = \sum_{s=1}^{t-1}\langle \hat{\ell}_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p}),$$

we first note that

$$\begin{aligned}
\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle &= -F_{T+1}(\mathbf{p}^\dagger) + \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) + \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t\rangle \\
&= -F_{T+1}(\mathbf{p}^\dagger) + \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) + F_{T+1}(\mathbf{p}_{T+1}) \\
&\quad + \sum_{t=1}^{T}\left(F_t(\mathbf{p}_t) - F_{t+1}(\mathbf{p}_{t+1})\right) + \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t\rangle \\
&= -F_{T+1}(\mathbf{p}^\dagger) + \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) + F_{T+1}(\mathbf{p}_{T+1}) \\
&\quad + \sum_{t=1}^{T}\Big(F_t(\mathbf{p}_t) + \langle \hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1})\Big).
\end{aligned}$$

By definition, there is

$$\begin{aligned} F_{T+1}(\mathbf{p}_{T+1}) - F_{T+1}(\mathbf{p}^\dagger) &= \min_{\mathbf{p}\in\Delta_n} F_{T+1}(\mathbf{p}) - F_{T+1}(\mathbf{p}^\dagger) \le 0 \\ \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) &= \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) - \min_{\mathbf{p}\in\Delta_n} \Psi_1(\mathbf{p}) \le \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger). \end{aligned}$$

Thus, we obtain

$$\sum_{t=1}^T \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle \le \frac{1}{\eta_{T+1}}\Psi(\mathbf{p}^\dagger) + \sum_{t=1}^T \Big(F_t(\mathbf{p}_t) + \langle \hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1})\Big)$$

Furthermore, we note that

$$\begin{aligned} F_t(\mathbf{p}_t) + \langle \hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1}) &= \sum_{s=1}^t \langle \hat{\ell}_s, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p}_t) - \frac{1}{\eta_{t+1}}\Psi(\mathbf{p}_t) \\ &\le \sum_{s=1}^t \langle \hat{\ell}_s, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p}_t) - \frac{1}{\eta_t}\Psi(\mathbf{p}_t) \\ &= \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}), \end{aligned}$$

where the first inequality is due to the assumption $\eta_{t+1} \le \eta_t$. By Taylor's expansion, we have

$$F_t(\mathbf{p}_{t+1}) - F_t(\mathbf{p}_t) = \langle \nabla F_t(\mathbf{p}_t), \mathbf{p}_{t+1} - \mathbf{p}_t\rangle + \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)}.$$

where $\xi_t = \alpha\mathbf{p}_t + (1-\alpha)\mathbf{p}_{t+1}$ for some $\alpha \in [0,1]$. By definition,

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} F_t(\mathbf{p}).$$

By KKT conditions, there exists some $\lambda_t \in \mathbb{R}$ such that

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\mathbb{R}} \Big(F_t(\mathbf{p}) + \lambda_t(1 - \sum_{k=1}^n p_{t,k})\Big).$$

By the optimality of $\mathbf{p}_t$, we have

$$\nabla F_t(\mathbf{p}_t) + \lambda_t \mathbf{1}_n = 0,$$

which implies

$$\langle \nabla F_t(\mathbf{p}), \mathbf{p}_{t+1} - \mathbf{p}_t\rangle = \langle -\lambda_t\mathbf{1}_n, \mathbf{p}_{t+1} - \mathbf{p}_t\rangle = 0.$$

Thus, there is

$$F_t(\mathbf{p}_{t+1}) - F_t(\mathbf{p}_t) = \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)}.$$

Using the above,

$$\begin{aligned}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &= \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle - \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)} \\ &\leq \max_{\mathbf{p}\in\mathbb{R}} \left(\langle \hat{\ell}_t, \mathbf{p}\rangle - \frac{1}{2}\|\mathbf{p}\|^2_{\nabla^2 F_t(\xi_t)}\right) \\ &\leq \frac{1}{2}\|\hat{\ell}_t\|^2_{(\nabla^2 F_t(\xi_t))^{-1}} = \frac{1}{2}\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2 \Psi(\xi_t))^{-1}},\end{aligned}$$

where the second inequality is because $\nabla^2\Psi(\xi_t)$ is a diagonal matrix and the second equality is due to $\nabla^2 F_t(\xi_t) = \nabla^2\Psi(\xi_t)/\eta_t$. Now we prove

$$\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \leq \frac{1}{2}\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}$$

if $\hat{\ell}_t \in \mathbb{R}^n_+$. Recall

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \sum_{k=1}^n \frac{\hat{\ell}^2_{t,k}}{\nabla^2_{k,k}\Psi(\xi_t)} = \frac{\hat{\ell}^2_{t,k_t}}{\nabla^2_{k_t,k_t}\Psi(\xi_t)}$$

and ξ_t is between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$, we prove case by case.

1. $(p_{t,k_t} - p_{t+1,k_t} < 0)$: In this case, we have

$$\begin{aligned}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &\leq \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle \\ &= \hat{\ell}_{t,k_t}(p_{t,k_t} - p_{t+1,k_t}) \\ &\leq 0 \leq \frac{1}{2}\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}.\end{aligned}$$

   The first inequality is due to $\mathbf{p}_t$ minimizing F_t. The second inequality is due to $\hat{\ell}_{t,k_t} \geq 0$.

2. $(p_{t,k_t} - p_{t+1,k_t} \geq 0)$: In this case, we have $p_{t,k_t} \geq \xi_{t,k_t}$. By assumption, there is $\nabla^2_{k_t,k_t}\Psi(\mathbf{p}_t) \leq \nabla^2_{k_t,k_t}\Psi(\xi_t)$. Thus

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \frac{\hat{\ell}^2_{t,k_t}}{\nabla^2_{k_t,k_t}\Psi(\xi_t)} \leq \frac{\hat{\ell}^2_{t,k_t}}{\nabla^2_{k_t,k_t}\Psi(\mathbf{p}_t)} = \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}$$

   completes the proof.

A.5. Proof of Lemma 12

We start by introducing a concentration result of the implicit exploration estimator based on Lemma 1 in Neu (2015).

Lemma 14 *Let $\gamma_1, \ldots, \gamma_T$ be a fixed non-increasing sequence with $\gamma_t \geq 0,\ \forall t \in [T]$ and $\alpha_{t,k}$ be non-negative $\mathcal{F}_{t-1}$ measurable random variables satisfying $\alpha_{t,k} \leq 2\gamma_t,\ \forall t \in [T], k \in [n]$. Then, with probability at least $1-\delta$,*

$$\sum_{t=1}^T \sum_{k=1}^n \alpha_{t,k}\left(\hat{\ell}_{t,k} - (\ell^c_{t,k} + C_t)\right) \leq 3\ell_\infty \log(1/\delta).$$

Given Lemma 14, we decompose the $\textcircled{1}$ into 4 terms.

$$\sum_{t=1}^{T}\langle \ell_t^c + C_t\mathbf{1}_n, \mathbf{q}_t - \mathbf{p}^\dagger\rangle = \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle + \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{q}_t - \mathbf{p}_t\rangle + \sum_{t=1}^{T}\langle \ell_t^c + C_t\mathbf{1}_n - \hat{\ell}_t, \mathbf{q}_t - \mathbf{p}^\dagger\rangle$$
$$= \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle + \sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{q}_t - \mathbf{p}_t\rangle + \sum_{t=1}^{T}\langle \ell_t^c + C_t\mathbf{1}_n - \hat{\ell}_t, \mathbf{q}_t\rangle + \sum_{t=1}^{T}\langle \hat{\ell}_t - (\ell_t^c + C_t\mathbf{1}_n), \mathbf{p}^\dagger\rangle.$$

