Near-Optimal Regret in Linear MDPs with Aggregate Bandit Feedback

Asaf Cassel [*] [1] **Haipeng Luo** [2] [3] **Aviv Rosenberg** [*] [4] **Dmitry Sotnikov** [3]

Abstract

In many real-world applications, it is hard to provide a reward signal in each step of a Reinforcement Learning (RL) process and more natural to give feedback when an episode ends. To this end, we study the recently proposed model of RL with Aggregate Bandit Feedback (RL-ABF), where the agent only observes the sum of rewards at the end of an episode instead of each reward individually. Prior work studied RL-ABF only in tabular settings, where the number of states is assumed to be small. In this paper, we extend ABF to linear function approximation and develop two efficient algorithms with near-optimal regret guarantees: a value-based optimistic algorithm built on a new randomization technique with a Q-functions ensemble, and a policy optimization algorithm that uses a novel hedging scheme over the ensemble.

1. Introduction

Reinforcement Learning (RL) has demonstrated remarkable empirical success in recent years (Mnih et al., 2015; Haarnoja et al., 2018; Ouyang et al., 2022), leading to increasing interest in understanding its theoretical guarantees. The standard model for RL is the Markov Decision Process (MDP) in which an agent interacts with an environment over multiple steps. In every step, the agent takes an action based on its observation of the current state and immediately gets reward feedback before transitioning to the next state. However, in many applications, it is either hard to provide accurate reward or the reward naturally arrives only at the end of an episode, e.g., in robotics (Jain et al., 2013) or Large Language Models (LLMs; Stiennon et al. (2020)).

To address this issue, several works (Efroni et al., 2021; Cohen et al., 2021; Chatterji et al., 2021) studied the Aggregate Bandit Feedback (ABF) model in which the agent only observes the sum of rewards at the end of an episode as feedback, instead of each reward individually. Yet, they all focused on tabular MDPs where the number of states is assumed to be finite and small. Although ABF is challenging even in the tabular setting, in all practical applications the state space is huge and RL algorithms use function approximation to approximate the value function efficiently.

In this paper, we tackle aggregate feedback in the presence of function approximation. Specifically, we consider ABF in Linear MDPs (Jin et al., 2020b), developing two efficient algorithms with near-optimal regret guarantees, i.e., $\tilde{O}(\text{poly}(dH)\sqrt{K})$ regret where K is the number of episodes, d is the dimension of the linear MDP, and H is the horizon. Our first algorithm, called Randomized Ensemble Least Squares Value Iteration (RE-LSVI), is an optimistic LSVI algorithm where the optimism is achieved via a new randomization technique. This algorithm is simple and gives our tightest regret bound. Our second algorithm, called Randomized Ensemble Policy Optimization (REPO), is based on policy optimization (PO) methods that have become extremely popular in recent years (Schulman et al., 2015; 2017) and especially in training LLMs (Ouyang et al., 2022). This is the first PO algorithm for aggregate feedback, which does not exist even in the tabular MDP setting. Theoretically, ABF is substantially more challenging under function approximation since it can no longer be formulated as a linear bandits problem (Efroni et al., 2021) (doing so would lead to a regret bound that scales with the number of states).

Our main technical contribution is the *ensemble randomization* technique. We use independently sampled Gaussian noise to compute multiple Q-functions from the same data, allowing for an optimistic estimate of the optimal value with controlled variance. In addition, we introduce a *loose truncation* mechanism to keep our estimates bounded without creating biases in the random walks induced by the added noise. While Efroni et al. (2021) already leveraged randomization to solve ABF in the tabular setting, they use a complex Thompson Sampling (TS) based analysis (Agrawal & Goyal, 2013). In contrast, our novel use of the ensemble unlocks a very simple regret analysis based on optimism. A concurrent work (Wu & Sun, 2023) uses randomization to solve Preference-based RL in linear MDPs, which is closely related to ABF (Chen et al., 2022). Applying their algo-

[*] Work partially done while at Amazon Science. [1] Blavatnik School of Computer Science, Tel Aviv University [2] University of Southern California [3] Amazon Science [4] Google Research. Correspondence to: Asaf Cassel <acassel@mail.tau.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

rithm in our setting again requires a complex TS analysis and yields regret with a worse dependence on d and H.

Another major contribution is a novel *hedging scheme* on top of the Q-functions ensemble, which enables us to obtain the first PO algorithm for ABF. It also demonstrates the effectiveness of our ensemble technique even for algorithms that are not based on optimism, as PO algorithms are notoriously difficult to analyze, albeit highly successful in practice. In fact, only very recently it was proven that PO achieves $\sqrt{K}$ regret in linear MDPs with standard reward feedback (Sherman et al., 2023). While our PO algorithm in linear MDPs involves a warm-up stage that uniformly explores the state space (similarly to Sherman et al. (2023)), in the tabular setting we develop an alternative analysis that avoids the need to perform this warm-up and may be of independent interest.

Related Work. There is a rich literature on regret minimization in RL, where the most popular algorithmic methods are optimism (achieved via exploration bonuses, e.g., Jaksch et al. (2010); Azar et al. (2017); Jin et al. (2020b)), policy optimization (e.g., Cai et al. (2020); Shani et al. (2020); Luo et al. (2021)), randomized exploration techniques (e.g., Osband et al. (2016); Xiong et al. (2022)) and global optimization methods based on either regularization (e.g., Zimin & Neu (2013); Rosenberg & Mansour (2019a;b); Jin et al. (2020a)) or optimism (e.g., Zanette et al. (2020)). They all rely on the standard assumption that each individual reward within an episode is revealed.

Efroni et al. (2021) introduced RL-ABF and gave the first regret bounds in tabular MDPs. Chatterji et al. (2021) then generalized aggregate feedback to a more realistic setting where only certain binary feedback is available, but their algorithm is efficient only under additional assumptions. Cohen et al. (2021) studied ABF in tabular adversarial environments, but their algorithm is built on a global mirror descent approach that cannot be extended to function approximation. Unlike all previous work (including this one) that focus on ABF in the context of online RL and exploration, Xu et al. (2022) recently studied ABF in offline RL which poses different challenges.

A closely related model is Preference-based RL (PbRL; Wirth et al. (2017)), where the agent receives feedback only in terms of preferences over a trajectory pair instead of absolute rewards. Saha et al. (2023); Chen et al. (2022) achieved sub-linear regret for PbRL, but their algorithms are computationally intractable even in tabular MDPs. As mentioned earlier, in a concurrent work, Wu & Sun (2023) utilize randomization to solve PbRL in linear MDPs efficiently. Additional works on PbRL study sample complexity and offline RL, and use the term RL from Human Feedback (RLHF; Wang et al. (2023); Zhan et al. (2023a;b)).

A different line of work studies delayed reward feedback in RL (Howson et al., 2023; Jin et al., 2022; Lancewicki et al., 2022; 2023). While they also deal with reward feedback that is not provided immediately after taking an action, they still observe each reward individually, thus avoiding the need to perform accurate credit assignment that our ABF model requires.

It is also worth mentioning that recently Tiapkin et al. (2023) utilized an ensemble of Q-functions for regret minimization in tabular MDPs with standard reward feedback. While we perturb the reward in a manner that is applicable to both value iteration and policy optimization, they perturb the learning rates in Q-learning.

2. Problem Setup

Markov Decision Process (MDP). A finite horizon MDP $\mathcal{M}$ is defined by a tuple $(\mathcal{X}, \mathcal{A}, x_1, r, P, H)$ with $\mathcal{X}$, a set of states, $\mathcal{A}$, a set of actions, H, decision horizon, $x_1 \in \mathcal{X}$, an initial state (which is assumed to be fixed for simplicity), $r = (r_h)_{h\in[H]}, r_h : \mathcal{X} \times \mathcal{A} \to [0, 1]$, a horizon dependent immediate reward function for taking action a at state x and horizon h, and $P = (P_h)_{h\in[H]}, P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, the transition probabilities. A single episode of an MDP is a sequence $(x_h, a_h, r_h)_{h\in[H]} \in (\mathcal{X} \times \mathcal{A} \times [0, 1])^H$ such that

$$\Pr[x_{h+1} = x' \mid x_h = x, a_h = a] = P_h(x' \mid x, a),$$

and $r_h \in [0, 1]$ is sampled such that $\mathbb{E}[r_h \mid x_h, a_h] = r_h(x_h, a_h)$. Notice the overloaded notation for r_h where $r_h(\cdot, \cdot)$ refers to the immediate (mean) reward function, and r_h (without inputs) refers to a sampled reward at horizon h.

Linear MDP. A linear MDP (Jin et al., 2020b) satisfies all the properties of the above MDP but has the following additional structural assumptions. There is a known feature mapping $\phi : \mathcal{X} \times A \to \mathbb{R}^d$ such that

$$r_h(x, a) = \phi(x, a)^\mathsf{T}\theta_h, \quad P_h(x' \mid x, a) = \phi(x, a)^\mathsf{T}\psi_h(x'),$$

where $\theta_h \in \mathbb{R}^d$, $\theta = (\theta_1^\mathsf{T}, \ldots, \theta_H^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{dH}$, and $\psi_h : \mathcal{X} \to \mathbb{R}^d$ are unknown parameters. We make the following normalization assumptions, common throughout the literature:

1. $\|\phi(x, a)\| \le 1$ for all $x \in X, a \in \mathcal{A}$;
2. $\|\theta_h\| \le \sqrt{d}$ for all $h \in [H]$;
3. $\||\psi_h|(\mathcal{X})\| = \|\sum_{x\in\mathcal{X}} |\psi_h(x)|\| \le \sqrt{d}$ for all $h \in [H]$;

where $|\psi_h(x)|$ is the entry-wise absolute value of $\psi_h(x) \in \mathbb{R}^d$. We follow the standard assumption in the literature that the action space $\mathcal{A}$ is finite. In addition, for ease of mathematical exposition, we also assume that the state space $\mathcal{X}$ is finite. This allows for simple matrix notation and avoids technical measure theoretic definitions. Importantly, our results are completely independent of the state space size $|\mathcal{X}|$, both computationally and in terms of regret. Thus, there is no particular loss of generality.

Policy and Value. A stochastic Markov policy $\pi = (\pi_h)_{h\in[H]} : [H] \times \mathcal{X} \mapsto \Delta(\mathcal{A})$ is a mapping from a step and a state to a distribution over actions. Such a policy induces a distribution over trajectories $\iota = (x_h, a_h)_{h\in[H]}$, i.e., sequences of H state-action pairs. For $f : (\mathcal{X} \times \mathcal{A})^H \to \mathbb{R}$, which maps trajectories to real values, we denote the expectation with respect to ι under dynamics P and policy π as $\mathbb{E}_{P,\pi}[f(\iota)]$. Similarly, we denote the probability under this distribution by $\mathbb{P}_{P,\pi}[\cdot]$. We denote the class of stochastic Markov policies as Π_M. For each policy π and horizon $h \in [H]$ we define its reward-to-go as $V_h^\pi(x) = \mathbb{E}_{P,\pi}[\sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \mid x_h = x]$, which is the expected reward if one starts from state x at horizon h and follows policy π onwards. The performance of a policy, also called its value, is measured by its expected cumulative reward, given by $V_1^\pi(x_1)$. When clear from context, we omit the 1 and simply write $V^\pi(x_1)$. The optimal policy is thus given by $\pi^\star \in \arg\max_{\pi\in\Pi_M} V^\pi(x_1)$, known to be optimal even among the class of stochastic history-dependent policies. Finally, we denote the optimal value as $V^\star = V_1^{\pi^\star}$.

Aggregate Feedback and Regret. We consider a standard episodic regret minimization setting where an algorithm performs K interactions with an MDP $\mathcal{M}$, but limit ourselves to *aggregate feedback* reward observations. Concretely, at the start of each interaction/episode $k \in [K]$, the agent specifies a stochastic Markov policy $\pi^k = (\pi_h^k)_{h\in[H]}$. Subsequently, it observes the trajectory ι^k sampled from the distribution $\mathbb{P}_{P,\pi^k}$, and the cumulative episode reward v^k (see Protocol 1). This is unlike standard bandit feedback where the individual rewards $(r_h^k)_{h\in[H]}$ are observed.

Protocol 1 Aggregate Feedback Interaction Protocol

```
1: for episode $k = 1, 2, \dots, K$ do
2:     Agent chooses policy $\pi^k$.
3:     Observes trajectory $\iota^k = (x_h^k, a_h^k)_{h\in[H]}$.
4:     Observes episode reward $v^k = \sum_{h\in[H]} r_h^k$.
```

We measure the quality of any algorithm via its *regret* – the difference between the value of the policies π^k generated by the algorithm and that of the optimal policy $\pi^\star$, i.e.,

$$\text{Regret} = \sum_{k\in[K]} V^\star(x_1) - V^{\pi^k}(x_1).$$

3. Algorithms and Main Results

We present two algorithms for regret minimization in aggregate feedback linear MDPs. The first, RE-LSVI (Algorithm 2), is a value-based optimistic algorithm, and the second, REPO (Algorithm 3), is a policy optimization routine. Both algorithms achieve regret with the optimal $\sqrt{K}$ dependence on the number of episodes, but RE-LSVI has a favorable dependence on H.

Notation. Throughout the paper $\phi_h^k = \phi(x_h^k, a_h^k) \in \mathbb{R}^d$ denotes the state-action features at horizon h of episode k, and $\phi^k = ({\phi_1^k}^\mathsf{T}, \dots, {\phi_H^k}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{dH}$ is their concatenation. Following a similar convention, for any vector $\zeta \in \mathbb{R}^{dH}$, let $\{\zeta_h\}_{h\in[H]} \in \mathbb{R}^d$ denote its H equally sized sub-vectors, i.e., $(\zeta_1^\mathsf{T}, \dots, \zeta_H^\mathsf{T})^\mathsf{T} = \zeta$. Notice that the same does not hold for matrices, e.g., Λ_h^k is not a sub-matrix of Λ^k. In addition, $\|v\|_A = \sqrt{v^\mathsf{T} A v}$, and clipping operator $\mathrm{clip}_\beta[z]$ for some $\beta > 0$ is defined as $\min\{\beta, \max\{-\beta, z\}\}$. Hyperparameters follow the notations β_z, λ_z, and η_z for some z and $\delta \in (0, 1)$ denotes a confidence parameter. Finally, in the context of an algorithm, $\leftarrow$ signs refer to compute operations whereas $=$ signs define operators, which are evaluated at specific points as part of compute operations.

3.1. Randomized Ensemble Least Squares Value Iteration (RE-LSVI)

At the start of episode k, similarly to algorithms for standard reward feedback (e.g., Jin et al. (2020b)), RE-LSVI defines a dynamics backup operator $\widehat{\psi}_h^k$ for each $h \in [H]$, which, when given a function $V : \mathcal{X} \mapsto \mathbb{R}$, estimates the vector $\psi_h V = \sum_{x\in\mathcal{X}} \psi_h(x) V(x)$ using least squares as follows:

$$\widehat{\psi}_h^k V = (\Lambda_h^k)^{-1} \sum_{\tau\in\mathcal{D}_h^k} \phi_h^\tau V(x_{h+1}^\tau), \tag{1}$$

where $\Lambda_h^k = \lambda_p I + \sum_{\tau\in\mathcal{D}_h^k} \phi_h^\tau (\phi_h^\tau)^\mathsf{T}$ for some parameter $\lambda_p > 0$ is the covariance matrix for horizon h, and $\mathcal{D}_h^k$ is a dataset. On the other hand, for reward estimates, while standard algorithms compute a least square estimator for each θ_h using the observed rewards at horizon h, this is no longer feasible in our ABF model. Instead, given the aggregate reward $v^\tau = \sum_{h\in[H]} r_h^\tau$ for $\tau < k$, it is natural to directly estimate the aggregate reward vector θ using least squares:

$$\widehat{\theta}^k = (\Lambda^k)^{-1} \sum_{\tau\in\mathcal{D}^k} \phi^\tau v^\tau, \tag{2}$$

where $\Lambda^k = \lambda_r I + \sum_{\tau\in\mathcal{D}^k} \phi^\tau (\phi^\tau)^\mathsf{T}$ for some parameter $\lambda_r > 0$ is the aggregate covariance matrix, and $\mathcal{D}^k$ is a dataset. The corresponding H sub-vectors $\{\widehat{\theta}_h^k\}_{h\in[H]}$ then estimate $\{\theta_h\}_{h\in[H]}$. For RE-LSVI, $\mathcal{D}^k$ and $\mathcal{D}_h^k$ are both simply $\{1, \dots, k-1\}$, that is, all previous data (but they will be different for our next algorithm).

To encourage exploration, we use the standard exploration bonuses $b_h^k(x,a) = \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}$ for some parameter $\beta_p > 0$ to handle uncertainty in the dynamics. To handle uncertainty in the rewards, however, it is crucial to deploy our proposed ensemble randomization technique (reasoning deferred to the end of this subsection): we draw $m \approx \log(K/\delta)$ independent Gaussian noise vectors $\zeta^{k,i} \sim \mathcal{N}(0, \beta_r^2(\Lambda^k)^{-1})$ for $i \in [m]$ whose covariance is attuned to that of the aggregate reward least squares estimate (with some parameter $\beta_r > 0$). Next, we calculate m value estimates $\{\hat{V}_h^{k,i}\}_{i\in[m]}$ using standard LSVI, each with a different perturbed reward vector $\widehat{\theta}^k + \zeta^{k,i}$ (Line 7 in Algorithm 2). Finally, we find the index $i_k \in \arg\max_{i\in[m]} \hat{V}_1^{k,i}(x_1)$ with the maximal initial value and follow the greedy policy with respect to the i_k-th Q-function.

One final important tweak we make to LSVI is the clipping of the value estimates. Specifically, standard algorithms truncate each $\hat{Q}_h^{k,i}(x,a)$ to $[0,H]$ (the range of the true Q-function), but we propose a looser clip that truncates these Q-estimates to $[-(H+1-h)\beta_{r\zeta}, (H+1-h)\beta_{r\zeta}]$ for some parameter $\beta_{r\zeta} > 0$ such that $(H+1-h)\beta_{r\zeta} \approx \sqrt{d}H\beta_r$. The symmetry in our clipping is to avoid biases in the random walk induced by $\zeta^{k,i}$, and the looser threshold is crucial for the analysis (more discussion to follow).

Algorithm 2 RE-LSVI with aggregate feedback

```
1: input: $\delta, \lambda_r, \lambda_p, \beta_r, \beta_p, \beta_{r\zeta} > 0; m \geq 1$.
2: for episode $k = 1, 2, \ldots, K$ do
3:     Define $\mathcal{D}^k = \mathcal{D}_h^k = \{1, \ldots, k-1\}$ for all $h \in [H]$.
4:     Compute $\widehat{\theta}^k$ and define $\widehat{\psi}_h^k$ (Eqs. (1) and (2)).
5:     Sample $\zeta^{k,i} \sim \mathcal{N}(0, \beta_r^2(\Lambda^k)^{-1})$ for all $i \in [m]$.
6:     Define $\hat{V}_{H+1}^{k,i}(x) = 0$ for all $i \in [m], x \in \mathcal{X}$.
7:     For every $i \in [m]$ and $h = H, \ldots, 1$:
           $w_h^{k,i} \leftarrow \widehat{\theta}_h^k + \zeta_h^{k,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}$,
           $\hat{Q}_h^{k,i}(x,a) = \mathrm{clip}_{(H+1-h)\beta_{r\zeta}}\left[\phi(x,a)^\mathsf{T} w_h^{k,i} + b_h^k(x,a)\right]$,
           $\hat{V}_h^{k,i}(x) = \max_{a\in\mathcal{A}} \hat{Q}_h^{k,i}(x,a)$
           $\pi_h^{k,i}(x) \in \arg\max_{a\in\mathcal{A}} \hat{Q}_h^{k,i}(x,a)$.
8:     $i_k \leftarrow \arg\max_{i\in[m]} \hat{V}_1^{k,i}(x_1)$ and play $\pi^k = \pi^{k,i_k}$.
9:     Observe episode reward $v^k$ and trajectory $\iota^k$.
```

Most of the operations in RE-LSVI are similar to algorithms for standard reward feedback (e.g., Jin et al. (2020b)). The only difference lies in drawing Gaussian noise vectors and calculating m value functions instead of 1. Thus, the computational complexity is similar to previous work up to a logarithmic $\log(K/\delta)$ factor. The following is our main result for Algorithm 2.

Theorem 1. *Suppose that we run RE-LSVI (Algorithm 2) with the parameters defined in Lemma 4 (in Appendix A). Then with probability at least $1-\delta$, we have*

$$\mathrm{Regret} = O\left(\sqrt{d^5 H^7 K \log^6(dHK/\delta)}\right).$$

Discussion. Algorithm 2 contains elements of both LSVI-UCB (Jin et al., 2020b) and UCBVI-TS (Efroni et al., 2021), but also new ideas, necessary for combining them. When constructing an optimistic planning procedure, Efroni et al. (2021) noticed that the reward bonuses have to be trajectory-dependent. This breaks the MDP structure and makes efficient planning impossible. Overcoming this, they suggest drawing a random bonus whose covariance scales with the uncertainty of the aggregate feedback value estimation. They then follow a standard planning procedure over the empirical MDP. Unfortunately, directly planning in the empirical linear MDP seems to be a difficult task since value backups are calculated via a least squares argument, which corresponds to a potentially non-valid transition kernel, i.e., one with negative entries and whose sum may exceed 1. Consequently, one has to contend either with value backups blowing up exponentially with H, or with an empirical optimal policy that is not greedy with respect to its Q-function.

This is typically overcome by truncating the value to $[0,H]$ before using it in the backup step of the dynamic program (see e.g., (Jin et al., 2020b)). The resulting policy is not optimal for the empirical Linear MDP but can be shown to be optimistic with respect to the true optimal policy. Unfortunately, this truncation introduces a bias to the aggregate reward estimate, breaking the correlation between the uncertainty of the individual estimates at each horizon $h \in [H]$.

We overcome this bias by introducing a loose truncation mechanism that clips the values of $\hat{Q}_h^{k,i}(h \in [H])$ to $[-(H+1-h)\beta_{r\zeta}, (H+1-h)\beta_{r\zeta}]$ where $\beta_{r\zeta}$ is a high probability bound on the immediate perturbed reward estimates. This does not immediately solve the issue since the dynamics are still invalid, but through a careful (high probability) analysis, we show that one can bound the error by the unclipped, thus unbiased, process (see the analysis in Section 4.1 from Eq. (6) to Eq. (7)).

A second important feature of our algorithm is the randomized ensemble. In Efroni et al. (2021) the equivalent noise term is sampled only once and a single value is calculated. Similar to other Thompson Sampling (TS) methods, this results in an estimator that is optimistic only with constant probability, thus requiring a careful and intricate analysis to handle the instances where it is not optimistic. In contrast, we draw $m \approx \log(K/\delta)$ noise terms and calculate their respective value functions. Each value estimate is thus optimistic with constant probability, and since the noise terms are i.i.d, it is straightforward to see that at least one of them is optimistic with high probability. We show that following the policy related to the maximal value is thus also optimistic with high ($\geq 1-\delta$) probability. This yields a clear and easy-to-follow optimism-based analysis (see Section 4.1). Moreover, in what follows we show how to extend the randomized ensemble idea to obtain a policy optimization-based algorithm for linear MDPs with ABF.

3.2. Randomized Ensemble Policy Optimization (REPO)

We now introduce our second algorithm REPO that is based on the more popular policy optimization scheme. REPO starts with a reward-free warm-up routine by Sherman et al. (2023), which is in turn based on Wagenmaker et al. (2022), to uniformly explore the state space. It lasts for $k_0 \approx \sqrt{K}$ episodes and outputs exploratory datasets $\{\mathcal{D}_h^0\}_{h\in[H]}$ that are used to initialize the least squares estimators for the value iteration backups (Eq. (1)). In addition, for each horizon $h \in [H]$, it defines a set of "known" states:

$$\mathcal{Z}_h = \left\{ x \in \mathcal{X} \mid \max_{a\in\mathcal{A}} \|\phi(x,a)\|_{(\Lambda_h^0)^{-1}} \leq \frac{1}{2\beta_w H} \right\}, \quad (3)$$

where $\Lambda_h^0 = \lambda_p I + \sum_{\tau\in\mathcal{D}_h^0} \phi_h^\tau (\phi_h^\tau)^\mathsf{T}$ and $\beta_w > 0$ is some parameter. The algorithm then proceeds in epochs, each one ending when the uncertainty (of either the dynamics or aggregate reward) shrinks by a multiplicative factor, as expressed by the determinant of their covariance (Line 5 in Algorithm 3). At the start of each epoch e, we draw $m \approx \log(K/\delta)$ reward perturbation vectors $\zeta^{k_e,i}, i \in [m]$, where k_e denotes the first episode in the epoch. We also initialize m PO sub-routines and a Multiplicative Weights (MW) sub-routine that arbitrates over the PO copies. Inside an epoch, at the start of each episode k, we compute the aggregate reward vector $\widehat{\theta}^k$ (Eq. (2)) and estimated dynamics backup operators $\widehat{\psi}_h^k$ (Eq. (1)), and run the m copies of PO, each differing only by their reward perturbation (see Line 16 in Algorithm 3). This involves running online mirror descent (OMD) updates over the estimated Q values. Finally, a policy is chosen by hedging over the values of each PO copy using MW (Line 17 in Algorithm 3).

Notice that, we do not use direct reward bonuses (b_h^k in Algorithm 2) to encourage exploration. Instead, we augment the covariance of the reward perturbations with an additional term that accounts for the dynamics uncertainty (Line 8 in Algorithm 3). Additionally, we replace the clipping mechanism in the value backups with an indicator that zeroes out the Q estimates outside the "known" states sets $\mathcal{Z}_h, h \in [H]$. Similarly to Sherman et al. (2023), this is crucial as PO performance scales with the complexity (log covering number) of the policy class, which scales with K under the clipping mechanism in Algorithm 2 (more discussion to follow).

We note that the computational complexity of Algorithm 3 is comparable to RE-LSVI (Algorithm 2). The following is our main result for Algorithm 3 (for the full analysis see Appendix B).

Algorithm 3 REPO with aggregate feedback

```
1: input: $\delta, \epsilon_{\text{cov}}, \beta_w, \eta_o, \eta_x, \lambda_r, \lambda_p, \beta_r, \beta_p > 0; m \geq 1$.
2: Run reward-free warm-up (Sherman et al., 2023, Algorithm 2) with $(\frac{\delta}{7}, \beta_w, \epsilon_{\text{cov}})$ to get index sets $\{\mathcal{D}_h^0\}_{h\in[H]}$, "known" states sets $\{\mathcal{Z}_h\}_{h\in[H]}$ (Eq. (3)), and let $k_0$ be the first episode after the warm-up.
3: initialize: $e \leftarrow -1$.
4: for episode $k = k_0, \ldots, K$ do
5:     if $k = k_0$ or $\exists h \in [H],\ \det(\Lambda_h^k) \geq 2\det(\Lambda_h^{k_e})$
           or $\det(\Lambda^k) \geq 2\det(\Lambda^{k_e})$ then
6:         $e \leftarrow e+1$ and $k_e \leftarrow k$.
7:         $\Lambda_{\text{diag}}^{k_e} \leftarrow \operatorname{diag}\left(\Lambda_1^{k_e}, \ldots, \Lambda_H^{k_e}\right)$
8:         $\Sigma_\zeta^{k_e} \leftarrow 2\beta_r^2 (\Lambda^{k_e})^{-1} + 2H\beta_p^2 \left(\Lambda_{\text{diag}}^{k_e}\right)^{-1}$
9:         Sample $\zeta^{k_e,i} \sim \mathcal{N}\left(0, \Sigma_\zeta^{k_e}\right)$ for all $i \in [m]$.
10:        Reset $p^k(i) \leftarrow 1/m, \pi_h^{k,i}(a \mid x) \leftarrow 1/|\mathcal{A}|$.
11:    Sample $i_k$ according to $p^k(\cdot)$ and play $\pi^k = \pi^{k,i_k}$.
12:    Observe episode reward $v^k$ and trajectory $\iota^k$.
13:    Define $\mathcal{D}^k = \{k_0, \ldots, k-1\}$ and $\mathcal{D}_h^k = \mathcal{D}_h^0 \cup \mathcal{D}^k$.
14:    Compute $\widehat{\theta}^k$ and define $\widehat{\psi}_h^k$ (Eqs. (1) and (2)).
15:    Define $\hat{V}_{H+1}^{k,i}(x) = 0$ for all $i \in [m], x \in \mathcal{X}$.
16:    For every $i \in [m]$ and $h = H, \ldots, 1$:
           $w_h^{k,i} \leftarrow \widehat{\theta}_h^k + \zeta_h^{k_e,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}$
           $\hat{Q}_h^{k,i}(x,a) = \phi(x,a)^\mathsf{T} w_h^{k,i} \cdot \mathbb{1}_{\{x\in\mathcal{Z}_h\}}$
           $\hat{V}_h^{k,i}(x) = \sum_{a\in\mathcal{A}} \pi_h^{k,i}(a \mid x) \hat{Q}_h^{k,i}(x,a)$
           $\pi_h^{k+1,i}(a \mid x) \propto \pi_h^{k,i}(a \mid x) \exp(\eta_o \hat{Q}_h^{k,i}(x,a))$.
17:    $p^{k+1}(i) \propto p^k(i) \exp(\eta_x \hat{V}_1^{k,i}(x_1))$ for all $i \in [m]$.
```

Theorem 2. *Suppose that we run REPO (Algorithm 3) with the parameters defined in Theorem 15 (in Appendix B). Then with probability at least $1-\delta$, we have*

$$\text{Regret} = O\left(\sqrt{d^5 H^9 K \log^9(dH|\mathcal{A}|K/\delta)}\right).$$

Discussion. There are several conceptual differences between REPO and RE-LSVI. In RE-LSVI, the ensemble's decision is greedy with respect to the current value estimates (Line 8 in Algorithm 2). This works because the individual policies $\pi_h^{k,i}$ are greedy with respect to their values. In contrast, the PO sub-routines in REPO produce non-greedy policies $\pi_h^{k,i}$ whose performance is not competitive alone, but only as an entire sequence. Thus, the ensemble decisions must ensure that the chosen policies are competitive with the best contiguous sequence in hindsight. This is ensured by hedging over the values, i.e., choosing randomly with probabilities governed by the MW rule (Line 17).

The second difference relating to the ensemble method is the use of an epoch schedule. This is a fairly standard doubling trick that allows us to keep the covariance matrix fixed in intervals (at negligible cost). We use it to draw the reward perturbations only once at the start of the epoch and keep them fixed throughout the epoch. Since the policies are only competitive as a sequence, they must also be optimistic as a sequence.[1] As the reward perturbations encourage this optimism, our analysis depends on their sum inside an epoch, which is simply a single perturbation multiplied by the length of the epoch. Had we drawn a new perturbation for each episode, their covariances would be correlated, thus their sum would be non-Gaussian, which breaks our optimistic guarantees. We suspect that a Berry-Esseen type argument for martingales might show that this distribution would converge to Gaussian, thus obviating the need for an epoch schedule. We were unable to verify this and leave it for future work.

As we mentioned previously, REPO replaces the value clipping with an indicator mechanism by Sherman et al. (2023), which zeroes out the Q values outside "known" states sets, obtained during a warm-up period. This change is necessary to control the log covering number of the policy class, on which the regret has a square root dependence. Notice that our policies are a soft-max over the sum of past Q functions, i.e., $\pi_h^{k,i}(a|x) \propto \exp(\eta_o \sum_{k=k_e}^{k-1} \hat{Q}_h^{k,i}(x,a))$. Each $\hat{Q}_h^{k,i}$ has d parameters ($w_h^{k,i}$), thus the policy class may have dK parameters in the worst case. The log covering number would also scale similarly, significantly deteriorating the regret. When Q values are calculated using clipping (Line 7 of Algorithm 2), we cannot avoid this scenario. However, using the indicator mechanism, the policy class may be summarized as $\pi_h^{k,i}(a|x) \propto \exp(\eta_o \phi(x,a)^\mathsf{T} w \cdot \mathbb{1}_{\{x \in \mathcal{Z}_h\}})$, for some $w \in \mathbb{R}^d$, which has only d parameters.

Finally, we elaborate on our choice of purely stochastic exploration bonuses. Stochastic bonuses are typically larger than their deterministic counterparts by a factor of $\sqrt{d}$. While this usually makes them unfavorable, in our context they compensate for this deficit by reducing the complexity of the policy class. Concretely, it reduces its log covering number from Hd^3 to d, yielding an overall improvement to the regret bound. We note that the same cannot be said for RE-LSVI, where our clipping mechanism necessitates the use of deterministic bonuses for the dynamics uncertainty. To better understand this issue, see the analysis in Section 4.1.