For the first term, recall the definition of (negative) Shannon Entropy

$$\Psi(\mathbf{p}_t) = \log n + \sum_{k=1}^{n} p_{t,k}\log(p_{t,k}).$$

Notice that $0 \le \Psi(\mathbf{p}) \le \log n$ for every $\mathbf{p} \in \Delta_n$ and $\nabla_{k,k}\Psi(\mathbf{p}) = p_k^{-1} \le q_k^{-1} = \nabla_{k,k}\Psi(\mathbf{q})$ when $p_k \ge q_k$. Using Lemma 11, there is

$$\begin{aligned}\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle &\le \frac{\log n}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\eta_t p_{t,k_t}(\hat{\ell}_{t,k_t})^2\\ &\le \frac{\log n}{\eta_{T+1}} + 2\sum_{t=1}^{T}\eta_t C_t \hat{\ell}_{t,k_t}\\ &\le 3\ell_\infty\sqrt{n(T+1)\log(n)} + 4\sum_{t=1}^{T}\gamma_t\hat{\ell}_{t,k_t}.\end{aligned}$$

The second inequality is due to $p_{t,k_t}\hat{\ell}_{t,k_t} = p_{t,k_t}(\ell_{t,k_t}^c + C_t)/(q_{t,k_t} + \gamma_t) \le 4C_t$. Since $\eta_t C_t \le 2\gamma_t$, using Lemma 14 and setting $\alpha_{t,k} = \gamma_t$ for every $k \in [n]$, we have

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t\hat{\ell}_{t,k_t} &\le \sum_{t=1}^{T}\gamma_t\sum_{k=1}^{n}(\ell_{t,k}^c + C_t) + 3\ell_\infty\log(1/\delta)\\ &\le 3\ell_\infty\sqrt{\log(n)}\sum_{t=1}^{T}\sqrt{\frac{n}{t}} + 3\ell_\infty\log(1/\delta)\\ &\le 6\ell_\infty\sqrt{nT\log(n)} + 3\ell_\infty\log(1/\delta).\end{aligned}$$

The second inequality is due to $\ell_{t,k}^c + C_t \le 3\ell_\infty$ for all $t \in [T]$. Combining the above, there is

$$\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle \le \Theta\left(\ell_\infty\sqrt{nT\log(n)} + \ell_\infty\log(1/\delta)\right)$$

with probability at least $1-\delta$.

For the second term, we note that

$$
\begin{aligned}
\sum_{t=1}^{T}\langle\hat{\ell}_t,\mathbf{q}_t-\mathbf{p}_t\rangle &\le \sum_{t=1}^{T}\beta_t\left\langle\hat{\ell}_t,\frac{\mathbf{1}_n}{n}\right\rangle\\
&\le \sum_{t=1}^{T}\beta_t\left\langle\hat{\ell}_t-(\ell_t^c+C_t\mathbf{1}_n),\frac{\mathbf{1}_n}{n}\right\rangle+\sum_{t=1}^{T}\beta_t\left\langle\ell_t^c+C_t\mathbf{1}_n,\frac{\mathbf{1}_n}{n}\right\rangle\\
&\le \sum_{t=1}^{T}\beta_t\left\langle\hat{\ell}_t-(\ell_t^c+C_t\mathbf{1}_n),\frac{\mathbf{1}_n}{n}\right\rangle+6\ell_\infty\sqrt{nT\log(n)}.
\end{aligned}
$$

Since $\beta_t/n\le 2\gamma_t$, using Lemma 14 and setting $\alpha_{t,k}=\beta_t/n$ for every $k\in[n]$, we have

$$
\sum_{t=1}^{T}\beta_t\left\langle\hat{\ell}_t-(\ell_t^c+C_t\mathbf{1}_n),\frac{\mathbf{1}_n}{n}\right\rangle\le\sum_{t=1}^{T}2\gamma_t\sum_{k=1}^{n}\left(\hat{\ell}_{t,k}-(\ell_{t,k}^c+C_t)\right)\le 3\ell_\infty\log(1/\delta).
$$

Thus we have

$$
\sum_{t=1}^{T}\langle\hat{\ell}_t,\mathbf{q}_t-\mathbf{p}_t\rangle\le\Theta\left(\ell_\infty\sqrt{nT\log(n)}+\ell_\infty\log(1/\delta)\right)
$$

with probability at least $1-\delta$.

For the third term, there is

$$
\begin{aligned}
\sum_{t=1}^{T}\langle\ell_t^c+C_t\mathbf{1}_n-\hat{\ell}_t,\mathbf{q}_t\rangle &= \sum_{t=1}^{T}\sum_{k=1}^{n}q_{t,k}(\ell_{t,k}^c+C_t-\hat{\ell}_{t,k})\\
&=\sum_{t=1}^{T}\sum_{k=1}^{n}q_{t,k}\left(\ell_{t,k}^c+C_t-\frac{\mathbb{1}\{k=k_t\}}{q_{t,k}+\gamma_t}(\ell_{t,k}^c+C_t)\right)\\
&=\sum_{t=1}^{T}\sum_{k=1}^{n}q_{t,k}\left(\ell_{t,k}^c+C_t-\frac{q_{t,k}}{q_{t,k}+\gamma_t}(\ell_{t,k}^c+C_t)\right)\\
&+\sum_{t=1}^{T}\sum_{k=1}^{n}q_{t,k}\left(\frac{q_{t,k}}{q_{t,k}+\gamma_t}(\ell_{t,k}^c+C_t)-\frac{\mathbb{1}\{k=k_t\}}{q_{t,k}+\gamma_t}(\ell_{t,k}^c+C_t)\right)\\
&\le\sum_{t=1}^{T}\sum_{k=1}^{n}\gamma_t(\ell_{t,k}^c+C_t)+\sum_{t=1}^{T}\langle\widetilde{\ell}_t,\mathbf{q}_t-\mathbf{e}_{k_t}\rangle,
\end{aligned}
$$

where $\widetilde{\ell}_t$ denotes the implicit loss vector such that $\widetilde{\ell}_{t,k}=\frac{q_{t,k}}{q_{t,k}+\gamma_t}(\ell_{t,k}^c+C_t)$ for every $k\in[n]$. Notice that $\|\widetilde{\ell}_t\|_\infty\le 3\ell_\infty$. Thus with probability at least $1-\delta$,

$$
\begin{aligned}
\sum_{t=1}^{T}\sum_{k=1}^{n}\gamma_t(\ell_{t,k}^c+C_t)+\sum_{t=1}^{T}\langle\widetilde{\ell}_t,\mathbf{q}_t-\mathbf{e}_{k_t}\rangle &\le 6\ell_\infty\sqrt{nT\log(n)}+3\ell_\infty\sqrt{2T\log(2/\delta)}\\
&\le\Theta\left(\ell_\infty\sqrt{nT\log(n)}+\ell_\infty\sqrt{T\log(1/\delta)}\right).
\end{aligned}
$$

For the last term,

$$\begin{aligned}\sum_{t=1}^{T}\langle\hat{\ell}_t-(\ell_t^c+C_t\mathbf{1}_n),\mathbf{p}^\dagger\rangle &\leq \frac{1}{2\gamma_T}\sum_{t=1}^{T}2\gamma_t\langle\hat{\ell}_t-(\ell_t^c+C_t\mathbf{1}_n),\mathbf{p}^\dagger\rangle\\ &\leq \frac{3}{2\gamma_T}\ell_\infty\log(1/\delta)\\ &\leq \Theta\left(\ell_\infty\sqrt{\frac{nT}{\log n}}\log(1/\delta)\right).\end{aligned}$$

Summing up the above we can bound $\textcircled{1}$ by $\Theta\left(\ell_\infty\sqrt{\frac{nT}{\log n}}\log(1/\delta)+\ell_\infty\sqrt{nT\log(n)}\right)$.