REPO for Tabular MDPs with ABF. As mentioned, a policy optimization algorithm with ABF did not exist before our work even for the tabular setting. In Appendix C, we show a simplification of REPO for tabular MDPs with ABF. There, we do not restrict the Q values at all (no indicator or clipping), and thus do not need the warm-up routine by Sherman et al. (2023). Otherwise, the algorithms are essentially identical (up to parameter settings). Our analysis for the tabular case is also novel and follows a regret decomposition that has not been applied in the context of policy optimization to the best of our knowledge. It allows us to incorporate the optimal value $V^\star$ instead of the estimated value $\hat{V}^{k,i}$ in some of the terms, thus avoiding complications with bounding the covering number of the value functions class. However, it relies on the estimated dynamics backup operators being (nearly) valid distributions, i.e., have non-negative entries and sum to less than 1, and thus cannot be applied to our current implementation for linear MDPs unfortunately.

4. Analysis

In this section we sketch our main proof ideas. For full details see Appendix A (RE-LSVI) and Appendix B (REPO).

4.1. Analysis of RE-LSVI

Overview. We start by decomposing the regret as:

$$\text{Regret} = \sum_{k\in[K]} \underbrace{V^\star(x_1) - \hat{V}_1^{k,i_k}(x_1)}_{(i)} + \sum_{k\in[K]} \underbrace{\hat{V}_1^{k,i_k}(x_1) - V^{\pi^k}(x_1)}_{(ii)} .$$

Term (i) (optimism) reflects the difference between the performance of the optimal policy $\pi^\star$ and the optimistically estimated performance of the agent's policy π^k. Term (ii) (cost of optimism) reflects the difference between the true and estimated performance of π^k.

Next, we show in Lemmas 6 and 7 that, conditioned on a "good event" that holds with probability $1-\delta$ (see Lemma 4), $(i) \le 0$, i.e., the algorithm is indeed optimistic, and

$$(ii) \le \mathbb{E}_{P,\pi^k}\left[(\beta_r + \beta_\zeta)\|\phi^k\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h\in[H]} \|\phi_h^k\|_{(\Lambda_h^k)^{-1}}\right],$$

where β_ζ is a high-probability bound on the magnitude of the noise $\zeta^{k,i}$. Applying Azuma's inequality to the sum of

[1] We note that the notion of optimism in PO algorithms is slightly different compared to value iteration methods.

(ii) over k (a part of the good event), we conclude that

$$\begin{aligned}\text{Regret} &\le (\beta_r + \beta_\zeta) \sum_{k\in[K]} \|\phi^k\|_{(\Lambda^k)^{-1}} \\ &\quad + 2\beta_p \sum_{h\in[H]} \sum_{k\in[K]} \|\phi^k_h\|_{(\Lambda^k_h)^{-1}} \\ &\quad + (\beta_r + \beta_\zeta + 2H\beta_p)\sqrt{2K\log(5/\delta)}.\end{aligned}$$

The proof is concluded by bounding the terms $\sum_k \|\phi^k\|_{(\Lambda^k)^{-1}}$ and $\sqrt{H}\sum_k \|\phi^k_h\|_{(\Lambda^k_h)^{-1}}$ as $\tilde{O}(\sqrt{dHK})$ using the Cauchy-Schwarz inequality and Lemma 26, a standard elliptical potential lemma (see, e.g., Cohen et al., 2019, Lemma 13).

Optimism. In the remainder of this section, we explain the main claims showing that $(i) \le 0$. The proof for term (ii) follows similar arguments. First, we use a value difference lemma by Shani et al. (2020) to get that for every $i \in [m]$:

$$\begin{aligned}&V^\star(x_1) - \hat{V}^{k,i}_1(x_1) \qquad (4)\\ &= \mathbb{E}_{P,\pi^\star} \sum_{h\in[H]} \underbrace{\hat{Q}^{k,i}_h(x_h, \pi^\star_h(x_h)) - \hat{Q}^{k,i}_h(x_h, \pi^{k,i}_h(x_h))}_{(**)} \\ &\underbrace{+ \mathbb{E}_{P,\pi^\star} \sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}^{k,i}_{h+1}) - \hat{Q}^{k,i}_h(x_h,a_h)}_{(*)}.\end{aligned}$$

By the greedy definition of $\pi^{k,i}_h$, $(**) \le 0$. Next, suppose that, for all $k \in [K], h \in [H], i \in [m]$, the estimation error of the dynamics backup operator is well-concentrated

$$\|(\psi_h - \widehat{\psi}^k_h)\hat{V}^{k,i}_{h+1}\|_{\Lambda^k_h} \le \beta_p, \tag{5}$$

and the perturbed estimated reward is bounded

$$|\phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h)| \le \beta_{r\zeta}, \forall x \in \mathcal{X}, a \in \mathcal{A}, \tag{6}$$

both consequences of the good event. Then we have that

$$\begin{aligned}&\phi(x,a)^\mathsf{T} w^{k,i}_h + \beta_p \|\phi(x,a)\|_{(\Lambda^k_h)^{-1}} \\ &\quad = \phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h + \widehat{\psi}^k_h \hat{V}^{k,i}_{h+1}) + \beta_p\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}} \\ &\quad \ge \phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h + \psi_h \hat{V}^{k,i}_{h+1}) \quad \text{(Cauchy-Schwarz)} \\ &\quad + (\beta_p - \|(\widehat{\psi}^k_h - \psi_h)\hat{V}^{k,i}_{h+1}\|_{\Lambda^k_h})\|\phi(x,a)\|_{(\Lambda^k_h)^{-1}} \\ &\quad \ge \phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h + \psi_h \hat{V}^{k,i}_{h+1}). \qquad \text{(Eq. (5))}\end{aligned}$$

Due to the clipping of $\hat{V}^{k,i}_{h+1}$ and Eq. (6), the last term is absolutely bounded by $(H+1-h)\beta_{r\zeta}$. Combined with the fact that clipping is non-decreasing, we conclude that

$$\hat{Q}^{k,i}_h(x,a) \ge \phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h + \psi_h \hat{V}^{k,i}_{h+1}),$$

and plugging this back into $(*)$, we get that

$$\begin{aligned}(*) &\le \mathbb{E}_{P,\pi^\star} \sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}(\theta_h - \widehat{\theta}^k_h - \zeta^{k,i}_h) \\ &= \phi^{\pi^\star\mathsf{T}}(\theta - \widehat{\theta}^k - \zeta^{k,i}),\end{aligned} \tag{7}$$

where $\phi^{\pi^\star} = \mathbb{E}_{P,\pi^\star}(\phi(x_1,a_1)^\mathsf{T} \ldots, \phi(x_H,a_H)^\mathsf{T})^\mathsf{T}$.

Now, suppose that, for all $k \in [K]$, the aggregate reward estimation error is well concentrated

$$\|\theta - \widehat{\theta}^k\|_{\Lambda^k} \le \beta_r, \tag{8}$$

and the perturbations are optimistic in the sense that

$$\max_{i\in[m]} \phi^{\pi^\star\mathsf{T}}\zeta^{k,i} \ge \beta_r\|\phi^{\pi^\star}\|_{(\Lambda^k)^{-1}}, \tag{9}$$

both also a part of the good event. Recalling the definition of i_k (Line 8 in Algorithm 2) and putting everything together:

$$\begin{aligned}(i) &= \min_{i\in[m]} V^\star(x_1) - \hat{V}^{k,i}_1(x_1) \\ &\le \phi^{\pi^\star\mathsf{T}}(\theta - \widehat{\theta}^k) - \max_{i\in[m]} \phi^{\pi^\star\mathsf{T}}\zeta^{k,i} && \text{(Eq. (7))} \\ &\le \beta_r\|\phi^{\pi^\star}\|_{(\Lambda^k)^{-1}} - \max_{i\in[m]} \phi^{\pi^\star\mathsf{T}}\zeta^{k,i} && \text{(Eq. (8))} \\ &\le 0. && \text{(Eq. (9))}\end{aligned}$$

Proof (sketch) of Eq. (9). We conclude this section with some intuition regarding Eq. (9). Notice that its right-hand side looks like the desired deterministic bonus. On the other hand, the argument inside the max operation is the effective bonus of each member in the ensemble. Eq. (9) may thus be interpreted as a requirement that at least one perturbation yields the correct bonus under the optimal policy, thus will have an optimistic value. As we choose to follow the ensemble member with the largest value, it too must be optimistic. As for verifying that Eq. (9) holds with high probability, while this may appear a complex argument, it follows from the following fundamental result. Let $g_i = \phi^{\pi^\star\mathsf{T}}\zeta^{k,i}$ and notice that, conditioned on Λ^k, $\{g_i\}_{i\in[m]}$ are i.i.d $\mathcal{N}(0, \beta_r^2\|\phi^{\pi^\star}\|^2_{(\Lambda^k)^{-1}})$ variables. Eq. (9) thus holds with high probability by the following anti-concentration result.

Lemma 3. *Let $\sigma, m \ge 0$. Suppose that $g_i \sim \mathcal{N}(0,\sigma^2)$, $i \in [m]$ are i.i.d Gaussian random variables. With probability at least $1 - e^{-m/9}$*

$$\max_{i\in[m]} g_i \ge \sigma.$$

Proof. Recall that for a standard Gaussian random variable $G \sim \mathcal{N}(0,1)$ we have that $\Pr[G \ge 1] \ge 1/9$. Since g_i are independent, we conclude that

$$\Pr[\max_{i\in[m]} g_i \le \sigma] = \Pr[G \le 1]^m \le (8/9)^m \le e^{-m/9},$$

and taking the complement concludes the proof. ■

4.2. Analysis of REPO

Regret decomposition. We start with a coarse bound on the regret incurred during the warm-up period and decompose the remaining term as in RE-LSVI (Section 4.1), i.e.,

$$\text{Regret} \le Hk_0 + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} V_1^\star(x_1) - \hat{V}_1^{k,i_k}(x_1)}_{(i)} + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1)}_{(ii)},$$

where $E \approx dH\log K$ is the number of epochs and K_e is the set of episodes within epoch e. Recalling that k_0 is the length of the warm-up routine, a result by Sherman et al. (2023) (see Lemma 17) guarantees that $k_0 \approx 1/\epsilon_{\text{cov}} \approx \sqrt{K}$.

As in Section 4.1, we focus on bounding (i) since bounding (ii) uses a subset of the necessary techniques. We would have liked to decompose (i) for each episode separately (as in Eq. (4)). However, $\pi_h^{k,i}$ are no longer greedy, thus $(**)$ in Eq. (4) would not be non-positive anymore. We thus perform the same decomposition but over the entire epoch to get that for every $e \in [E], i \in [m]$:

$$\sum_{k\in K_e} V^\star(x_1) - \hat{V}_1^{k,i}(x_1) = \mathbb{E}_{P,\pi^\star}\sum_{h\in[H]} R_{Q,h}^{i}(x_h) + \sum_{k\in K_e} R_{Opt}^{k,i},$$

where $R_{Opt}^{k,i}$ is defined as $(*)$ in Eq. (4) and $R_{Q,h}^{i}(x_h)$ is

$$\sum_{k\in K_e}\sum_{a\in\mathcal{A}} \hat{Q}_h^{k,i}(x_h,a)(\pi_h^\star(a \mid x_h) - \pi_h^{k,i}(a \mid x_h)).$$

OMD. Since $\pi_h^{k,i}$ are updated using OMD with learning rate $\eta_o > 0$ (see Line 16 in Algorithm 3), we can invoke a standard OMD argument (Lemma 13) to get that

$$R_{Q,h}^{i}(x_h) \le \frac{\log|\mathcal{A}|}{\eta_o} + \eta_o|K_e|\beta_Q^2, \quad \forall x_h \in \mathcal{X},$$

where β_Q is a bound on $\max_{x\in\mathcal{X},a\in\mathcal{A}}|\hat{Q}_h^{k,i}(x,a)|$. Notice that, unlike RE-LSVI, $\hat{Q}_h^{k,i}$ are not bounded by definition in REPO. Nonetheless, we adapt arguments from Sherman et al. (2023) to show that $\beta_Q \approx \beta_r H\sqrt{d}$ as part of a good event, which includes the events already defined in Section 4.1.

Hedge. Next, we show (at the end of this section) that on the good event, there exists $\hat{j}_e \in [m]$ such that for $\epsilon_{\text{cov}} \approx 1/\sqrt{K}$ and all $k \in K_e$

$$R_{Opt}^{k,\hat{j}_e} \le 2\epsilon_{\text{cov}} H\beta_Q. \tag{10}$$

The connection between i_k and $\hat{j}_e$ is done through the MW update rule (Line 17 in Algorithm 3). For learning rate $\eta_x > 0$, a standard result (Lemma 12) implies that

$$\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,j_e}(x_1) - \hat{V}_1^{k,i_k}(x_1) \le \frac{E\log m}{\eta_x} + \eta_x K\beta_Q^2 + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}.$$

Bounding term (i). Putting everything together, we have

$$\begin{aligned}(i) &= \sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,\hat{j}_e}(x_1) - \hat{V}_1^{k,i_k}(x_1) \\ &\quad + \sum_{e\in[E]}\sum_{k\in K_e} V_1^\star(x_1) - \hat{V}_1^{k,\hat{j}_e}(x_1) \\ &\le \frac{E\log m}{\eta_x} + \eta_x K\beta_Q^2 + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}} \\ &\quad + \frac{EH\log|\mathcal{A}|}{\eta_o} + \eta_o H\beta_Q^2 K + 2\epsilon_{\text{cov}} H\beta_Q K,\end{aligned}$$

and plugging the parameter choices concludes the desired $O(\sqrt{K})$ bound. Notice that calculating $\hat{j}_e$ can only be done at the end of an epoch as it essentially maximizes the sum of estimated values inside an epoch. However, in every episode we need to choose a member of the ensemble whose policy we follow. This demonstrates the necessity of hedging over the ensemble in REPO (Line 17 in Algorithm 3), as opposed to the greedy method in RE-LSVI (Line 8 in Algorithm 2).

Proof (sketch) of Eq. (10). To begin, we show in Lemma 9 that the reward-free warm-up explores the state space well in the sense that, for any policy $\pi \in \Pi_M$ and vector $v \in \mathbb{R}^d$

$$\left|\mathbb{E}_{P,\pi}(\phi(x_h,a_h) - \bar{\phi}_h(x_h,a_h))^\mathsf{T} v\right| \le \epsilon_{\text{cov}} \max_{x,a}|\phi(x,a)^\mathsf{T} v|,$$

where $\bar{\phi}_h(x,a) = \phi(x,a)\mathbb{1}_{\{x\in\mathcal{Z}_h\}}$ are truncated features according to the known states set $\mathcal{Z}_h$. We then rewrite $R_{Opt}^{k,i}$ in the following way using the definition of $\hat{Q}_h^{k,i}(x_h,a_h)$:

$$R_{Opt}^{k,i} = \underbrace{\sum_{h\in[H]} (\bar{\phi}_h^{\pi^\star})^\mathsf{T}\Big(\theta_h - \widehat{\theta}_h^k - \zeta_h^{k_e,i} + (\psi_h - \widehat{\psi}_h^k)\hat{V}_{h+1}^{k,i}\Big)}_{(a)} + \underbrace{\mathbb{E}_{P,\pi^\star}\sum_{h\in[H]}(\phi(x_h,a_h) - \bar{\phi}_h(x_h,a_h))^\mathsf{T}(\theta_h + \psi_h\hat{V}_{h+1}^{k,i})}_{(b)},$$

where $\bar{\phi}_h^{\pi^\star} = \mathbb{E}_{P,\pi^\star}\bar{\phi}_h(x_h,a_h)$. Recall that, on the good event, $|\hat{V}_h^{k,i}(x)| \le \beta_Q$, thus the above argument bounds (b) by $2\epsilon_{\text{cov}}H\beta_Q$. Letting $\bar{\phi}^{\pi^\star} = ((\bar{\phi}_1^{\pi^\star})^\mathsf{T},\ldots,(\bar{\phi}_H^{\pi^\star})^\mathsf{T})^\mathsf{T}$, we apply the Cauchy-Schwarz inequality together with the least squares estimation bounds (Eqs. (5) and (8)) to get that

$$
\begin{aligned}
(a) &\le \beta_r \|\bar{\phi}^{\pi^\star}\|_{(\Lambda^k)^{-1}} + \beta_p \sum_{h \in [H]} \|\bar{\phi}_h^{\pi^\star}\|_{(\Lambda_h^k)^{-1}} - (\bar{\phi}^{\pi^\star})^\mathsf{T} \zeta^{k_e,i} \\
&\le \|\bar{\phi}^{\pi^\star}\|_{\Sigma_\zeta^{k_e}} - (\bar{\phi}^{\pi^\star})^\mathsf{T} \zeta^{k_e,i},
\end{aligned}
$$

where the last inequality also uses $\Lambda_h^{k_e} \preceq \Lambda_h^k, \Lambda^{k_e} \preceq \Lambda^k$, the definition of $\Sigma_\zeta^{k_e}$, and the Cauchy-Schwarz inequality. Finally, similarly to Eq. (9), we have that with high probability there exists $\hat{j}_e \in [m]$ such that the above is at most zero.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant number 2549/19), the Len Blavatnik and the Blavatnik Family Foundation, and the Israeli VATAT data science scholarship.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chatterji, N., Pacchiano, A., Bartlett, P., and Jordan, M. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34:3401–3412, 2021.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pp. 1300–1309, 2019.

Cohen, A., Kaplan, H., Koren, T., and Mansour, Y. Online markov decision processes with aggregate bandit feedback. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1301–1329. PMLR, 15–19 Aug 2021.

Efroni, Y., Merlis, N., and Mannor, S. Reinforcement learning with trajectory feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7288–7295, May 2021.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Howson, B., Pike-Burke, C., and Filippi, S. Optimism and delays in episodic reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6061–6094. PMLR, 2023.

Jain, A., Wojcik, B., Joachims, T., and Saxena, A. Learning trajectory preferences for manipulators via iterative improvement. *Advances in neural information processing systems*, 26, 2013.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7281–7289, 2022.

Lancewicki, T., Rosenberg, A., and Sotnikov, D. Delay-adapted policy optimization and improved regret for adversarial MDP with delayed bandit feedback. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18482–18534. PMLR, 2023.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pp. 2209–2218, 2019a.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019b.

Saha, A., Pacchiano, A., and Lee, J. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 6263–6289. PMLR, 2023.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Sherman, U., Cohen, A., Koren, T., and Mansour, Y. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Munos, R., Naumov, A., pierre perrault, Valko, M., and MENARD, P. Model-free posterior sampling via learning rate randomization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pp. 22430–22456. PMLR, 2022.

Wang, Y., Liu, Q., and Jin, C. Is RLHF more difficult than standard RL? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46, 2017.

Wu, R. and Sun, W. Making rl with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2023.

Xiong, Z., Shen, R., Cui, Q., Fazel, M., and Du, S. S. Near-optimal randomized exploration for tabular markov decision processes. *Advances in Neural Information Processing Systems*, 35:6358–6371, 2022.

Xu, T., Wang, Y., Zou, S., and Liang, Y. Provably efficient offline reinforcement learning with trajectory-wise reward. *arXiv preprint arXiv:2206.06426*, 2022.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023a.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. How to query human feedback efficiently in rl? *arXiv preprint arXiv:2305.18505*, 2023b.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 1583–1591, 2013.

A. Proofs of Randomized Ensemble Least Squares Value Iteration (RE-LSVI)

A.1. Proof of Theorem 1

We begin by defining a so-called "good event", followed by optimism and cost of optimism, and conclude with the proof of Theorem 1.

Good event. We define the following good event $E_g = E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5$, over which the regret is deterministically bounded:

$$E_1 = \left\{\forall k \in [K] : \|\theta - \widehat{\theta}^k\|_{\Lambda^k} \le \beta_r\right\}; \tag{11}$$

$$E_2 = \left\{\forall k \in [K], h \in [H], i \in [m] : \|(\psi_h - \widehat{\psi}^k_h)\hat{V}^{k,i}_{h+1}\|_{\Lambda^k_h} \le \beta_p\right\}; \tag{12}$$

$$E_3 = \{\forall k \in [K], i \in [m] : \|\zeta^{k,i}\|_{\Lambda^k} \le \beta_\zeta \beta_r\}; \tag{13}$$

$$E_4 = \left\{\forall k \in [K] : \max_{i \in [m]} (\phi^{\pi^\star})^\mathsf{T} \zeta^{k,i} \ge \beta_r \|\phi^{\pi^\star}\|_{(\Lambda^k)^{-1}}\right\}; \tag{14}$$

$$E_5 = \left\{\sum_{k \in [K]} \mathbb{E}_{P,\pi^k}[Y_k] \le \sum_{k \in [K]} Y_k + (\beta_r(1 + \beta_\zeta) + 2H\beta_p)\sqrt{2K \log \frac{5}{\delta}}\right\}; \tag{15}$$

where $\phi^{\pi^\star} = \mathbb{E}_{P,\pi^\star}[\phi(x_{1:H}, a_{1:H})]$, $\phi(x_{1:H}, a_{1:H}) = (\phi(x_1, a_1)^\mathsf{T}, \ldots, \phi(x_H, a_H)^\mathsf{T})^\mathsf{T}$, and $Y_k = (\beta_r + \beta_\zeta)\|\phi^k\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h \in [H]} \|\phi^k_h\|_{(\Lambda^k_h)^{-1}}$.

Lemma 4 (Good event). *Consider the following parameter setting:*

$$\lambda_p = 1, \lambda_r = H, m = 9\log(5K/\delta), \beta_r = 2H\sqrt{2dH\log(10K/\delta)},$$
$$\beta_\zeta = \sqrt{\frac{11dH}{2} \log \frac{5mK}{\delta}}, \beta_{r\zeta} = 2\beta_\zeta\beta_r/\sqrt{H}, \beta_p = 40\beta_\zeta\beta_r d\sqrt{H\log(163KdH/\delta)}$$

Then $\Pr[E_g] \ge 1 - \delta$.

Lemma 5. *Under the parameter choices in Lemma 4 and the good event E_g (Eqs. (11) to (15)), we have*

$$|\phi(x,a)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k,i}_h + \psi_h\hat{V}^{k,i}_{h+1})| \le (H + 1 - h)\beta_{r\zeta} \qquad , \forall h \in [H].$$

This second result is a straightforward calculation that follows from the first (proofs in Appendix A.2).

Optimism and its cost. The following two results show that our value construction is optimistic concerning the true performance of the optimal policy but not overly optimistic compared with the performance of its induced policy.

Lemma 6 (Optimism). *Suppose that the good event E_g holds (Eqs. (11) to (15)), then*

$$V^\star(x_1) - \hat{V}^{k,i_k}_1(x_1) \le 0 \quad , \forall k \in [K].$$

Proof. First, we use Lemma 25, a value difference lemma by Shani et al. (2020), to get that for every $i \in [m]$

$$\begin{aligned} V^\star(x_1) - \hat{V}^{k,i}_1(x_1) &= \mathbb{E}_{P,\pi^\star} \sum_{h \in [H]} \hat{Q}^{k,i}_h(x_h, \pi^\star_h(x_h)) - \hat{Q}^{k,i}_h(x_h, \pi^{k,i}_h(x_h)) \\ &\quad + \mathbb{E}_{P,\pi^\star} \sum_{h \in [H]} \phi(x_h, a_h)^\mathsf{T}(\theta_h + \psi_h\hat{V}^{k,i}_{h+1}) - \hat{Q}^{k,i}_h(x_h, a_h) \\ &\le \mathbb{E}_{P,\pi^\star} \sum_{h \in [H]} \phi(x_h, a_h)^\mathsf{T}(\theta_h + \psi_h\hat{V}^{k,i}_{h+1}) - \hat{Q}^{k,i}_h(x_h, a_h), \end{aligned}$$

where the inequality is by the greedy definition of $\pi_h^{k,i}$. Next, because clip is a non-decreasing operator, we have that

$$\begin{aligned}
\hat{Q}_h^{k,i}(x,a) &= \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T} w_h^{k,i} + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&= \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}) + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&\geq \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}) + (\beta_p - \|(\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i}\|_{\Lambda_h^k})\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy-Schwarz)} \\
&\geq \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})\Big] \qquad \text{(Eq. (12))} \\
&= \phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}). \qquad \text{(Lemma 5)}
\end{aligned}$$

Plugging this back into the above, we get that

$$V^\star(x_1) - \hat{V}_1^{k,i}(x_1) \leq \mathbb{E}_{P,\pi^\star} \sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}(\theta_h - \widehat{\theta}_h^k - \zeta_h^{k,i}) = {\phi^{\pi^\star}}^\mathsf{T}(\theta - \widehat{\theta}^k - \zeta^{k,i}),$$

where $\phi^{\pi^\star} = \mathbb{E}_{P,\pi^\star}(\phi(x_1,a_1)^\mathsf{T}\dots,\phi(x_H,a_H)^\mathsf{T})^\mathsf{T}$. Finally, recalling the definition of i_k (Line 8 in Algorithm 2), we get

$$\begin{aligned}
V^\star(x_1) - \hat{V}_1^{k,i_k}(x_1) = \min_{i\in[m]} V^\star(x_1) - \hat{V}_1^{k,i}(x_1) &\leq {\phi^{\pi^\star}}^\mathsf{T}(\theta - \widehat{\theta}^k) - \max_{i\in[m]} {\phi^{\pi^\star}}^\mathsf{T}\zeta^{k,i} \\
&\leq \beta_r \|\phi^{\pi^\star}\|_{\Lambda^{k-1}} - \max_{i\in[m]} {\phi^{\pi^\star}}^\mathsf{T}\zeta^{k,i} \qquad \text{(Cauchy-Schwarz, Eq. (11))} \\
&\leq 0. \qquad \text{(Eq. (14))}
\end{aligned}$$

■

Lemma 7 (Cost of optimism). *Suppose that the good event E_g holds (Eqs. (11) to (15)), then*

$$\hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1) \leq \mathbb{E}_{P,\pi^k}\left[\beta_r(1+\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \quad ,\forall k\in[K].$$

Proof. First, because clip is a non-decreasing operator, we have that

$$\begin{aligned}
\hat{Q}_h^{k,i}(x,a) &= \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T} w_h^{k,i} + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&= \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}) + \beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&\leq \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}) + (\beta_p + \|(\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i}\|_{\Lambda_h^k})\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy-Schwarz)} \\
&\leq \mathrm{clip}_{(H+1-h)\beta_r\zeta}\Big[\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}) + 2\beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\Big] \qquad \text{(Eq. (12))} \\
&\leq \phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i}) + 2\beta_p \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(Lemma 5, } \mathrm{clip}_\beta[z] \leq z, \forall z \geq -\beta\text{)}
\end{aligned}$$

Next, we use Lemma 25, a value difference lemma by Shani et al. (2020), to get that

$$
\begin{aligned}
\hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1) &= \mathbb{E}_{P,\pi^\star}\left[\sum_{h\in[H]} \hat{Q}_h^{k,i}(x_h,a_h) - \phi(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}_{h+1}^{k,i})\right] \\
&\le \mathbb{E}_{P,\pi^\star}\left[\sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}(\theta_h - \widehat{\theta}_h^k - \zeta_h^{k,i}) + 2\beta_p\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\
&= \mathbb{E}_{P,\pi^\star}\left[\phi(x_{1:H},a_{1:H})^\mathsf{T}(\theta - \widehat{\theta}^k - \zeta^{k,i}) + 2\beta_p \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\
&\le \mathbb{E}_{P,\pi^\star}\left[\beta_r(1+\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right]. \\
&\qquad \text{(Cauchy-Schwarz, Eqs. (11) and (13))}
\end{aligned}
$$

■

Regret bound. The following is our main result for Algorithm 2.

Theorem (restatement of Theorem 1). *Suppose that we run RE-LSVI (Algorithm 2) with the parameters defined in Lemma 4. Then with probability at least $1-\delta$, we have*

$$
\text{Regret} \le 1088\sqrt{d^5H^7K}\log^2\big(1467KdH\log(5K/\delta)/\delta\big) = \tilde{O}(\sqrt{d^5H^7K}).
$$

Proof. Suppose that the good event E_g holds (Eqs. (11) to (15)). By Lemma 4, this holds with probability at least $1-\delta$. We conclude that

$$
\begin{aligned}
\text{Regret} &= \sum_{k\in[K]} V^\star(x_1) - \hat{V}_1^{k,i_k}(x_1) + \sum_{k\in[K]} \hat{V}_1^{k,i_k}(x_1) - V^{\pi^k}(x_1) \\
&\le \sum_{k\in[K]} \mathbb{E}_{P,\pi^k}\left[\beta_r(1+\beta_\zeta)\|\phi^k\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h\in[H]} \|\phi_h^k\|_{(\Lambda_h^k)^{-1}}\right] && \text{(Lemmas 6 and 7)} \\
&\le \beta_r(1+\beta_\zeta)\sum_{k\in[K]} \|\phi^k\|_{(\Lambda^k)^{-1}} + 2\beta_p \sum_{h\in[H]}\sum_{k\in[K]} \|\phi_h^k\|_{(\Lambda_h^k)^{-1}} + (\beta_r(1+\beta_\zeta) + 2H\beta_p)\sqrt{2K\log\frac{5}{\delta}} && \text{(Eq. (15))} \\
&\le \beta_r(1+\beta_\zeta)\sqrt{2dHK\log(2K)} + 2\beta_p H\sqrt{2dK\log(2K)} + (\beta_r(1+\beta_\zeta) + 2H\beta_p)\sqrt{2K\log\frac{5}{\delta}} \\
&\qquad \text{(Lemma 26, } \|\phi^k\|^2 \le H = \lambda_r, \|\phi_h^k\|^2 \le 1 = \lambda_p) \\
&\le 2\beta_r(1+\beta_\zeta)\sqrt{dHK\log(10K/\delta)} + 4\beta_p H\sqrt{dK\log(10K/\delta)} && (\sqrt{x}+\sqrt{y} \le \sqrt{2x+2y}) \\
&\le 4\beta_\zeta\beta_r\sqrt{dHK\log(10K/\delta)} + 160\beta_\zeta\beta_r\sqrt{d^3H^3K}\log(163KdH/\delta) \\
&\le 164\beta_\zeta\beta_r\sqrt{d^3H^3K}\log(163KdH/\delta) \\
&\le 1088\sqrt{d^5H^7K}\log^2(163mKdH/\delta) \\
&\le 1088\sqrt{d^5H^7K}\log^2\big(1467KdH\log(5K/\delta)/\delta\big),
\end{aligned}
$$

where the last four transitions used our parameter choices. ■

A.2. Proofs of good event (RE-LSVI)

Lemma (restatement of Lemma 5). *Under the parameter choices in Lemma 4 and the good event E_g (Eqs. (11) to (15)), we have*

$$
|\phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})| \le (H+1-h)\beta_{r\zeta} \qquad ,\forall h \in [H].
$$

Proof. Under E_g we have for all $k \in [K], h \in [H], i \in [m], a \in \mathcal{A}, x \in \mathcal{X}$:

$$
\begin{aligned}
|\phi(x,a)^\intercal(\widehat{\theta}_h^k + \zeta_h^{k,i})| &\le 1 + \|\widehat{\theta}_h^k - \theta_h + \zeta_h^{k,i}\| && (\phi(x,a)^\intercal\theta_h \in [0,1], \|\phi(x,a)\| \le 1) \\
&\le 1 + \|\widehat{\theta}^k - \theta\| + \|\zeta^{k,i}\| && \text{(triangle inequality)} \\
&\le 1 + (\|\widehat{\theta}^k - \theta\|_{\Lambda^k} + \|\zeta^{k,i}\|_{\Lambda^k})/\sqrt{\lambda_r} && (\Lambda^k \succeq \lambda_r I) \\
&\le 1 + \beta_r(1+\beta_\zeta)/\sqrt{\lambda_r} && \text{(Eqs. (11) and (13))} \\
&\le \beta_{r\zeta}, && (\beta_\zeta \ge 2, \lambda_r = H)
\end{aligned}
$$

and thus we also get that

$$
\begin{aligned}
|\phi(x,a)^\intercal(\widehat{\theta}_h^k + \zeta_h^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})| &\le \beta_{r\zeta} + |\phi(x,a)^\intercal \psi_h \hat{V}_{h+1}^{k,i}| \\
&= \beta_{r\zeta} + |\mathbb{E}_{x' \sim P_h(\cdot|x,a)} \hat{V}_{h+1}^{k,i}(x')| \\
&\le \beta_{r\zeta} + \max_{x' \in \mathcal{X}, a' \in \mathcal{A}} |\hat{Q}_{h+1}^{k,i}(x',a')| \le (H+1-h)\beta_{r\zeta},
\end{aligned}
$$

where the last inequality is due to the clipping of $\hat{Q}_{h+1}^{k,i}$ (Line 7 and Algorithm 2). ■

Lemma (restatement of Lemma 4). *Consider the following parameter setting:*

$$
\lambda_p = 1, \lambda_r = H, m = 9\log(5K/\delta), \beta_r = 2H\sqrt{2dH\log(10K/\delta)},
$$
$$
\beta_\zeta = \sqrt{\frac{11dH}{2}\log\frac{5mK}{\delta}}, \beta_{r\zeta} = 2\beta_\zeta\beta_r/\sqrt{H}, \beta_p = 40\beta_\zeta\beta_r d\sqrt{H\log(163KdH/\delta)}
$$

Then $\Pr[E_g] \ge 1 - \delta$.