A.6. Proof of Lemma 13

Remind

$$\mathbb{P}\left\{\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i\neq\mathbf{0}_n\}\mathbb{1}\{C_t<2^i\}>\frac{n}{\beta_t}\log(1/\delta)\right\}\leq\delta$$

for every $i\leq K$. We first note that

$$\sum_{i=-\infty}^{K}2^i\left(\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i\neq\mathbf{0}_n\}\mathbb{1}\{C_t<2^i\}\right)\leq\sum_{i=-\infty}^{K}2^i\left(\sqrt{\frac{n^2+nT}{\log n}}\log(2^{K-i+2}/\delta)\right)$$

with probability at least $1-\delta$. This is because $\sum_{i=-\infty}^{K}\delta/2^{K-i+2}\leq\delta$. Denote by

$$S_i=2^i\left(\sqrt{\frac{n^2+nT}{\log n}}\log(2^{K-i+2}/\delta)\right),$$

we then prove $S_{i-1}/S_i\leq 3/4$ for every $i\leq K$.

$$\frac{S_{i-1}}{S_i}\leq\frac{1}{2}\frac{\log(2^{K-i+2}/\delta)+\log 2}{\log(2^{K-i+2}/\delta)}\leq\frac{3}{4}.$$

Thus,

$$\begin{aligned}\sum_{i=-\infty}^{K}2^i\left(\sum_{t=1}^{T}\mathbb{1}\{\ell_t^i\neq\mathbf{0}_n\}\mathbb{1}\{C_t<2^i\}\right) &\leq\sum_{i=-\infty}^{K}S_i\leq S_K\sum_{i=-\infty}^{K}(\frac{3}{4})^{K-i}=4S_K\\ &\leq 2^{K+2}\left(\sqrt{\frac{n^2+nT}{\log n}}\log(4/\delta)\right)\\ &\leq 8\ell_\infty\sqrt{\frac{n^2+nT}{\log n}}\log(4/\delta)\\ &=\Theta\left(\ell_\infty\sqrt{\frac{n^2+nT}{\log n}}\log(1/\delta)\right)\end{aligned}$$

A.7. Proof of Lemma 14

Notice that

$$\sum_{t=1}^{T}\sum_{k=1}^{n}\alpha_{t,k}\left(\hat{\ell}_{t,k}-(\ell^c_{t,k}+C_t)\right)=3\ell_\infty\sum_{t=1}^{T}\sum_{k=1}^{n}\alpha_{t,k}\left(\frac{\hat{\ell}_{t,k}}{3\ell_\infty}-\frac{\ell^c_{t,k}+C_t}{3\ell_\infty}\right).$$

Since $0\le\ell^c_{t,k}+C_t\le 3\ell_\infty$ for all $t\in[T]$, by Lemma 1 in Neu (2015), we complete the proof.

Appendix B. Omitted details for Section 4

B.1. Omitted details of Occupancy measure

In this subsection, we briefly explain the concept of "occupancy measure" and show how to reformulate adversarial MDP problems to adversarial MAB problems (for more details see Jin et al. (2019) and Lee et al. (2020)). For any $(s,a)\in[S]\times[A]$, the probability that policy π visits the state-action pair (s,a) with transition function P can be denoted by

$$q^{P,\pi}(s,a)=\mathbb{P}\left\{s_{h(s)}=s,a_{h(s)}=a|P,\pi\right\},$$

where $h(s)$ denotes the index of the layer to which state s belongs. Here, $q^{P,\pi}\in R^{S\times A}$ is a valid occupancy measure. Following Jin et al. (2019), we denote $\Delta(P)$ by the set of occupancy measures whose induced transition function is P, i.e., the set of $q^{P,\pi}$ for all policy π with transition function P, and $\Delta(\mathcal{P})$ by the set of occupancy measures whose induced transition function belongs to the set of transition functions $\mathcal{P}$, i.e., the set of $q^{P,\pi}$ for all policy π with transition function $P\in\mathcal{P}$. Assuming P is the underlying transition function, the total expected regret (w.r.t. randomness of the transition function) can be written as

$$\begin{aligned}\mathcal{R}(T)&=\sum_{t=1}^{T}\ell_t(\pi_t)-\sum_{t=1}^{T}\ell_t(\pi^\star)\\&=\sum_{t=1}^{T}\sum_{s\in[S],a\in[A]}(q^{P,\pi_t}(s,a)-q^{P,\pi^\star}(s,a))\ell_t(s,a)\\&=\sum_{t=1}^{T}\langle q^{P,\pi_t}-q^{P,\pi^\star},\ell_t\rangle.\end{aligned}$$

When the regret is written in this way, it is clear that the adversarial MDP problems can be reduced to the adversarial MAB problems.

B.2. Omitted details of `UOB-REPS-EX`

In this section, we introduce the algorithm `UOB-REPS-EX`, as illustrated in Algorithm 5. The algorithm is mainly the same to `UOB-REPS` in Jin et al. (2019), except for the following three differences. First, `UOB-REPS-EX` uses the clipping loss with offset $\{\ell^c_t(s_h,a_h)+C_{t,h}\}_{h\in[H]}$ instead of $\{\ell_t(s_h,a_h)\}_{h\in[H]}$ as the input. Secondly, in each episode, Algorithm 5 applies FTRL with an