Proof. First, notice that $\|\phi^k\|_{(\Lambda^k)^{-1}}, \|\phi_h^k\|_{(\Lambda_h^k)^{-1}} \le 1$, thus $0 \le Y_k \le \beta_r(1+\beta_\zeta) + 2H\beta_p$. Using Azuma's inequality, we conclude that E_5 (Eq. (15)) holds with probability at least $1 - \delta/5$. Next, by Lemma 31 and our choice of parameters, E_1 (Eq. (11)) holds with probability at least $1-\delta/5$. Now, suppose that the noise is generated such that $\zeta^{k,i} = \beta_r(\Lambda_h^k)^{-1/2}g^{k,i}$ where $g^{k,i} \sim \mathcal{N}(0, I_dH)$ are i.i.d for all $k \in [K], i \in [m]$. Indeed, notice that

$$
\mathbb{E}(\zeta^{k,i})(\zeta^{k,i})^\intercal = \beta_r^2(\Lambda_h^k)^{-1/2}\mathbb{E}\left[(g^{k,i})(g^{k,i})^\intercal\right](\Lambda_h^k)^{-1/2} = \beta_r^2(\Lambda_h^k)^{-1}.
$$

Taking a union bound over Lemma 29 with $\delta/5mK$, we have that with probability at least $1-\delta/5$, simultaneously for all $i \in [m], k \in [K]$

$$
\|\zeta^{k,i}\|_{\Lambda_h^k} = \beta_r\|g^{k,i}\| \le \beta_r\sqrt{\frac{11dH}{2}\log\frac{5mK}{\delta}} = \beta_r\beta_\zeta,
$$

thus establishing E_3 (Eq. (13)). Next, notice that conditioned on Λ^k, $(\phi^{\pi^\star})^\intercal\zeta^{k,i}, i \in [m]$ are i.i.d $\mathcal{N}(0, \beta_r^2\|\phi^{\pi^\star}\|^2_{(\Lambda^k)^{-1}})$. Applying Lemma 28 with $m = 9\log(5K/\delta)$ and taking a union bound, we have that E_4 (Eq. (14)) holds with probability at least $1-\delta/5$.

Now, for any $h \in [H]$ consider the function class $\mathcal{V}_h \subseteq \mathbb{R}^{\mathcal{X}}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ with the following parametric form

$$
V(\cdot) = \mathrm{clip}_{(H+1-h)\beta_{r\zeta}}\left[\max_a w^\intercal\phi(\cdot,a) + \beta\sqrt{\phi(\cdot,a)^\intercal\Lambda^{-1}\phi(\cdot,a)}\right]
$$

where (w, β, Λ) are parameters satisfying $\|w\| \le 4KH\beta_{r\zeta}, \beta \in [0, \bar{\beta}]$ where $\bar{\beta} = 876\beta_{r\zeta}d^{7/4}HK\log(5H^2/\delta)$, and $\lambda_{\min}(\Lambda) \ge 1$ where $\lambda_{\min}$ denotes the minimal eigenvalue. Let $\mathcal{N}_{\epsilon,h}$ be the ϵ-covering number of $\mathcal{V}_h$ with respect to the supremum distance. Then for $\epsilon = \beta_{r\zeta} H\sqrt{d}/2K$, Lemma 34 says that

$$\begin{aligned}
\log \mathcal{N}_{\epsilon,h} &\leq d \log\left(1 + \frac{16KH\beta_{r\zeta}}{\epsilon}\right) + d^2 \log\left(1 + \frac{8\sqrt{d}\bar{\beta}^2}{\epsilon^2}\right) \\
&= d \log\left(1 + \frac{32K^2}{\sqrt{d}}\right) + d^2 \log\left(1 + \frac{32K^2\bar{\beta}^2}{\sqrt{d}H^2\beta_{r\zeta}^2}\right) \\
&\leq 4d^2 \log\left(\frac{6K\bar{\beta}}{d^{1/4}H\beta_{r\zeta}}\right) \\
&= 4d^2 \log\left(5256K^2 d^{3/2} \log\frac{5H^2}{\delta}\right) \\
&\leq 8d^2 \log(163KdH/\delta).
\end{aligned}$$

Applying Lemma 32 to $\mathcal{V}_h, h \in [H]$, we have that with probability at least $1 - \delta/5$ simultaneously for all $k \in [K], h \in [H], V \in \mathcal{V}_{h+1}$

$$\begin{aligned}
\|(\psi - \widehat{\psi}_h^k)V\|_{\Lambda_h^k} &\leq 4\beta_{r\zeta} H\sqrt{d\log(K+1) + 2\log(5H\mathcal{N}_\epsilon/\delta)} \\
&\leq 4\beta_{r\zeta} H\sqrt{d\log(K+1) + 2\log(5H^2/\delta) + 16d^2\log(163KdH/\delta)} \\
&\leq 20\beta_{r\zeta} dH\sqrt{\log(163KdH/\delta)} \\
&= 40\beta_\zeta\beta_r d\sqrt{H\log(163KdH/\delta)} \\
&= \beta_p.
\end{aligned}$$

Taking a union bound, all of the events so far hold with probability at least $1 - \delta$. Finally, we show that these events also imply that $\hat{V}_{h+1}^{k,i} \in \mathcal{V}_{h+1}$, thus E_2 (Eq. (12)) holds. $\hat{V}_{h+1}^{k,i}$ has the correct functional form. It remains to show that its parameters are within the range of $\mathcal{V}_{h+1}$. First,

$$\begin{aligned}
\|w_h^{k,i}\| &= \left\|\widehat{\theta}_h^k + \zeta_h^{k,i} + {\Lambda_h^k}^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \hat{V}_{h+1}^{k,i}(x_{h+1}^\tau)\right\| \\
&\leq \left\|(\Lambda^k)^{-1}\right\| \sum_{\tau=1}^{k-1} \|\phi^\tau\| |v^\tau| + \|\zeta^{k,i}\| + \left\|(\Lambda_h^k)^{-1}\right\| \sum_{\tau=1}^{k-1} \|\phi_h^\tau\| |\hat{V}_{h+1}^{k,i}(x_{h+1}^\tau)| \\
&\leq HK + (\|\zeta^{k,i}\|_{\Lambda^k}/\sqrt{H}) + KH\beta_{r\zeta} && (\|\phi_h^\tau\| \leq 1, \Lambda_h^k \succeq I, \Lambda^k \succeq HI, \|\hat{V}_{h+1}^{k,i}\|_\infty \leq H\beta_{r\zeta}) \\
&\leq HK + (\beta_r\beta_\zeta/\sqrt{H}) + KH\beta_{r\zeta} && \text{(Eq. (13))} \\
&\leq 4KH\beta_{r\zeta},
\end{aligned}$$

where the last transition is due to our parameter choices. Finally, we have that

$$\begin{aligned}
\beta_p &= 20\beta_{r\zeta} dH\sqrt{\log(163KdH/\delta)} \\
&\leq 20\beta_{r\zeta} dH\sqrt{2Kd + \log(82H/\delta)} && (\log(x) \leq x) \\
&\leq 20\beta_{r\zeta} dH\sqrt{2Kd\log(82H/\delta)} && (a + b \leq ab, \forall a, b \geq 2) \\
&\leq 20\beta_{r\zeta} d^{7/4} HK\sqrt{8\log(4H/\delta)} && (H, d, K \geq 1) \\
&\leq 876\beta_{r\zeta} d^{7/4} HK\log(5H^2/\delta) \\
&= \bar{\beta},
\end{aligned}$$

as desired. ■

B. Randomized Ensemble Policy Optimization (REPO) with Reward-Free Warm-Up

B.1. Proof of Theorem 2

We begin by defining a so-called "good event", followed by optimism, cost of optimism, Ensemble Hedging cost, and Policy Optimization cost. We conclude with the proof of Theorem 2.

Good event. Define the truncated features $\bar{\phi}_h(x,a) = \mathbb{1}_{\{x \in \mathcal{Z}_h\}}\phi(x,a)$ and their concatenation $\bar{\phi}(x_{1:H}, a_{1:H}) = (\bar{\phi}_1(x_1,a_1)^\mathsf{T}, \dots, \bar{\phi}_H(x_H,a_H)^\mathsf{T})^\mathsf{T}$. We also define the expected truncated feature occupancy of a policy π as $\bar{\phi}^\pi = \mathbb{E}_{P,\pi}\bar{\phi}(x_{1:H}, a_{1:H})$. In addition, to simplify presentation, we denote $\zeta^{k,i} = \zeta_r^{k,i} + \zeta_p^{k,i}$ where $\zeta_r^{k,i} \sim \mathcal{N}\Big(0, 2\beta_r^2(\Lambda^k)^{-1}\Big)$ and $\zeta_p^{k,i} \sim \mathcal{N}\Big(0, 2H\beta_p^2 \mathrm{diag}\big(\Lambda_1^k, \dots, \Lambda_H^k\big)^{-1}\Big)$. We define the following good event $E_g = \bigcap_{i=1}^7 E_i$, over which the regret is deterministically bounded:

$$E_1 = \Big\{\forall e \in [E], k \in K_e : \|\theta - \widehat{\theta}^k\|_{\Lambda^k} \le \beta_r\Big\}; \tag{16}$$

$$E_2 = \Big\{\forall e \in [E], k \in K_e, i \in [m], h \in [H] : \|(\psi_h - \widehat{\psi}_h^k)\hat{V}_{h+1}^{k,i}\|_{\Lambda_h^k} \le \beta_p, \|\hat{Q}_{h+1}^{k,i}\|_\infty \le \beta_Q\Big\}; \tag{17}$$

$$E_3 = \Big\{\forall e \in [E], i \in [m], h \in [H] : \|\zeta_r^{k_e,i}\|_{\Lambda^{k_e}} \le \beta_\zeta\beta_r, \|\zeta_{p,h}^{k_e,i}\|_{\Lambda_h^{k_e}} \le \beta_\zeta\beta_p\Big\}; \tag{18}$$

$$E_4 = \left\{\forall e \in [E] : \max_{i \in [m]} (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,i} \ge \|\bar{\phi}^{\pi^\star}\|_{\Sigma_\zeta^{k_e}}\right\}; \tag{19}$$

$$E_5 = \{\forall \pi \in \Pi_M, h \in [H] : \mathbb{P}_{P,\pi}[x_h \notin \mathcal{Z}_h] \le \epsilon_{\text{cov}}\}; \tag{20}$$

$$E_6 = \left\{\sum_{k=k_0}^K \mathbb{E}_{P,\pi^k}[Y_k] \le \sum_{k=k_0}^K Y_k + (1+\sqrt{2}\beta_\zeta)(\beta_r + H\beta_p)\sqrt{2K\log\frac{7}{\delta}}\right\}. \tag{21}$$

$$E_7 = \left\{\sum_{k=k_0}^K \sum_{i=1}^m p^k(i)\hat{V}_1^{k,i}(x_1) \le \sum_{k=k_0}^K \hat{V}_1^{k,i_k}(x_1) + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}\right\}; \tag{22}$$

where $Y_k = (1+\sqrt{2}\beta_\zeta)\Big[\beta_r\|\phi^k\|_{(\Lambda^k)^{-1}} + \beta_p\sum_{h\in[H]}\|\phi_h^k\|_{(\Lambda_h^k)^{-1}}\Big]$.

Lemma 8 (Good event). *Consider the following parameter setting:*

$$\lambda_p = 1, \lambda_r = H, m = 9\log(7K/\delta), \beta_r = 2H\sqrt{2dH\log(14K/\delta)}, \beta_\zeta = \sqrt{11dH\log(14mHK/\delta)},$$
$$\beta_p = 216\beta_\zeta\beta_r\sqrt{dH\log\frac{60K^3H\beta_Q}{\delta\sqrt{d}}}, \beta_Q = 18\beta_\zeta\beta_r\sqrt{H}, \beta_w = 36\beta_\zeta\sqrt{d\log\frac{60K^3H\beta_Q}{\delta\sqrt{d}}}, \eta_o \le 1, \epsilon_{cov} \ge 1/K.$$

Then $\Pr[E_g] \ge 1-\delta$.

Proof in Appendix B.2.

Optimism and its cost. We start with a result that bounds the cost of transitioning to the truncated state-action features.

Lemma 9. *Suppose that the good event E_g holds (Eqs. (16) to (22)), then for all $v \in \mathbb{R}^d, h \in [H], \pi \in \Pi_M$, we have*

$$\big|\mathbb{E}_{P,\pi}\big[(\phi(x_h,a_h) - \bar{\phi}_h(x_h,a_h))^\mathsf{T}v\big]\big| \le C\epsilon_{cov},$$

where $C = \max_{x,a}|\phi(x,a)^\mathsf{T}v|$. Additionally, if $v = \psi_h\hat{V}_{h+1}^{k,i}$ then $C \le \beta_Q$, and if $v = \widehat{\theta}_h^k + \zeta_{r,h}^{k_e,i}$ then $C \le 2\beta_r\beta_\zeta/\sqrt{H}$.

Proof. We have that

$$\big|\mathbb{E}_{P,\pi}\big[(\phi(x_h,a_h) - \bar{\phi}_h(x_h,a_h))^\mathsf{T}v\big]\big| = \big|\mathbb{E}_{P,\pi}\big[(1 - \mathbb{1}_{\{x_h \in \mathcal{Z}_h\}})\phi(x_h,a_h)^\mathsf{T}v\big]\big| \le C\mathbb{P}_{P,\pi}[x_h \notin \mathcal{Z}_h] \le C\epsilon_{\text{cov}},$$

where the last inequality is by Eq. (20). Now, if $v = \psi_h\hat{V}_{h+1}^{k,i}$ then

$$C = \max_{x,a}|\phi(x,a)^\mathsf{T}\psi_h\hat{V}_{h+1}^{k,i}| \le \max_{x,a}\|\phi(x,a)^\mathsf{T}\psi_h\|_1\|\hat{V}_{h+1}^{k,i}\| = \|\hat{V}_{h+1}^{k,i}\| \le \beta_Q,$$

where the last equality used that $\phi(x,a)^\mathsf{T}\psi_h$ is a distribution over $\mathcal{X}$, thus has ℓ_1-norm of 1 and the last inequality used Eq. (17). Finally, if $v = \widehat{\theta}^k_h + \zeta^{k_e,i}_{r,h}$ then using Cauchy-Schwarz

$$C \le \max_{x,a} \|\phi(x,a)\| \|\widehat{\theta}^k_h + \zeta^{k_e,i}_{r,h}\| \le \|\theta_h\| + \|\widehat{\theta}^k - \theta\| + \|\zeta^{k_e,i}_r\| \le \sqrt{d} + \frac{\beta_r(1+\beta_\zeta)}{\sqrt{H}} \le \frac{2\beta_r\beta_\zeta}{\sqrt{H}},$$

where the second inequality used that $\|\phi(x,a)\| \le 1$ and the triangle-inequality, the third used $\Lambda^k \succeq HI, \|\theta_h\| \le \sqrt{d}$ and Eqs. (16) and (18), and the fourth used $\beta_r \ge \sqrt{dH}, \beta_\zeta \ge 3$. ■

Lemma 10 (Optimism). *Suppose that the good event E_g holds (Eqs. (16) to (22)) and define $\hat{j}_e = \arg\max_{i\in[m]} (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,i}$, $e \in [E]$, then for all $e \in [E], k \in K_e$*

$$\sum_{h\in[H]} \mathbb{E}_{P,\pi^\star}\Big[\phi(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}^{k,\hat{j}_e}_{h+1}) - \hat{Q}^{k,\hat{j}_e}_h(x_h,a_h)\Big] \le \frac{9}{8}\epsilon_{cov} H\beta_Q \quad ,\forall k \in [K].$$