Algorithm 5: UOB-REPS-EX: Upper Occupancy Bound Relative Entropy Policy Search with Explicit Exploration

```
Initialize: state space $S$, action space $A$, episode number $T$, learning rate $\eta$, implicit
            exploration rate $\gamma$, explicit exploration rate $\beta$, confidence parameter $\delta$, Shannon
            Entropy $\{\Psi_h\}_{h\in[H]}$, and Comp-UOB as Algorithm 3 in Jin et al. (2019)
Initialize: epoch index $i = 1$, confidence set $\mathcal{P}_1$ as the set of all transition functions, counters
            $N_0(s,a) = N_1(s,a) = M_0(s'|s,a) = M_1(s'|s,a) = 0,\ \forall(s,a)$, occupancy
            measure $\hat{q}_1(s,a,s') = \frac{1}{|S_h||A||S_{h+1}|},\ \forall(s,a,s')$ and corresponding policy $\pi_1 = \pi^{\hat{q}_1}$
Input: State exploration policies $\pi^s$, $\forall s$, (streaming) trajectories $\{s_h, a_h, \ell_t^+(s_h,a_h)\}_{h\in[H]}$
       and clipping threshold $\{C_{t,h}\}_{h\in[H]}$ for $t \in [T]$
Output: (Streaming) policies $\pi_t$ for $t \in [T]$
for $t = 1$ to $T$ do
    Send policy $\pi_t$ to SCB-RL. Receive trajectory $\{s_h, a_h, \ell_t^c(s_h,a_h) + C_{t,h}\}_{h\in[H]}$ and
     clipping threshold $\{C_{t,h}\}_{h\in[H]}$
    Compute upper occupancy bound: $u_t(s_h,a_h) = \text{Comp-UOB}(\pi_t, s_h, a_h, \mathcal{P}_i),\ \forall h \in [H]$
    Construct loss estimators:
        $$\hat{\ell}_t(s,a) = \frac{\ell_t^c(s,a) + C_{t,h(s)}}{u_t(s,a)+\gamma}\mathbb{1}\{s_{h(s)} = s, a_{h(s)} = a\},\ \forall(s,a) \in [S]\times[A]$$
    Update counters:
    $N_i(s_h,a_h) \leftarrow N_i(s_h,a_h) + 1,\ M_i(s_{h+1}|s_h,a_h) \leftarrow M_i(s_{h+1}|s_h,a_h) + 1,\ \forall h \in [H]$
    if $\exists h \in [H],\ N_i(s_h,a_h) \geq \max\{1, 2N_{i-1}(s_h,a_h)\}$ then
        Increase epoch index $i \leftarrow i+1$
        Initialize new counters: $N_i \leftarrow N_{i-1},\ M_i \leftarrow M_{i-1}$
        Update confidence set $\mathcal{P}_i$
            $$\mathcal{P}_i = \Big\{\hat{P} : |\hat{P}(s'|s,a) - \bar{P}_i(s'|s,a)| \leq \epsilon_i(s'|s,a), \forall(s,a,s') \in [S_h]\times[A]\times[S_{h+1}], h = 0,\dots,H-1\Big\},$$
         where $\bar{P}_i(s'|s,a) = \frac{M_i(s'|s,a)}{\max\{1,N_i(s,a)\}}$ and
            $$\epsilon_i(s'|s,a) = 4\sqrt{\frac{\bar{P}(s'|s,a)\ln\left(\frac{TSA}{\delta}\right)}{\max\{1,N_i(s,a)-1\}}} + \frac{28\ln\left(\frac{TSA}{\delta}\right)}{3\max\{1,N_i(s,a)-1\}}$$
    end
    Update occupancy measure and policy, get $\widetilde{q}_{t+1}$ and set $\widetilde{\pi}_{t+1} \leftarrow \pi^{\widetilde{q}_{t+1}}$
        $$\widetilde{q}_{t+1} = \arg\min_{q\in\Delta(\mathcal{P}_i)}\left(\sum_t \langle q, \hat{\ell}_t\rangle + \sum_h \frac{C_{t,h}}{\eta}\Psi_h(q)\right)$$
    Add extra exploration: $\pi_{t+1} = (1-\beta)\widetilde{\pi}_{t+1} + \beta\text{Uniform}(\pi^1,\dots,\pi^S)$
end
```

adaptive learning rate to update the occupancy measure $\widetilde{q}_{t+1}$, instead of using OMD with a fixed learning rate. Recall the definition of (negative) Shannon Entropy on an occupancy measure q is

$$\Psi_h(q) = \sum_{s\in[S_h],a\in[A]} q(s,a)\ln\frac{1}{q(s,a)},\ \forall h\in[H].$$

Lastly, the policy output by `UOB-REPS-EX` is a mixture of its FTRL output policy $\widetilde{\pi}_{t+1}$ and the exploration policies from `RF-ELP`. This step is to allow every state-action pair to have a probability of being visited, so that `SCB-RL` can perceive the change of loss scale and update the clipping threshold on time.

B.3. Proof of Lemma 8

We first note that

$$\mathbb{E}\left[\sum_h r(s_h,a_h)|P,\pi\right] = \mathbb{E}\left[\sum_h \mathbb{1}\{s_h = s\}|P,\pi\right] = q^{P,\pi}(s).$$

Denoted by $q^s = \max_{\pi\in\Pi} q^{P,\pi}(s)$, using the above, Lemma 7 implies that

$$q^s - q^{P,\pi^{s,N}}(s) \le \tilde{\mathcal{O}}\left(\sqrt{\frac{SA\mathrm{Var}^s}{N}} + \frac{SAH}{N}\right) \le \tilde{\mathcal{O}}\left(\sqrt{\frac{SAq^s}{N}} + \frac{SAH}{N}\right)$$

holds with probability at least $1-\delta$, where the last inequality is due to $\mathrm{Var}^s \le q^s$. By the appendix F.3.4 in Zhang et al. (2023), we can set $C = \mathcal{O}\left(\log^4(T)\log^2(SAH)\log(\frac{1}{\delta})\right)$ such that $q^s - q^{P,\pi^{s,N}}(s) \le C\left(\sqrt{\frac{SAq^s}{N}} + \frac{SAH}{N}\right)$. When $q^s > 9C^2\frac{SAH}{N}$, we note that

$$q^s > 2C\left(\sqrt{\frac{SAq^s}{N}} + \frac{SAH}{N}\right)$$

and thus

$$q^{P,\pi^{s,N}}(s) \ge q^s - C\left(\sqrt{\frac{SAq^s}{N}} + \frac{SAH}{N}\right) \ge \frac{q^s}{2}.$$

This completes the proof.

B.4. Proof of Lemma 9

We start by stating two key technical lemmas from Jin et al. (2019). The first outlines the reliability of the confidence sets. The second essentially describes how the confidence set shrinks over time.

Lemma 15 *(Lemma 2 in Jin et al. (2019)) With probability at least $1-4\delta$, there is $P\in\mathcal{P}_i$ for all i.*

Lemma 16 *(Lemma 4 in Jin et al. (2019)) With probability at least $1-\delta$, for any $h \in [H]$ and any collection of transition functions $\{P_t^s\}_{s\in[S]}$ such that $P_t^s \in \mathcal{P}_{i_t}$ for all $s \in [S]$, there is*

$$\sum_{t=1}^{T} \sum_{s\in[S_h],a\in[A]} |q^{P_t^s,\pi_t}(s,a) - q^{P,\pi_t}(s,a)| \leq \mathcal{O}\left(S\sqrt{AT\ln\left(\frac{TSA}{\delta}\right)}\right)$$

Recall $q_t = q^{P,\pi_t}$ and $\hat{q}_t = q^{\hat{P}_t,\pi_t}$. Given the above lemma, we decompose the regret into

$$\sum_{t=1}^{T}\langle \ell_t^+, q_t - q^*\rangle = \overbrace{\sum_{t=1}^{T}\langle \ell_t^+, q_t - \hat{q}_t\rangle}^{\textsc{Error}} + \overbrace{\sum_{t=1}^{T}\langle \ell_t^+ - \hat{\ell}_t, \hat{q}_t\rangle}^{\textsc{Bias}_1} + \overbrace{\sum_{t=1}^{T}\langle \hat{\ell}_t, \hat{q}_t - q^\star\rangle}^{\textsc{Reg}} + \overbrace{\sum_{t=1}^{T}\langle \hat{\ell}_t - \ell_t^+, q^*\rangle}^{\textsc{Bias}_2}$$

Bounding ERROR : By Lemma 16, we immediately obtain the following bound.