Proof. Define $\bar{\phi}^\pi_h = \mathbb{E}_{P,\pi}\bar{\phi}_h(x_h,a_h)$ and recall that $\bar{\phi}^\pi = ((\bar{\phi}^\pi_1)^\mathsf{T}, \dots, (\bar{\phi}^\pi_H)^\mathsf{T})^\mathsf{T}$ Then we have that

$$\begin{aligned}
&\sum_{h\in[H]} \mathbb{E}_{P,\pi^\star}\Big[\phi(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}^{k,\hat{j}_e}_{h+1}) - \hat{Q}^{k,\hat{j}_e}_h(x_h,a_h)\Big] \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \sum_{h\in[H]} \mathbb{E}_{P,\pi^\star}\Big[\bar{\phi}_h(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}^{k,\hat{j}_e}_{h+1}) - \hat{Q}^{k,\hat{j}_e}_h(x_h,a_h)\Big] && \text{(Lemma 9)} \\
&= \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \sum_{h\in[H]} \mathbb{E}_{P,\pi^\star}\big[\bar{\phi}_h(x_h,a_h)\big]^\mathsf{T}\Big(\theta_h - \widehat{\theta}^k_h - \zeta^{k_e,\hat{j}_e}_h + (\psi_h - \widehat{\psi}^k_h)\hat{V}^{k,\hat{j}_e}_{h+1}\Big) \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \beta_r\|\bar{\phi}^{\pi^\star}\|_{(\Lambda^{k_e})^{-1}} + \sum_{h\in[H]} \beta_p \|\bar{\phi}^{\pi^\star}_h\|_{(\Lambda^{k_e}_h)^{-1}} - (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} && \text{(Eqs. (16) and (17), } \Lambda^{k_e} \preceq \Lambda^k, \Lambda^{k_e}_h \preceq \Lambda^k_h) \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \|\bar{\phi}^{\pi^\star}\|_{\beta_r^2(\Lambda^{k_e})^{-1}} + \|\bar{\phi}^{\pi^\star}\|_{H\beta_p^2 \text{diag}\left(\Lambda^{k_e}_1,\dots,\Lambda^{k_e}_H\right)^{-1}} - (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} && \text{(Cauchy-Schwarz)} \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \|\bar{\phi}^{\pi^\star}\|_{2\beta_r^2(\Lambda^{k_e})^{-1} + 2H\beta_p^2 \text{diag}\left(\Lambda^{k_e}_1,\dots,\Lambda^{k_e}_H\right)^{-1}} - (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} && \text{(Cauchy-Schwarz)} \\
&= \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \|\bar{\phi}^{\pi^\star}\|_{\Sigma^{k_e}_\zeta} - (\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q, && \text{(Eq. (19))}
\end{aligned}$$

as desired. ■

Lemma 11 (Cost of optimism). *Suppose that the good event E_g holds (Eqs. (16) to (22)), then $\forall e \in [E], k \in K_e$*

$$\hat{V}^{k,i_k}_1(x_1) - V^{\pi^k}_1(x_1) \le \frac{9}{8}\epsilon_{cov} H\beta_Q + (1+\sqrt{2}\beta_\zeta)\mathbb{E}_{P,\pi^k}\left[\beta_r\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \beta_p \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda^k_h)^{-1}}\right].$$

Proof. By Lemma 25, a value difference lemma by Shani et al. (2020),

$$\begin{aligned}
\hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1) &= \sum_{h\in[H]} \mathbb{E}_{P,\pi^k}\left[\hat{Q}_k^{k,i_k}(x_h,a_h) - \phi(x_h,a_h)^\mathsf{T}\left(\theta_h + \psi_h \hat{V}_{h+1}^{k,i_k}\right)\right] \\
&= \sum_{h\in[H]} \mathbb{E}_{P,\pi^k}\left[\bar{\phi}_h(x_h,a_h)^\mathsf{T}\left(\widehat{\theta}_h^k + \zeta_{r,h}^{k_e,i_k} + \zeta_{p,h}^{k_e,i_k} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i_k}\right) - \phi(x_h,a_h)^\mathsf{T}\left(\theta_h + \psi_h \hat{V}_{h+1}^{k,i_k}\right)\right] \\
&\le \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \mathbb{E}_{P,\pi^k}\left[\sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}\left(\widehat{\theta}_h^k - \theta_h + \zeta_{r,h}^{k_e,i_k}\right) + \sum_{h\in[H]} \bar{\phi}_h(x_h,a_h)^\mathsf{T}\left(\zeta_{p,h}^{k_e,i_k} + (\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i_k}\right)\right] \\
&\qquad \text{(Lemma 9)} \\
&= \frac{9}{8}\epsilon_{\text{cov}} H\beta_Q + \mathbb{E}_{P,\pi^k}\left[\underbrace{\phi(x_{1:H},a_{1:H})^\mathsf{T}(\widehat{\theta}^k - \theta + \zeta_r^{k_e,i_k})}_{(i)} + \sum_{h\in[H]} \underbrace{\bar{\phi}_h(x_h,a_h)^\mathsf{T}\left(\zeta_{p,h}^{k_e,i_k} + (\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i_k}\right)}_{(ii)}\right].
\end{aligned}$$

We conclude the proof by bounding (i), (ii).

$$\begin{aligned}
(i) &\le \beta_r \|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \beta_r\beta_\zeta \|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^{k_e})^{-1}} && \text{(Eqs. (16) and (18))} \\
&\le \beta_r(1+\sqrt{2}\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}}. && \text{(Lemma 27)}
\end{aligned}$$

$$\begin{aligned}
(ii) &\le \beta_p \|\bar{\phi}_h(x_h,a_h)\|_{(\Lambda_h^k)^{-1}} + \beta_p\beta_\zeta \|\bar{\phi}_h(x_h,a_h)\|_{(\Lambda_h^{k_e})^{-1}} && \text{(Eqs. (17) and (18))} \\
&\le \beta_p(1+\sqrt{2}\beta_\zeta)\|\bar{\phi}_h(x_h,a_h)\|_{(\Lambda_h^k)^{-1}} && \text{(Lemma 27)} \\
&\le \beta_p(1+\sqrt{2}\beta_\zeta)\|\phi_h(x_h,a_h)\|_{(\Lambda_h^k)^{-1}},
\end{aligned}$$

where the last inequality used that for $x_h \in \mathcal{Z}_h$, $\phi(x_h,a_h) = \bar{\phi}_h(x_h,a_h)$ and for $x_h \notin \mathcal{Z}_h$ $\|\bar{\phi}_h(x_h,a_h)\|_{(\Lambda_h^k)^{-1}} = 0 \le \|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}$. ■

Hedge over the ensemble. We use standard online mirror arguments to bound the regret against each single $j \in [m]$ in every epoch $e \in [E]$.

Lemma 12 (Hedge). *For every epoch $e \in [E]$, let $j_e \in [m]$. Suppose that the good event E_g holds (Eqs. (16) to (22)) and set $\eta_x \le 1/\beta_Q$, then*

$$\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,j_e}(x_1) - \hat{V}_1^{k,i_k}(x_1) \le \frac{E\log m}{\eta_x} + \eta_x K\beta_Q^2 + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}.$$

Proof. Note that the distribution $p^k(\cdot)$ is reset in the beginning of every epoch. Then, the lemma follows by first using Eq. (22) and then applying Lemma 24 for each epoch individually with $y_t(a) = -\hat{V}_1^{k,i}(x_1), x_t(a) = p^k(i)$ and noting that $|\hat{V}_1^{k,i}(x_1)| \le \beta_Q$ by Eq. (17). ■

Policy online mirror descent. We use standard online mirror descent arguments to bound the local regret in each state.

Lemma 13 (OMD). *Suppose that the good event E_g holds (Eqs. (16) to (22)) and set $\eta_o \le 1/\beta_Q$, then*

$$\sum_{k\in K_e}\sum_{a\in\mathcal{A}} \hat{Q}_h^{k,i}(x,a)(\pi_h^\star(a \mid x) - \pi_h^{k,i}(a \mid x)) \le \frac{\log|\mathcal{A}|}{\eta_o} + \eta_o \sum_{k\in K_e} \beta_Q^2 \quad ,\forall e \in [E], i \in [m], h \in [H], x \in \mathcal{X}.$$

Proof. Note that the policy $\pi^{k,i}$ is reset in the beginning of every epoch. Then, the lemma follows directly by Lemma 24 with $y_t(a) = -\hat{Q}_h^{k,i}(x,a), x_t(a) = \pi_h^{k,i}(a \mid x)$ and noting that $|\hat{Q}_h^{k,i}(x,a)| \le \beta_Q$ by Eq. (17). ■

Epoch schedule. The algorithm operates in epochs. At the beginning of each epoch, the noise is re-sampled and the policies are reset to be uniformly random. We denote the total number of epochs by E, the first episode within epoch e by k_e, and the set of episodes within epoch e by K_e. The following lemma bounds the number of epochs.

Lemma 14. *The number of epochs E is bounded by $3dH\log(2K)$.*

Proof. Let $\mathcal{T}_h = \{e_h^1, e_h^2, \ldots\}$ be the epochs where the condition $\det(\Lambda_h^k) \geq 2\det(\Lambda_h^{k_e})$ was triggered in Line 5 of Algorithm 3. Then we have that

$$\det(\Lambda_h^{k_e}) \geq \begin{cases} 2\det(\Lambda_h^{k_{e-1}}) & , e \in \mathcal{T}_h \\ \det(\Lambda_h^{k_{e-1}}) & , \text{otherwise.} \end{cases}$$

Unrolling this relation, we get that

$$\det(\Lambda_h^K) \geq 2^{|\mathcal{T}_h|-1}\det I = 2^{|\mathcal{T}_h|-1},$$

and changing sides, and taking the logarithm we get that

$$\begin{aligned} |\mathcal{T}_h| &\leq 1 + \log_2\det\big(\Lambda_h^K\big) \\ &\leq 1 + d\log_2\|\Lambda_h^K\| && (\det(A) \leq \|A\|^d) \\ &\leq 1 + d\log_2\left(1 + \sum_{k=1}^{K-1}\|\phi_h^k\|^2\right) && \text{(triangle inequality)} \\ &\leq 1 + d\log_2 K && (\|\phi_h^k\| \leq 1) \\ &\leq (3/2)d\log 2K. \end{aligned}$$

Similarly, if $\bar{\mathcal{T}}$ are the epochs where the condition $\det(\Lambda^k) \geq 2\det(\Lambda^{k_e})$ was triggered, the $|\bar{\mathcal{T}}| \leq (3/2)dH\log 2K$. We conclude that

$$E = |\bar{\mathcal{T}} \cup \big(\cup_{h\in[H]}\mathcal{T}_h\big)| \leq |\bar{\mathcal{T}}| + \sum_{h\in[H]}|\mathcal{T}_h| \leq 3dH\log(2K). \qquad \blacksquare$$

Regret bound.

Theorem 15 (restatement of Theorem 2). *Suppose that we run Algorithm 3 with*

$$\eta_x = \sqrt{\frac{3dH\log(2K)\log m}{K\beta_Q^2}}, \eta_o = \sqrt{\frac{3dH\log(2K)\log|\mathcal{A}|}{K\beta_Q^2}}, \epsilon_{cov} = \frac{2d^2H^2}{3\sqrt{C\beta_Q K}}\log^4\left(\frac{28H^2K\beta_w^2}{\delta}\right),$$

where $C > 0$ is a universal constant defined in Lemma 17, and the other parameters as in Lemma 8. Then with probability at least $1-\delta$ we incur regret at most

$$\text{Regret} \leq 39\sqrt{Cd^5H^9K\log^9\left(\frac{28H^2K\beta_w^2}{\delta}\right)} + 67204\sqrt{d^5H^8K\log^5\frac{60K^3|\mathcal{A}|Hm\beta_Q}{\delta}} = \tilde{O}\left(\sqrt{d^5H^9K\log^9\frac{|\mathcal{A}|}{\delta}}\right).$$

Proof. Suppose that the good event E_g holds (Eqs. (16) to (22)). By Lemma 8, this holds with probability at least $1-\delta$. For every epoch $e \in [E]$, let $\hat{j}_e = \arg\max_{i\in[m]}(\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,i}$. Now decompose the regret using Lemma 25, an extended value difference lemma by Shani et al. (2020).

$$
\begin{aligned}
\text{Regret} &= \sum_{k\in[K]} V_1^{\pi^\star}(x_1) - V_1^{\pi^k}(x_1) \\
&\le Hk_0 + \sum_{k=k_0}^{K} V_1^{\pi^\star}(x_1) - V_1^{\pi^k}(x_1) \\
&= Hk_0 + \underbrace{\sum_{k=k_0}^{K} \hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1)}_{(i)} + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,\hat{j}_e}(x_1) - \hat{V}_1^{k,i_k}(x_1)}_{(ii)} \\
&\quad + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e}\sum_{h\in[H]}\sum_{a\in\mathcal{A}} \mathbb{E}_{P,\pi^\star}\Big[\hat{Q}_h^{k,\hat{j}_e}(x_h,a)(\pi_h^\star(a\mid x_h) - \pi_h^{k,\hat{j}_e}(a\mid x_h))\Big]}_{(iii)} \\
&\quad + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e}\sum_{h\in[H]} \mathbb{E}_{P,\pi^\star}\Big[\phi(x_h,a_h)^\mathsf{T}(\theta_h + \psi_h \hat{V}_{h+1}^{k,\hat{j}_e}) - \hat{Q}_h^{k,\hat{j}_e}(x_h,a_h)\Big]}_{(iv)}.
\end{aligned}
$$

By Lemma 17, for $\beta_w = \tilde{O}(d\sqrt{H})$ and $\epsilon_{\text{cov}} \ge 1/K$, we have that $k_0 \le C\cdot\left(\frac{d^4H^4}{\epsilon_{\text{cov}}}\log^8\left(\frac{28H^2K\beta_w^2}{\delta}\right)\right)$. By Lemma 10 $(iv) \le \frac{9}{8}\epsilon_{\text{cov}}H\beta_Q K$. By Lemmas 12 and 14 (with our choice of η_x), we have that

$$
(ii) \le 2\beta_Q\sqrt{3KdH\log(m)\log(2K)} + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}} \le 8\beta_Q\sqrt{KdH}\log\left(\frac{7Km}{\delta}\right).
$$

Similarly, by Lemmas 13 and 14 (with our choice of η_o) we have

$$
(iii) \le \sum_{h\in[H]}\sum_{e\in[E]} \mathbb{E}_{P,\pi^\star}\left[\frac{\log A}{\eta_o} + \eta_o\sum_{k\in K_e}\beta_Q^2\right] \le 4H\beta_Q\sqrt{KdH\log(2K)\log|\mathcal{A}|}.
$$

For term (i), we use Lemma 11 as follows.

$$
\begin{aligned}
(i) &\le \frac{9}{8}\epsilon_{\text{cov}}H\beta_Q K + (1+\sqrt{2}\beta_\zeta)\sum_{k\in[K]}\mathbb{E}_{P,\pi^k}\left[\beta_r\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \beta_p\sum_{h\in[H]}\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\
&\le \frac{9}{8}\epsilon_{\text{cov}}H\beta_Q K + \beta_r(1+\sqrt{2}\beta_\zeta)\sum_{k\in[K]}\|\phi^k\|_{(\Lambda^k)^{-1}} + \beta_p(1+\sqrt{2}\beta_\zeta)\sum_{h\in[H]}\sum_{k\in[K]}\|\phi_h^k\|_{(\Lambda_h^k)^{-1}} && \text{(Eq. (21))} \\
&\quad + (1+\sqrt{2}\beta_\zeta)(\beta_r + H\beta_p)\sqrt{2K\log\frac{7}{\delta}} \\
&\le \frac{9}{8}\epsilon_{\text{cov}}H\beta_Q K + (1+\sqrt{2}\beta_\zeta)(\beta_r\sqrt{H} + \beta_p H)\sqrt{2Kd\log(2K)} + (1+\sqrt{2}\beta_\zeta)(\beta_r + H\beta_p)\sqrt{2K\log\frac{7}{\delta}} && \text{(Lemma 26)} \\
&\le \frac{9}{8}\epsilon_{\text{cov}}H\beta_Q K + 8\beta_p\beta_\zeta H\sqrt{Kd\log\frac{7K}{\delta}}. && (\beta_\zeta \ge 3, \beta_p \ge 216\beta_r)
\end{aligned}
$$

Putting all bounds together, we get that

$$\begin{aligned}
\mathrm{Regret} &\le C \cdot \left(\frac{d^4 H^5}{\epsilon_{\text{cov}}} \log^8\left(\frac{28H^2 K \beta_w^2}{\delta}\right)\right) + \frac{9}{4}\epsilon_{\text{cov}} H \beta_Q K + 8\beta_p \beta_\zeta H \sqrt{K d \log \frac{7K}{\delta}} \\
&+ 8\beta_Q \sqrt{KdH} \log\left(\frac{7Km}{\delta}\right) + 4H\beta_Q \sqrt{KdH \log(2K) \log|\mathcal{A}|} \\
&\le 3d^2 H^3 \log^4\left(\frac{28H^2 K \beta_w^2}{\delta}\right) \sqrt{C\beta_Q K} + 12H\beta_Q \sqrt{KdH} \log\left(\frac{7K|\mathcal{A}|m}{\delta}\right) \\
&+ 108\beta_Q \beta_\zeta H d \sqrt{K} \log \frac{60K^3 H \beta_Q}{\delta\sqrt{d}} \\
&\le 3d^2 H^3 \log^4\left(\frac{28H^2 K \beta_w^2}{\delta}\right) \sqrt{C\beta_Q K} + 120\beta_Q \beta_\zeta H d \sqrt{K} \log \frac{60K^3 |\mathcal{A}| H m \beta_Q}{\delta} \\
&\le 3d^2 H^3 \log^4\left(\frac{28H^2 K \beta_w^2}{\delta}\right) \sqrt{18C\beta_r \beta_\zeta H K} + 2160\beta_r \beta_\zeta^2 d \sqrt{H^3 K} \log \frac{60K^3 |\mathcal{A}| H m \beta_Q}{\delta} \\
&\le 39\sqrt{C d^5 H^9 K \log^9\left(\frac{28H^2 K \beta_w^2}{\delta}\right)} + 67204\sqrt{d^5 H^8 K \log^5 \frac{60K^3 |\mathcal{A}| H m \beta_Q}{\delta}},
\end{aligned}$$

where the last transition also used that

$$\beta_r \beta_\zeta^2 \le 22\sqrt{2d^3 H^5 \log^3(14mHK/\delta)}, \beta_r \beta_\zeta \le \sqrt{88} d H^2 \log(14mHK/\delta). \qquad \blacksquare$$

B.2. Proofs of good event (REPO)

We begin by defining function classes and properties necessary for the uniform convergence arguments over the value functions. We then proceed to define a proxy good event, whose high probability occurrence is straightforward to prove. We then show that the proxy event implies the desired good event.

Value and policy classes. We define the following class of restricted Q-functions:

$$\widehat{\mathcal{Q}}(\mathcal{Z}, W, C) = \left\{\hat{Q}(\cdot, \cdot; w, \mathcal{Z}) \mid \|w\| \le W, \|\hat{Q}(\cdot, \cdot; w, \mathcal{Z})\|_\infty \le C\right\},$$

where $\hat{Q}(x, a; w, \mathcal{Z}) = w^\mathsf{T} \phi(x, a) \cdot \mathbb{1}_{\{x \in \mathcal{Z}\}}$. Next, we define the following class of soft-max policies:

$$\Pi(\mathcal{Z}, W) = \{\pi(\cdot \mid \cdot; w, \mathcal{Z}) \mid \|w\| \le W\},$$

where $\pi(a \mid x; w, \mathcal{Z}) = \frac{\exp(\hat{Q}(x,a;w,\mathcal{Z}))}{\sum_{a' \in \mathcal{A}} \exp(\hat{Q}(x,a';w,\mathcal{Z}))}$. Finally, we define the following class of restricted value functions:

$$\widehat{\mathcal{V}}(\mathcal{Z}, W, C) = \left\{\hat{V}(\cdot; \pi, \hat{Q}) \mid \pi \in \Pi(\mathcal{Z}, WK), \hat{Q} \in \widehat{\mathcal{Q}}(\mathcal{Z}, W, C)\right\}, \tag{23}$$

where $\hat{V}(x; \pi, \hat{Q}) = \sum_{a \in \mathcal{A}} \pi(a \mid x) \hat{Q}(x, a)$. The following lemma provides the bound on the covering number of the value function class defined above (see proof in Appendix B.2).

Lemma 16. *For any $\epsilon, W > 0, C \ge 1$ and $\mathcal{Z} \subseteq \mathcal{X}$ we have $\log \mathcal{N}_\epsilon\left(\widehat{\mathcal{V}}(\mathcal{Z}, W, C)\right) \le 2d \log(1 + 7WCK/\epsilon)$ where $\mathcal{N}_\epsilon$ is the covering number of a class in supremum distance.*

Proof. First, notice that our policy class $\Pi(\mathcal{Z}, W)$ fits Lemma 12 of (Sherman et al., 2023) with $f_\theta(y) = y^\mathsf{T} \theta \cdot \mathbb{1}_{\{y \in \mathcal{Z}\}}$. Since $\|y\| = \|\phi(x, a)\| \le 1$, f_θ is $1-$Lipschitz with respect to θ and thus the policy class is $2-$Lipschitz, in ℓ_1-norm, i.e.,

$$\|\pi(\cdot \mid x; w) - \pi(\cdot \mid x; w')\|_1 \le 2\|w - w'\|.$$

Similarly, $\widehat{\mathcal{Q}}(\mathcal{Z}, W, C)$ is $1-$Lipschitz in supremum norm, i.e.,

$$\|\hat{Q}(\cdot, \cdot; w) - \hat{Q}'(\cdot, \cdot; w')\|_\infty \le \|w - w'\|.$$

Now, let $V, V' \in \widehat{\mathcal{V}}(\mathcal{Z}, W, C)$ and $w = (w_1, w_2), w' = (w_1', w_2') \in \mathbb{R}^{2d}$ be their respective parameters. We have that

$$|V(x;\pi,\hat{Q}) - V(x;\pi',\hat{Q}')| \le \underbrace{|V(x;\pi,\hat{Q}) - V(x;\pi,\hat{Q}')|}_{(i)} + \underbrace{|V(x;\pi,\hat{Q}') - V(x;\pi',\hat{Q}')|}_{(ii)}.$$

For the first term

$$\begin{aligned}
(i) &= \left|\sum_{a\in\mathcal{A}} \pi(a \mid x)(\hat{Q}(x,a;w_2) - \hat{Q}(x,a;w_2'))\right| \\
&\le \sum_{a\in\mathcal{A}} \pi(a \mid x)\left|\hat{Q}(x,a;w_2) - \hat{Q}(x,a;w_2')\right| && \text{(triangle inequality)} \\
&\le \sum_{a\in\mathcal{A}} \pi(a \mid x)\|w_2 - w_2'\| && \text{($\hat{Q}$ is 1-Lipschitz)} \\
&= \|w_2 - w_2'\|.
\end{aligned}$$

For the second term

$$(ii) = \left|\sum_{a\in\mathcal{A}} \hat{Q}'(x,a)(\pi(a \mid x; w_1) - \pi(a \mid x; w_1'))\right| \le C\|\pi(\cdot \mid x; w_1) - \pi(\cdot \mid x; w_1')\|_1 \le 2C\|w_1 - w_1'\|,$$

where the first transition used that $\|Q\|_\infty \le C$ for all $Q \in \widehat{\mathcal{Q}}(\mathcal{Z}, W, C)$ and the second used the Lipschitz property of the policy class shown above. Combining the terms we get that

$$|V(x;\pi,\hat{Q}) - V(x;\pi',\hat{Q}')| \le \|w_2 - w_2'\| + 2C\|w_1 - w_1'\| \le \sqrt{1+4C^2}\|w - w'\|,$$

implying that $\widehat{\mathcal{V}}(\mathcal{Z}, W, C)$ is $\sqrt{1+4C^2}$−Lipschitz in supremum norm. Finally, notice that $\|w\| = \sqrt{\|w_1\|^2 + \|w_2\|^2} \le \sqrt{2}KW$. Applying Lemma 35 together with our assumption that $C \ge 1$ concludes the proof. ■

Proxy good event. We first need the following result regarding the reward free exploration.

Lemma 17 (Lemma 16 in (Sherman et al., 2023)). *Assume we run Algorithm 2 of (Sherman et al., 2023) with* $\epsilon_{cov} \ge 1/K$*. Then it will terminate after* $C \cdot \left(\frac{dH^3}{\epsilon_{cov}} \max\{\beta_w^2, d^3H\} \log^8\left(\frac{28H^2K\beta_w^2}{\delta}\right)\right)$ *episodes where* $C > 0$ *is a numerical constant, and with probability at least* $1 - \delta/7$*, outputs* $\Lambda_h^0, h \in [H]$ *such that*

$$\forall \pi \in \Pi_M, h \in [H] : \mathbb{P}_{P,\pi}[x_h \notin \mathcal{Z}_h] \le \epsilon_{cov}$$

where $\mathcal{Z}_h = \left\{x \in \mathcal{X} \mid \forall a, \|\phi(x,a)\|_{(\Lambda_h^0)^{-1}} \le 1/(2\beta_w H)\right\}$.

Next, we define a proxy good event $\bar{E}_g = E_1 \cap \bar{E}_2 \cap E_3 \cap E_4 \cap E_5 \cap E_6 \cap \bar{E}_7$ where

$$\bar{E}_2 = \left\{\forall k \in [K], h \in [H], V \in \widehat{\mathcal{V}}(\mathcal{Z}_{h+1}, W, \beta_{Q,h+1}) : \|(\psi_h - \widehat{\psi}_h^k)V\|_{\Lambda_h^k} \le \beta_{p,h}\right\} \tag{24}$$

$$\bar{E}_7 = \left\{\sum_{k=k_0}^{K}\sum_{i=1}^{m} p^k(i)\mathrm{clip}_{\beta_Q}\left[\hat{V}_1^{k,i}(x_1)\right] \le \sum_{k=k_0}^{K} \mathrm{clip}_{\beta_Q}\left[\hat{V}_1^{k,i_k}(x_1)\right] + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}\right\}, \tag{25}$$

and $\beta_{Q,h}, \beta_{p,h}, h \in [H]$ are such that $\beta_{p,h} = 12\beta_{Q,h+1}\sqrt{d\log\frac{15K^2HW}{\delta\sqrt{d}}}$, $W = 4K\beta_Q$ and

$$\beta_{Q,h} = \beta_{Q,H} + \left(1 + \frac{1}{H}\right)\beta_{Q,h+1}, \quad \beta_{Q,H} = \frac{6\beta_\zeta\beta_r}{\sqrt{H}}, \quad \beta_{Q,H+1} = 0. \tag{26}$$

We have the following result for the proxy good event.

Lemma 18 (Proxy good event). *Consider the parameter setting of Lemma 8. Then* $\Pr[\bar{E}_g] \ge 1 - \delta$*.*

Proof. First, by Lemma 31 and our choice of parameters, E_1 (Eq. (16)) holds with probability at least $1-\delta/7$. Next, as explained in Sherman et al. (2023, Lemma 7), the set $\mathcal{Z}_{h+1}$, which determines the value function class, is independent of the warm-up dataset $\mathcal{D}^0_h$ used in the dynamics estimator $\widehat{\psi}^k_h$. This is because their warm-up routine runs independent algorithms for each $h \in [H]$. As a result, conditioning on $\mathcal{Z}_{h+1}$ does not break the martingale structure of the self-normalized process (Lemma 30) and we can apply Lemmas 16 and 32 to get that with probability at least $1-\delta/7$ simultaneously for all $k \in [K], h \in [H], V \in \widehat{\mathcal{V}}(\mathcal{Z}_{h+1}, W, \beta_{Q,h+1})$

$$\begin{aligned}
\|(\psi_h - \widehat{\psi}^k_h)V\|_{\Lambda^k_h} &\le 4\beta_{Q,h+1}\sqrt{d\log(K+1) + 2\log(5H/\delta) + 4d\log(1 + 14K^2W\beta_Q/\beta_Q\sqrt{d})} \\
&\le 4\beta_{Q,h+1}\sqrt{d\log(K+1) + 2\log(5H/\delta) + 4d\log\frac{15K^2W}{\sqrt{d}}} \\
&\le 12\beta_{Q,h+1}\sqrt{d\log\frac{15K^2HW}{\delta\sqrt{d}}} \\
&= \beta_{p,h},
\end{aligned}$$

implying $\bar{E}_2$ (Eq. (24)). Now, suppose that the noise is generated such that $\zeta^{k_e,i}_{p,h} = \sqrt{2H\beta_p^2}(\Lambda^{k_e}_h)^{-1/2}g^{e,i}_h$ where $g^{e,i}_{p,h} \sim \mathcal{N}(0, I_d)$ are i.i.d for all $e \in [E], i \in [m], h \in [H]$. Indeed, notice that

$$\mathbb{E}(\zeta^{k_e,i}_{p,h})(\zeta^{k_e,i}_{p,h})^\mathsf{T} = 2H\beta_p^2(\Lambda^{k_e}_h)^{-1/2}\mathbb{E}\Big[(g^{e,i}_{p,h})(g^{e,i}_{p,h})^\mathsf{T}\Big](\Lambda^{k_e}_h)^{-1/2} = 2H\beta_p^2(\Lambda^{k_e}_h)^{-1}.$$

Taking a union bound over Lemma 29 with $\delta/14mHK$, we have that with probability at least $1-\delta/14$, simultaneously for all $i \in [m], h \in [H], e \in [E]$

$$\|\zeta^{k_e,i}_{p,h}\|_{\Lambda^{k_e}_h} = \sqrt{2H}\beta_p\|g^{e,i}_{p,h}\| \le \beta_p\sqrt{11dH\log(14mHK/\delta)} = \beta_p\beta_\zeta.$$

Similarly, defining $\zeta^{k_e,i}_r = \sqrt{2\beta_r^2}(\Lambda^k)^{-1/2}g^{e,i}_r$, with $g^{e,i}_r \sim \mathcal{N}(0, I_{dH})$, and taking a union bound over Lemma 29 with $\delta/14mK$, we have that with probability at least $1-\delta/14$, simultaneously for all $i \in [m], k \in [K]$

$$\|\zeta^{k_e,i}_r\|_{\Lambda^{k_e}} = \sqrt{2}\beta_r\|g^{e,i}_r\| \le \beta_r\sqrt{11dH\log(14mHK/\delta)} = \beta_r\beta_\zeta.$$

Taking a union bound over the last two events shows that E_3 (Eq. (18)) holds with probability at least $1-\delta/7$.

Next, for any $e \in [E]$ notice that conditioned on $\mathcal{Z}_h, h \in [H]$ and $\Sigma^{k_e}_\zeta$ we have that $(\bar{\phi}^{\pi^\star})^\mathsf{T}\zeta^{k_e,i}, i \in [m]$ are i.i.d $\mathcal{N}(0,\sigma^2)$ variables with $\sigma^2 = \left\|\bar{\phi}^{\pi^\star}\right\|^2_{\Sigma^{k_e}_\zeta}$. Applying Lemma 28 with $m = 9\log(7K/\delta)$ and taking a union bound, we have that with probability at least $1-\delta/7$ E_4 (Eq. (19)) holds.

Next, by Lemma 17, E_5 (Eq. (20)) holds with probability at least $1-\delta/7$. Finally, notice that $\|\phi^k\|_{(\Lambda^k)^{-1}}, \|\phi^k_h\|_{(\Lambda^k_h)^{-1}} \le 1$, thus $0 \le Y_k \le (1+\sqrt{2}\beta_\zeta)(\beta_r + H\beta_p)$. Using Azuma's inequality, we conclude that E_6 (Eq. (21)) holds with probability at least $1-\delta/7$. $\bar{E}_7$ (Eq. (25)) holds with probability at least $1-\delta/7$ by a similar argument. ■

The good event. The following results show that the proxy good event implies the good event.