Lemma 17 *With probability at least $1-\delta$, there is*

$$\textsc{Error} \leq \mathcal{O}\left(\sum_{h\in[H]} \ell_{\infty,h} S\sqrt{AT\ln\left(\frac{TSA}{\delta}\right)}\right)$$

Bounding BIAS$_1$: The high level idea of bounding BIAS$_1$ is to show that $\hat{\ell}_t$ is not underestimating ℓ_t^+ by too much, which is ensured due to the fact that the confidence set becomes more and more accurate for frequently visited state-action pairs.

Lemma 18 *With probability at least $1-\delta$, there is*

$$\textsc{Bias}_1 \leq \mathcal{O}\left(\sum_{h\in[H]} \ell_{\infty,h} S\sqrt{AT\ln\left(\frac{SAT}{\delta}\right)} + \sum_{h\in[H]} \gamma \ell_{\infty,h} S_h AT\right)$$

Bounding REG : In this part, we build the proof based on the ideas of Neu (2015) and Jin et al. (2019). The main challenge is that our loss estimator $\hat{\ell}_t$ corresponds to the policy π_t rather than the FTRL output $\widetilde{\pi}_t$, which makes some regular proof tricks no longer applicable.

Lemma 19 *With probability at least $1-\delta$, there is*

$$\textsc{Reg} \leq \mathcal{O}\left(\frac{\ln(SA)}{\eta}\sum_{h\in[H]}\ell_{\infty,h} + \frac{\eta}{1-\beta}AT\sum_{h\in[H]}\ell_{\infty,h}S_h + \frac{\ln\left(\frac{H}{\delta}\right)}{\gamma}\sum_{h\in[H]}\ell_{\infty,h} + \beta T\sum_{h\in[H]}\ell_{\infty,h}\right).$$

Bounding BIAS$_2$: BIAS$_2$ can be bounded via a direct application of Lemma 21.

Lemma 20 *With probability at least $1-\delta$, there is*

$$\textsc{Bias}_2 \leq \mathcal{O}\left(\frac{1}{\gamma}\sum_{h\in[H]}\ell_{\infty,h}\ln\left(\frac{H}{\delta}\right)\right)$$

Summing up the above and setting $\eta = \gamma = \mathcal{O}\left(\sqrt{\frac{H\ln(SAT/\delta)}{SAT}}\right)$ and $\beta \geq 1/2$, we get

$$\sum_{t=1}^{T}\langle \ell_t^+, q_t - q^*\rangle \leq \mathcal{O}\left(\sum_{h\in[H]}\ell_{\infty,h}S\sqrt{AT\ln\left(\frac{SAT}{\delta}\right)} + \beta T\sum_{h\in[H]}\ell_{\infty,h}\right)$$

with probability $1-\delta$.

B.5. Proof of Lemma 10

Recall $q^s = \max_{\pi \in \Pi} q^{P,\pi}(s)$ and $\ell'_t = |\ell_t - \ell^c_t|$. By the clipping rule, there is $\ell'(s,a) \le |\ell_t(s,a)| \mathbb{1}\{C_{t,h} < 2^i\}$ for every (s,a) pair. Fix $h \in [H]$, it suffices to prove

$$\sum_{s \in [S_h], a \in [A]} \sum_{t=1}^{T} |\ell_t(s,a)||q_t(s,a) - q^*(s,a)| \mathbb{1}\left\{q^s > \tilde{\mathcal{O}}(SAH/\xi T)\right\} \mathbb{1}\left\{C_{t,h} < 2^i\right\} \le \tilde{\mathcal{O}}\left(\frac{\ell_{\infty,h} SA}{\beta}\right)$$

Define $K := \arg\min_{j \in \mathbb{N}} \left\{\ell_{\infty,h} \le 2^j\right\}$. Define $\ell^i_t(s,a) = \ell_t(s,a)\mathbb{1}\{2^{i-1} < |\ell_t(s,a)| \le 2^i\}\mathbb{1}\{q^s > \tilde{\mathcal{O}}(SAH/\xi T)\}$. We note that

$$\begin{aligned}
&\sum_{s \in [S_h], a \in [A]} \sum_{t=1}^{T} |\ell_t(s,a)||q_t(s,a) - q^*(s,a)| \mathbb{1}\left\{q^s > \tilde{\mathcal{O}}(SAH/\xi T)\right\} \mathbb{1}\left\{C_{t,h} < 2^i\right\} \\
\le &\sum_{i=-\infty}^{K} \sum_{s \in [S_h], a \in [A]} \sum_{t=1}^{T} |\ell^i_t(s,a)||q_t(s,a) - q^*(s,a)| \mathbb{1}\left\{C_{t,h} < 2^i\right\} \\
\le &\sum_{i=-\infty}^{K} \sum_{s \in [S_h], a \in [A]} \sum_{t=1}^{T} |\ell^i_t(s,a)| q_s \mathbb{1}\left\{C_{t,h} < 2^i\right\} \\
\le &\sum_{i=-\infty}^{K} 2^i \sum_{t=1}^{T} \mathbb{1}\left\{C_{t,h} < 2^i\right\} \sum_{s \in [S_h], a \in [A]} q_s \mathbb{1}\left\{\ell^i_t(s,a) \ne 0\right\}.
\end{aligned}$$

For brevity, we denote by $X^i_t = \sum_{s \in [S_h], a \in [A]} q_s \mathbb{1}\{\ell^i_t(s,a) \ne 0\}$. By Lemma 8, for any $s \in [S_h]$, if $q_s \mathbb{1}\left\{\ell^i_t(s,a) \ne 0\right\} \ne 0$, state s can be well explored by `RF-ELP` with high probability. As shown in Algorithm 5, the exploration policy has probability at least $\tilde{\mathcal{O}}(\beta/SA)$ to be played in episode t for every state. Thus, the algorithm is able to observe the outlier and update $C_{t+1,h}$ to 2^i with probability at least $\tilde{\mathcal{O}}(\beta X^i_t/SA)$. Thus, for every integer $m \ge 1$, we have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{t=1}^{T} \mathbb{1}\left\{C_{t,h} < 2^i\right\} X^i_t \ge \sum_{t=1}^{m} X^i_t\right\} &\le \prod_{t=1}^{m} \left(1 - \tilde{\mathcal{O}}(\beta X^i_t/SA)\right) \\
&\le \tilde{\mathcal{O}}\left(\left(1 - \frac{\beta}{SA}\right)^{\sum_{t=1}^{m} X^i_t}\right).
\end{aligned}$$

This implies that with probability at least $1-\delta$, there is

$$\sum_{t=1}^{T} \mathbb{1}\left\{C_{t,h} < 2^i\right\} \sum_{s \in [S_h], a \in [A]} q_s \mathbb{1}\left\{\ell^i_t(s,a) \ne 0\right\} \le \tilde{\mathcal{O}}\left(\frac{SA}{\beta}\right).$$

Then, using the same idea of Lemma 13, we can achieve a high probability union bound on all $i \leq K$, i.e., with probability at least $1-\delta$

$$\begin{aligned}
&\sum_{i=-\infty}^{K} 2^i \sum_{t=1}^{T} \mathbb{1}\left\{C_{t,h} < 2^i\right\} \sum_{s\in[S_h],a\in[A]} q_s \mathbb{1}\left\{\ell_t^i(s,a) \neq 0\right\} \\
\leq &\sum_{i=-\infty}^{K} 2^i \tilde{\mathcal{O}}\left(\frac{SA}{\beta}\right) \leq 2^{K+1}\tilde{\mathcal{O}}\left(\frac{SA}{\beta}\right) = \tilde{\mathcal{O}}\left(\frac{\ell_{\infty,h}SA}{\beta}\right),
\end{aligned}$$

which completes the proof.

B.6. Proof of Lemma 17

$$\begin{aligned}
\sum_{t=1}^{T}\langle \ell_t^+, q_t - \hat{q}_t\rangle &\leq \sum_{h\in[H]}\left(\max_{s\in[S_h],a\in[A]} \ell_t^+(s,a)\left(\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]} |q^{\hat{P}_t,\pi_t}(s,a) - q_t(s,a)|\right)\right) \\
&\leq \sum_{h\in[H]}\left(\max_{s\in[S_h],a\in[A]} \ell_t^+(s,a)\left(\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]} |q^{P_t^s,\pi_t}(s,a) - q_t(s,a)|\right)\right) \\
&\leq \mathcal{O}\left(\sum_{h\in[H]} C_{T,h} S\sqrt{AT\ln\left(\frac{TSA}{\delta}\right)}\right) \\
&\leq \mathcal{O}\left(\sum_{h\in[H]} \ell_{\infty,h} S\sqrt{AT\ln\left(\frac{TSA}{\delta}\right)}\right)
\end{aligned}$$

The second inequality is by setting $\hat{P}_t = P_t^s \in \mathcal{P}_{i_t}$ as in Lemma 16. The third and last inequalities are due to $\ell_t^+(s,a) \leq 2C_{t,h(s)} \leq 2C_{T,h(s)}$ and $C_{T,h} \leq 2\max_{s\in[S_h],a\in[A]} \ell_t^+(s,a) = 4\ell_{\infty,h}$ for all $h \in [H]$.

B.7. Proof of Lemma 18

We first note that

$$\mathbb{E}\left[\hat{\ell}_t(s,a)\right] = \frac{\ell_t^c(s,a) + C_{t,h(s)}}{u_t(s,a)+\gamma} q_t(s,a) = \frac{q_t(s,a)}{u_t(s,a)+\gamma}\ell_t^+(s,a).$$

Then

$$\begin{aligned}
\sum_{t=1}^{T}\langle \ell_t^+ - \hat{\ell}_t, \hat{q}_t\rangle &= \sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]} \hat{q}_t(s,a)\left(\ell_t^+(s,a) - \hat{\ell}_t(s,a)\right) \\
&= \sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]} \hat{q}_t(s,a)\left(\ell_t^+(s,a) - \mathbb{E}\left[\hat{\ell}_t(s,a)\right]\right) \\
&+ \sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]} \hat{q}_t(s,a)\left(\mathbb{E}\left[\hat{\ell}_t(s,a)\right] - \hat{\ell}_t(s,a)\right)
\end{aligned}$$

For the first term, there is

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]}\hat{q}_t(s,a)\left(\ell_t^+(s,a)-\mathbb{E}\left[\hat{\ell}_t(s,a)\right]\right)\\
=&\sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]}\hat{q}_t(s,a)\ell_t^+(s,a)\left(1-\frac{q_t(s,a)}{u_t(s,a)+\gamma}\right)\\
\leq&3\sum_{h\in[H]}\ell_{\infty,h}\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]}\frac{\hat{q}_t(s,a)}{u_t(s,a)+\gamma}\left(u_t(s,a)+\gamma-q_t(s,a)\right)\\
\leq&3\sum_{h\in[H]}\ell_{\infty,h}\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]}|u_t(s,a)-q_t(s,a)|+3\sum_{h\in[H]}\gamma\ell_{\infty,h}S_hAT.
\end{aligned}$$

Recall

$$u_t(s,a)=\pi_t(a|s)\max_{\hat{P}\in\mathcal{P}_{i_t}}q^{\hat{P},\pi_t}(s),$$

the last inequality is due to $\hat{q}_t(s,a)\leq u_t(s,a)$. Moreover, since $q^{P_t^x,\pi_t}(s,a)=\pi_t(a|s)q^{P_t^x,\pi_t}(s)\leq u_t(s,a)$ for all $(s,a)\in[S]\times[A]$, it suffices to bound $\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]}|u_t(s,a)-q_t(s,a)|$ by Lemma 16. Thus we can conclude

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]}\hat{q}_t(s,a)\left(\ell_t^+(s,a)-\mathbb{E}\left[\hat{\ell}_t(s,a)\right]\right)\\
\leq&\mathcal{O}\left(\sum_{h\in[H]}\ell_{\infty,h}S\sqrt{AT\ln\left(\frac{SAT}{\delta}\right)}+\sum_{h\in[H]}\gamma\ell_{\infty,h}S_hAT\right).
\end{aligned}$$

For the second term, notice that $|\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]}\hat{q}_t(s,a)\hat{\ell}_t(s,a)|\leq 3\sum_{h\in[H]}\ell_{\infty,h}$ for all $t\in[T]$. Using Azuma's inequality, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T}\sum_{h\in[H]}\sum_{s\in[S_h],a\in[A]}\hat{q}_t(s,a)\left(\mathbb{E}\left[\hat{\ell}_t(s,a)\right]-\hat{\ell}_t(s,a)\right)\leq\mathcal{O}\left(\sum_{h\in[H]}\ell_{\infty,h}\sqrt{T\ln\frac{1}{\delta}}\right).$$

Summing up the two terms and resize δ completes the proof.

B.8. Proof of Lemma 19

We start by decomposing REG into

$$\begin{aligned}
\textsc{Reg}&=(1-\beta)\sum_{t=1}^{T}\langle\hat{\ell}_t,\widetilde{q}_t-q^\star\rangle+\beta\sum_{t=1}^{T}\langle\hat{\ell}_t,q^{\hat{P}_t,\text{Uniform}(\pi^1,\dots,\pi^S)}-q^\star\rangle\\
&\leq(1-\beta)\sum_{t=1}^{T}\langle\hat{\ell}_t,\widetilde{q}_t-q^\star\rangle+\beta\sum_{t=1}^{T}\langle\hat{\ell}_t-\ell_t^+,q^{\hat{P}_t,\text{Uniform}(\pi^1,\dots,\pi^S)}\rangle+\beta\sum_{t=1}^{T}\langle\ell_t^+,q^{\hat{P}_t,\text{Uniform}(\pi^1,\dots,\pi^S)}\rangle\\
&\leq(1-\beta)\sum_{t=1}^{T}\langle\hat{\ell}_t,\widetilde{q}_t-q^\star\rangle+\beta\sum_{t=1}^{T}\langle\hat{\ell}_t-\ell_t^+,q^{\hat{P}_t,\text{Uniform}(\pi^1,\dots,\pi^S)}\rangle+3\beta T\sum_{h\in[H]}\ell_{\infty,h}
\end{aligned}$$

To bound the first and second term, we propose a variant of Lemma 11 in Jin et al. (2019).

Lemma 21 *For any sequence of functions $\alpha_1, \dots, \alpha_T$ such that $\alpha_t \in [0, 2\gamma]^{S\times A}$ and $\mathcal{F}_t$-measurable for all $t \in [T]$, with probability at least $1-\delta$, there is*

$$\sum_{t=1}^{T} \sum_{s\in[S],a\in[A]} \alpha_t(s,a) \left(\hat{\ell}_t(s,a) - \frac{q_t(s,a)}{u_t(s,a)} \ell_t^+(s,a) \right) \le \mathcal{O}\left(\sum_{h\in[H]} \ell_{\infty,h} \ln\left(\frac{H}{\delta}\right)\right).$$

Without loss of generality, in the following, we assume that $u_t(s,a) \ge q_t(s,a)$ for all $(s,a) \in [S]\times[A]$, which holds true with probability at least $1-\delta$. Using Lemma 21, since $q^{\hat{P}_t,\mathrm{Uniform}(\pi^1,\dots,\pi^S)}$ is independent to ℓ_t and thus be $\mathcal{F}_t$-measurable, we can immediately bound the second term by

$$\begin{aligned}\beta \sum_{t=1}^{T} \langle \hat{\ell}_t - \ell_t^+, q^{\hat{P}_t,\mathrm{Uniform}(\pi^1,\dots,\pi^S)} \rangle &= \frac{\beta}{2\gamma} \sum_{t=1}^{T} \langle \hat{\ell}_t - \ell_t^+, 2\gamma q^{\hat{P}_t,\mathrm{Uniform}(\pi^1,\dots,\pi^S)} \rangle \\ &\le \frac{\beta}{2\gamma} \sum_{h\in[H]} \ell_{\infty,h} \ln\left(\frac{H}{\delta}\right)\end{aligned}$$

with probability at least $1-\delta$. It suffices to focus on the first term. By standard analysis of FTRL algorithm, there is

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, \widetilde{q}_t - q^\star \rangle \le \sum_{h\in[H]} \frac{C_{T,h}}{\eta} \ln(SA) + \sum_{t=1}^{T} \sum_{h\in[H]} \frac{\eta}{C_{t,h}} \sum_{s\in[S_h],a\in[A]} \widetilde{q}_t(s,a) \hat{\ell}_t^2(s,a).$$

We further note that

$$\widetilde{q}_t(s,a)\hat{\ell}_t^2(s,a) \le \widetilde{q}_t(s,a) \frac{\ell_t^+(s,a)}{u_t(s,a)+\gamma} \hat{\ell}_t(s,a) \le 2C_{t,h} \frac{\widetilde{q}_t(s,a)}{u_t(s,a)+\gamma} \hat{\ell}_t(s,a).$$

Different to the proof of Jin et al. (2019), we cannot immediately conclude $\widetilde{q}_t(s,a)/(u_t(s,a)+\gamma) \le 1$. This is because $u_t(s,a)$ is the upper bound of $q^{\hat{P}_t,\pi_t}(s,a)$ rather than $q^{\hat{P}_t,\widetilde{\pi}_t}(s,a)$. Here we prove by showing

$$u_t(s,a) \ge q^{\hat{P}_t,\pi_t}(s,a) \ge (1-\beta) q^{\hat{P}_t,\widetilde{\pi}_t}(s,a),$$

which implies that $\widetilde{q}_t(s,a)/(u_t(s,a)+\gamma) \le 1/(1-\beta)$. Thus we have

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, \widetilde{q}_t - q^\star \rangle \le \sum_{h\in[H]} \frac{C_{T,h}}{\eta} \ln(SA) + \frac{2\eta}{1-\beta} \sum_{t=1}^{T} \sum_{h\in[H]} \sum_{s\in[S_h],a\in[A]} \hat{\ell}_t(s,a).$$

Applying Lemma 21 obtains

$$\begin{aligned}&\sum_{t=1}^{T} \langle \hat{\ell}_t, \widetilde{q}_t - q^\star \rangle \\ &\le \sum_{h\in[H]} \frac{C_{T,h}}{\eta} \ln(SA) + \frac{2\eta}{1-\beta} \sum_{t=1}^{T} \sum_{h\in[H]} \sum_{s\in[S_h],a\in[A]} \frac{q_t(s,a)}{u_t(s,a)} \ell_t(s,a) + \sum_{h\in[H]} \frac{\ell_{\infty,h}}{2\gamma} \ln\left(\frac{H}{\delta}\right) \\ &\le \sum_{h\in[H]} \frac{2\ell_{\infty,h}}{\eta} \ln(SA) + \frac{2\eta}{1-\beta} \sum_{h\in[H]} \ell_{\infty,h} S_h A T + \sum_{h\in[H]} \frac{\ell_{\infty,h}}{2\gamma} \ln\left(\frac{H}{\delta}\right)\end{aligned}$$

with probability at least $1-\delta$. Combining with the above and resize δ we finally get

$$\textsc{Reg} \le \mathcal{O}\left(\frac{\ln(SA)}{\eta}\sum_{h\in[H]}\ell_{\infty,h} + \frac{\eta}{1-\beta}AT\sum_{h\in[H]}\ell_{\infty,h}S_h + \frac{\ln\left(\frac{H}{\delta}\right)}{\gamma}\sum_{h\in[H]}\ell_{\infty,h} + \beta T\sum_{h\in[H]}\ell_{\infty,h}\right)$$

completing the proof.

B.9. Proof of Lemma 20

Assuming $u_t(s,a) \ge q_t(s,a)$ for all $(s,a)\in[S]\times[A]$ be true. Using Lemma 21, we have

$$\begin{aligned}\sum_{t=1}^{T}\langle\hat{\ell}_t-\ell_t^+, q^*\rangle &\le \sum_{t=1}^{T}\sum_{s\in[S],a\in[A]} q^*(s,a)\left(\frac{q_t(s,a)}{u_t(s,a)}\ell_t^+(s,a)-\ell_t^+(s,a)\right)+\frac{1}{2\gamma}\sum_{h\in[H]}\ell_{\infty,h}\ln\left(\frac{H}{\delta}\right)\\ &\le \frac{1}{2\gamma}\sum_{h\in[H]}\ell_{\infty,h}\ln\left(\frac{H}{\delta}\right)\end{aligned}$$

with probability at least $1-\delta$. Resize δ completes the proof.

B.10. Proof of Lemma 21

We note that

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{s\in[S],a\in[A]}\alpha_t(s,a)\left(\hat{\ell}_t(s,a)-\frac{q_t(s,a)}{u_t(s,a)}\ell_t^+(s,a)\right)\\ &=\sum_{h\in[H]}3\ell_{\infty,h}\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]}\alpha_t(s,a)\left(\frac{\hat{\ell}_t(s,a)}{3\ell_{\infty,h}}-\frac{q_t(s,a)}{u_t(s,a)}\frac{\ell_t^+(s,a)}{3\ell_{\infty,h}}\right).\end{aligned}$$

Now $\ell_t^+(s,a)/3\ell_{\infty,h}$ is within $[0,1]$ for all $t\in[T]$ and $(s,a)\in[S]\times[A]$. By the results of Lemma 11 in Jin et al. (2019), there is

$$\sum_{t=1}^{T}\sum_{s\in[S_h],a\in[A]}\alpha_t(s,a)\left(\frac{\hat{\ell}_t(s,a)}{3\ell_{\infty,h}}-\frac{q_t(s,a)}{u_t(s,a)}\frac{\ell_t^+(s,a)}{3\ell_{\infty,h}}\right)\le\ln\left(\frac{H}{\delta}\right)$$

for all $h\in[H]$ with probability at least $1-\delta$. Thus we have

$$\sum_{t=1}^{T}\sum_{s\in[S],a\in[A]}\alpha_t(s,a)\left(\hat{\ell}_t(s,a)-\frac{q_t(s,a)}{u_t(s,a)}\ell_t^+(s,a)\right)\le\mathcal{O}\left(\sum_{h\in[H]}\ell_{\infty,h}\ln\left(\frac{H}{\delta}\right)\right),$$

which completes the proof.

B.11. Omitted details of Remark 6

In this section, we describe how to reduce the regret of `SCB-RL` to $\tilde{\mathcal{O}}(\sum_{h\in[H]}\ell_{\infty,h}S\sqrt{AT})$. Recall `RF-ELP`, we fix the number of episodes used to find an exploration policy for state s as $\mathcal{O}(\sqrt{SAT})$. This is actually not necessary, that is, if the exploration algorithm has already found a good exploration policy for state s, it should stop searching and take the policy as output. In this case, the number of episodes used to find an exploration policy will be independent of T. Inspired by this, we design `RF-ELP-ES`, as illustrated in Algorithm 6. We will elucidate the details of the algorithm in the following section.

By Lemma 7 and Lemma F.3.4 in Zhang et al. (2023), there exists $C = \mathcal{O}(\log^3(T)\log^2(SAH)\log(\frac{1}{\delta}))$ such that

$$q^s - q^{P,\pi^{s,N}}(s) \le C\left(\sqrt{\frac{SAq^s}{N}} + \frac{SAH}{N}\right)$$

for all $N \ge 1$ with probability at least $1-\delta$. Taking the above as a quadratic function of $\sqrt{q^s}$, we have

$$\sqrt{q^s} \le \sqrt{q^{P,\pi^{s,N}}(s)} + 2C\sqrt{\frac{SAH}{N}},$$

which immediately implies

$$q^s - q^{P,\pi^{s,N}}(s) \le 3C^2\left(\sqrt{\frac{SAq^{P,\pi^{s,N}}(s)}{N}} + \frac{SAH}{N}\right). \tag{3}$$

Given $q^{P,\pi^{s,N}}(s) = \mathbb{E}[\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}]/N$, by empirical Bernstein's inequality, there exists $C' = \mathcal{O}(\log(\frac{T}{\delta}))$ such that with probability at least $1-\delta$, for all $N \ge 1$, there is

$$\left|q^{P,\pi^{s,N}}(s) - \frac{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}{N}\right| \le C'\left(\frac{\sqrt{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}}{N} + \frac{1}{N}\right). \tag{4}$$

Combining inequalities (3) and (4), it suffices to show that

$$q^s \le \frac{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}{N} + C''\left(\frac{\sqrt{SA\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}}{N} + \frac{SAH}{N}\right) \tag{5}$$

where $C'' = \mathcal{O}(\log^7(T)\log^4(SAH)\log^3(\frac{1}{\delta}))$. Furthermore, by inequality 4, we further have

$$q^{P,\pi^{s,N}}(s) \ge \frac{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}{N} - C'\left(\frac{\sqrt{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}}}{N} + \frac{1}{N}\right).$$

This means

$$\frac{q^s}{q^{P,\pi^{s,N}}(s)} \le \frac{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\} + C''\left(\sqrt{SA\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}} + SAH\right)}{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\} - C'\left(\sqrt{\sum_{t=1}^N \mathbb{1}_t\{s|P,\pi_t^s\}} + 1\right)}.$$

```
Algorithm 6: Reward free exploration in RL with Early Stopping (RF-ELP-ES)
Input: State $s$; Upper bound of exploration episodes number $N$
Output: Policy $\pi \in \Pi$; Exploration episodes number $N'$
Initialize reward: $r^s(s', a') \leftarrow \mathbb{1}\{s' = s\}$ for all $(s', a') \in [S] \times A$
Run MVP (Zhang et al., 2023) $N'$ episodes, where
  $N' = \min\{N, \arg\min_M \sum_{t=1}^{M} \mathbb{1}_t\{s|P, \pi_t^s\} \geq 9{C''}^2 SAH\}$, get policies
  $\{\pi_1^s, \dots, \pi_{N'}^s\} \leftarrow$ MVP$(r^s, N')$, set $\pi^{s,N'}(\cdot|s) \leftarrow$ Uniform$(\pi_1^s, \dots, \pi_{N'}^s)$
Set $\pi^{s,N'}(\cdot|s) \leftarrow$ Uniform$(A)$
Return $\pi^{s,N'}$, $N'$
```

When $\sum_{t=1}^{N} \mathbb{1}_t\{s|P, \pi_t^s\} \geq 9{C''}^2 SAH$, it suffices to say that $q^s / q^{P,\pi^{s,N}}(s) \leq 4$ with probability at least $1 - 2\delta$, that is, policy $\pi^{s,N}$ is good enough to explore state s, thus we can stop RF-ELP-ES in advance.

The last question is how large does N need to make $\sum_{t=1}^{N} \mathbb{1}_t\{s|P, \pi_t^s\} \geq 9{C''}^2 SAH$ with probability at least $1 - \delta$. Taking inequality (5) as a quadratic function of $\sqrt{\sum_{t=1}^{N} \mathbb{1}_t\{s|P, \pi_t^s\}}$, we note that

$$\sqrt{\sum_{t=1}^{N} \mathbb{1}_t\{s|P, \pi_t^s\}} \geq \frac{-C''\sqrt{SA} + \sqrt{{C''}^2 SA + 4Nq^s - 4SAH}}{2}.$$

RF-ELP-ES will end when

$$\frac{-C''\sqrt{SA} + \sqrt{{C''}^2 SA + 4Nq^s - 4SAH}}{2} \geq 3{C''}^2\sqrt{SAH}.$$

The above holds if $N \geq 16{C''}^2 SAH/q_s$. Therefore, RF-ELP-ES requires at most $\tilde{\mathcal{O}}(SAH/q_s)$ episodes to find an exploration policy for every state s.

Regret analysis Denote by

$$q_{\min} = \min_{s \in [S]} \{q^s\} = \min_{s \in [S]} \left\{ \max_{\pi \in \Pi} q^{P,\pi}(s) \right\}.$$

Consider $T \geq \tilde{\mathcal{O}}\left(\frac{SAH^2}{q_{\min}^2}\right)$ such that $\xi T \geq 16{C''}^2 SAH/q_s$ for all $s \in [S]$. In this case, every state is thoroughly explored, thus the term $\tilde{\mathcal{O}}\left(\frac{\sum_{h\in[H]} \ell_{\infty,h} S_h SHA}{\xi}\right)$ in the regret can be eliminated. Moreover, we can reduce the error incurred by the exploration phase from $\mathcal{O}\left(\sum_{h\in[H]} \ell_{\infty,h} \xi ST\right)$ to $\tilde{\mathcal{O}}\left(\sum_{h\in[H]} \ell_{\infty,h} \frac{S^2AH}{q_{\min}}\right)$ since RF-ELP-ES operates at most $\tilde{\mathcal{O}}\left(\frac{S^2AH}{q_{\min}}\right)$ episodes. Combining with other terms in the regret, we finally have

$$.\mathcal{R}(T) \leq \tilde{\mathcal{O}}\left(\sum_{h\in[H]} \ell_{\infty,h} \left[S\sqrt{AT} + \beta T + \frac{SA}{\beta} + \frac{S^2AH}{q_{\min}}\right]\right).$$

Setting $\beta = \mathcal{O}(\sqrt{SA/T})$ concludes the proof.