Lemma 19. *Suppose that $\bar{E}_g$ holds. If $\pi^{k,i}_h \in \Pi(\mathcal{Z}_h, WK)$ for all $h \in [H]$ then $\hat{Q}^{k,i}_h \in \widehat{\mathcal{Q}}(\mathcal{Z}_h, W, \beta_{Q,h}), \hat{V}^{k,i}_h \in \widehat{\mathcal{V}}(\mathcal{Z}_h, W, \beta_{Q,h})$ for all $h \in [H+1]$.*

Proof. We show that the claim holds for by backwards induction on $h \in [H+1]$.
Base case $h = H+1$: Since $V^{k,i}_{H+1} = 0$ it is also implied that $\hat{Q}^{k,i}_{H+1} = 0$. Because $w = 0 \in \widehat{\mathcal{Q}}(\mathcal{Z}_h, W, \beta_{Q,H+1} = 0)$ we have that $\hat{Q}^{k,i}_{H+1} \in \widehat{\mathcal{Q}}(\mathcal{Z}_h, W, \beta_{Q,H+1} = 0)$, and similarly $V^{k,i}_{H+1} \in \hat{V}^{k,i}_h \in \widehat{\mathcal{V}}(\mathcal{Z}_h, W, \beta_{Q,h})$.

Induction step: Now, suppose the claim holds for $h+1$ and we show it also holds for h. We have that

$$\begin{aligned}
|\hat{Q}_h^{k,i}(x,a)| &= |\phi(x,a)^\mathsf{T} w_h^{k,i} \mathbb{1}_{\{x\in\mathcal{Z}_h\}}| \\
&= \mathbb{1}_{\{x\in\mathcal{Z}_h\}} \cdot |\phi(x,a)^\mathsf{T}(\theta_h + (\widehat{\theta}_h^k - \theta_h) + \zeta_{r,h}^{k,i} + \zeta_{p,h}^{k,i} + (\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i} + \psi_h \hat{V}_{h+1}^{k,i})| \\
&\le 1 + \|\widehat{\theta}^k - \theta\| + \|\zeta_r^{k,i}\| + \|\hat{V}_{h+1}^{k,i}\|_\infty \qquad \text{(triangle inequality, Cauchy-Schwarz, } \|\phi(x,a)\| \le 1\text{)} \\
&\qquad + \mathbb{1}_{\{x\in\mathcal{Z}_h\}} \cdot \|\phi(x,a)\|_{(\Lambda_h^{k_e})^{-1}} \left[\|\zeta_{p,h}^{k,i}\|_{\Lambda_h^{k_e}} + \|(\widehat{\psi}_h^k - \psi_h)\hat{V}_{h+1}^{k,i}\|_{\Lambda_h^{k_e}} \right] \\
&\le 1 + \frac{\beta_r}{\sqrt{H}} + \frac{\beta_\zeta \beta_r}{\sqrt{H}} + \beta_{Q,h+1} + \frac{1}{2\beta_w H}\left[\beta_\zeta \beta_p + 12\beta_{Q,h+1}\sqrt{d \log \frac{15K^2HW}{\delta\sqrt{d}}} \right] \\
&\qquad \text{(induction hypothesis, Eqs. (16), (18), (20) and (24))} \\
&\le \beta_{Q,H} + \left(1 + \frac{1}{H}\right)\beta_{Q,h+1} \qquad \left(\beta_w \ge 36\beta_\zeta \sqrt{d\log \tfrac{15K^2HW}{\delta\sqrt{d}}}\right) \\
&= \beta_{Q,h}, \qquad \text{(Eq. (26))}
\end{aligned}$$

and

$$\|w_h^{k,i}\| = \|\widehat{\theta}_h^k + \zeta_{r,h}^{k,i} + \zeta_{p,h}^{k,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}\| \le HK + \frac{\beta_r \beta_\zeta}{\sqrt{H}} + \beta_p\beta_\zeta + \beta_Q K \le 4\beta_Q K = W.$$

Since $\pi_h^{k,i} \in \Pi(\mathcal{Z}_h, W)$, this proves the induction step and concludes the proof. ■

Lemma (restatement of Lemma 8). *Consider the following parameter setting:*

$$\lambda_p = 1, \lambda_r = H, m = 9\log(7K/\delta), \beta_r = 2H\sqrt{2dH\log(14K/\delta)}, \beta_\zeta = \sqrt{11dH\log(14mHK/\delta)},$$

$$\beta_p = 216\beta_\zeta\beta_r\sqrt{dH\log\frac{60K^3H\beta_Q}{\delta\sqrt{d}}}, \beta_Q = 16\beta_\zeta\beta_r\sqrt{H}, \beta_w = 36\beta_\zeta\sqrt{d\log\frac{60K^3H\beta_Q}{\delta\sqrt{d}}}, \eta_o \le 1, \epsilon_{cov} \ge 1/K.$$

Then $\Pr[E_g] \ge 1-\delta$.

Proof. Suppose that $\bar{E}_g$ holds. By Lemma 18, this occurs with probability at least $1-\delta$. We show that $\bar{E}_g$ implies E_g, thus concluding the proof. Notice that

$$\begin{aligned}
\pi_h^{k,i}(a|x) \propto \exp\left(\eta \sum_{k'=k_e}^{k-1} \hat{Q}_h^{k',i}(x,a)\right) &= \exp\left(\mathbb{1}_{\{x\in\mathcal{Z}_h\}}\phi(x,a)^\mathsf{T} \eta \sum_{k'=k_e}^{k-1} w_h^{k',i}\right) \\
&= \exp\left(\mathbb{1}_{\{x\in\mathcal{Z}_h\}}\phi(x,a)^\mathsf{T} q_h^{k,i}\right),
\end{aligned}$$

where $q_h^{k,i} = \eta \sum_{k'=k_e}^{k-1} w_h^{k',i}$. We thus have that $\|q_h^{k,i}\| \le WK$ implies $\pi_h^{k,i} \in \Pi(\mathcal{Z}_h, WK)$. We show by induction on $k \ge k_0$ that $\pi_h^{k,i} \in \Pi(\mathcal{Z}_h, WK)$ for all $h \in [H]$. For the base case, $k = k_0$, $\pi_h^{k,i}$ are uniform, implying that $q_h^{k_0,i} = 0$, thus $\pi_h^{k_0,i} \in \Pi(\mathcal{Z}_h, WK)$. Now, suppose the claim holds for all $k' < k$. Then by Lemma 19 we have that $\hat{Q}_h^{k',i} \in \widehat{\mathcal{Q}}(\mathcal{Z}_h, W, \beta_{Q,h+1})$ for all $k' < k$ and $h \in [H]$. This implies that $\|w_h^{k',i}\| \le W$ for all $k' < k$ and $h \in [H]$, thus $\|q_h^{k,i}\| \le WK\eta \le WK$ for all $h \in [H]$, concluding the induction step.

Now, since $\pi_h^{k,i} \in \Pi(\mathcal{Z}_h, WK)$ for all $k \ge k_0, h \in [H], i \in [m]$, we can apply Lemma 19 to get that $\hat{Q}_h^{k,i} \in \widehat{\mathcal{Q}}(\mathcal{Z}_h, W, \beta_{Q,h+1}), \hat{V}_h^{k,i} \in \widehat{\mathcal{V}}(\mathcal{Z}_h, W, \beta_{Q,h+1})$ for all $k \ge k_0, h \in [H], i \in [m]$. Notice that for any $h \in [H]$, $\beta_{Q,h} \le 3H\beta_{Q,H} = 18\sqrt{H}\beta_\zeta\beta_r = \beta_Q$, and

$$\begin{aligned}
\beta_{p,h} = 12\beta_{Q,h+1}\sqrt{d\log\frac{15K^2HW}{\delta\sqrt{d}}} &\le 36\beta_{Q,H}H\sqrt{d\log\frac{15K^2HW}{\delta\sqrt{d}}} \\
&= 216\beta_\zeta\beta_r\sqrt{dH\log\frac{15K^2HW}{\delta\sqrt{d}}} = \beta_p.
\end{aligned}$$

Using $\bar{E}_2$ (Eq. (24)) we conclude that E_2 (Eq. (17)) holds. This also implies that $\text{clip}_{\beta_Q}\left[\hat{V}_1^{k,i}(x_1)\right] = \hat{V}_1^{k,i}(x_1)$, thus $\bar{E}_7$ (Eq. (25)) implies E_7 (Eq. (22)). ■

C. Randomized Ensemble Policy Optimization (REPO) for Tabular MDPs

In this section, we present a simplified version of REPO (Algorithm 3) for tabular MDPs. For consistency of the presentation, we encode the tabular MDP as a linear MDP of dimension $d = |\mathcal{X}||\mathcal{A}|$ using a standard one-hot encoding (see Example 2.1 in Jin et al. (2020b)).

C.1. The Randomized Ensemble Policy Optimization Algorithm

We present REPO for tabular MDPs in Algorithm 4. The algorithm is identical to Algorithm 3 up to the following differences. First, we remove the reward-free warm-up step. Second, we remove the indicator mechanism used to keep the Q values bounded. Third, we affix the dynamics backup estimate at the start of each epoch (see the choice of dataset $\mathcal{D}_h^k$ in Line 11 of Algorithm 4). Otherwise, the algorithm is unchanged. As will be seen in the analysis, this simplification is made possible since the empirical tabular MDP, i.e., the one induced by the observed samples, has a sub-stochastic transition kernel, i.e., one with non-negative values that sum to at most one.

Algorithm 4 REPO with aggregate feedback for tabular MDPs

```
1: input: $\delta, \eta_o, \eta_x, \lambda_r, \lambda_p, \beta_r, \beta_p > 0; m \geq 1$.
2: initialize: $e \leftarrow -1$.
3: for episode $k = 1, 2, \ldots, K$ do
4:     if $k = 1$ or $\exists h \in [H],\ \det(\Lambda_h^k) \geq 2\det(\Lambda_h^{k_e})$ or $\det(\Lambda^k) \geq 2\det(\Lambda^{k_e})$ then
5:         $e \leftarrow e + 1$ and $k_e \leftarrow k$.
6:         $\Sigma_\zeta^{k_e} \leftarrow 2\beta_r^2 (\Lambda^{k_e})^{-1} + 2H\beta_p^2 \mathrm{diag}\left(\Lambda_1^{k_e}, \ldots, \Lambda_H^{k_e}\right)^{-1}$
7:         Sample $\zeta^{k_e,i} \sim \mathcal{N}\left(0, \Sigma_\zeta^{k_e}\right)$ for all $i \in [m]$.
8:         Reset $p^k(i) \leftarrow 1/m, \pi_h^{k,i}(a \mid x) \leftarrow 1/|\mathcal{A}|$.
9:     Sample $i_k$ according to $p^k(\cdot)$ and play $\pi^k = \pi^{k,i_k}$.
10:    Observe episode reward $v^k$ and trajectory $\iota^k$.
11:    Define $\mathcal{D}^k = \{1, \ldots, k-1\}$ and $\mathcal{D}_h^k = \{1, \ldots, k_e - 1\}$ for all $h \in [H]$.
12:    Compute $\widehat{\theta}^k$ and define $\widehat{\psi}_h^k$ (Eqs. (1) and (2)).
13:    Define $\hat{V}_{H+1}^{k,i}(x) = 0$ for all $i \in [m], x \in \mathcal{X}$.
14:    For every $i \in [m]$ and $h = H, \ldots, 1$:
           $w_h^{k,i} \leftarrow \widehat{\theta}_h^k + \zeta_h^{k_e,i} + \widehat{\psi}_h^k \hat{V}_{h+1}^{k,i}$
           $\hat{Q}_h^{k,i}(x,a) = \phi(x,a)^\mathsf{T} w_h^{k,i}$
           $\hat{V}_h^{k,i}(x) = \sum_{a \in \mathcal{A}} \pi_h^{k,i}(a \mid x) \hat{Q}_h^{k,i}(x,a)$
           $\pi_h^{k+1,i}(a \mid x) \propto \pi_h^{k,i}(a \mid x) \exp(\eta_o \hat{Q}_h^{k,i}(x,a))$.
15:    $p^{k+1}(i) \propto p^k(i) \exp(\eta_x \hat{V}_1^{k,i}(x_1))$ for all $i \in [m]$.
```

The following is our main result for Algorithm 4 (proof in the following Appendix C.2).

Theorem 20. *Suppose that we run Algorithm 4 with* $\eta_x = \sqrt{\frac{3dH\log(2K)\log m}{K\beta_Q^2}}$, $\eta_o = \sqrt{\frac{3dH\log(2K)\log|\mathcal{A}|}{K\beta_Q^2}}$ *and the other parameters as in Lemma 21. Then with probability at least* $1 - \delta$ *we incur regret at most*

$$\begin{aligned}\text{Regret} &\leq 1747\sqrt{Kd^3H^7\max\{H, |\mathcal{X}|\}}\log^2\frac{20KH|\mathcal{X}||\mathcal{A}|m}{\delta} \\ &= 1747\sqrt{K|\mathcal{X}|^3|\mathcal{A}|^3H^7\max\{H, |\mathcal{X}|\}}\log^2\frac{20KH|\mathcal{X}||\mathcal{A}|m}{\delta} \\ &= \tilde{O}(\sqrt{K|\mathcal{X}|^3|\mathcal{A}|^3H^7\max\{H, |\mathcal{X}|\}}).\end{aligned}$$

C.2. Analysis (REPO for tabular MDPs)

Good event. Let $\widehat{P}^k = (\widehat{P}^k_1, \ldots, \widehat{P}^k_H)$ be the empirical (sub-probability) transition kernel defined such that $\widehat{P}^k_h(\cdot|x,a) = \phi(x,a)\widehat{\psi}^k_h$ for all $h \in [H]$. Recall that we keep the kernel fixed throughout each epoch, thus $\widehat{P}^k = \widehat{P}^{k_e}$ for all $e \in [E], k \in K_e$. As in standard MDPs, the empirical transition kernel together with a policy π defines a (sub) probability measure over trajectories $\iota = (x_h, a_h)_{h\in[H]}$. We can therefore define an expectation operator $\mathbb{E}_{\widehat{P}^k,\pi}$ as we did for standard MDPs. While linearity of the expectation holds for any measure, the fact that it is a sub-probability measure implies that $\mathbb{E}_{\widehat{P}^k,\pi} f(\iota) \le \max_\iota f(\iota)$.

Finally, for a policy π and $k \in [K], h \in [H]$, we define the expected empirical feature occupancy as $\hat{\phi}^{k,\pi}_h = \mathbb{E}_{\widehat{P}^k,\pi}\phi(x_h, a_h)$ and their concatenation as $\hat{\phi}^{k,\pi} = ((\hat{\phi}^{k,\pi}_1)^\mathsf{T}, \ldots, (\hat{\phi}^{k,\pi}_H)^\mathsf{T})^\mathsf{T}$. In addition, to simplify presentation, we denote $\zeta^{k,i} = \zeta^{k,i}_r + \zeta^{k,i}_p$ where $\zeta^{k,i}_r \sim \mathcal{N}\Big(0, 2\beta_r^2(\Lambda^k)^{-1}\Big)$ and $\zeta^{k,i}_p \sim \mathcal{N}\Big(0, 2H\beta_p^2 \mathrm{diag}\big(\Lambda^k_1, \ldots, \Lambda^k_H\big)^{-1}\Big)$. We define the following good event $E_g = \bigcap_{i=1}^7 E_i$, over which the regret is deterministically bounded:

$$E_1 = \Big\{\forall e \in [E], k \in K_e : \|\theta - \widehat{\theta}^k\|_{\Lambda^k} \le \beta_r\Big\}; \tag{27}$$

$$E_2 = \Big\{\forall e \in [E], k \in K_e, i \in [m], h \in [H] : \|(\psi_h - \widehat{\psi}^k_h)V^\star_{h+1}\|_{\Lambda^{k_e}_h} \le \beta_p, \|\hat{Q}^{k,i}_{h+1}\|_\infty \le \beta_Q\Big\}; \tag{28}$$

$$E_3 = \Big\{\forall e \in [E], k \in K_e, i \in [m], h \in [H] : \|\zeta^{k_e,i}_r\|_{\Lambda^{k_e}} \le \beta_\zeta\beta_r, \|\zeta^{k_e,i}_{p,h}\|_{\Lambda^{k_e}_h} \le \beta_\zeta\beta_p\Big\}; \tag{29}$$

$$E_4 = \left\{\forall e \in [E] : \max_{i\in[m]}(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,i} \ge \|\hat{\phi}^{k_e,\pi^\star}\|_{\Sigma^{k_e}_\zeta}\right\}; \tag{30}$$

$$E_5 = \Big\{\forall e \in [E], k \in K_e, h \in [H], x \in \mathcal{X}, a \in \mathcal{A} : \|\phi(x,a)^\mathsf{T}(\psi_h - \widehat{\psi}^k_h)\|_1 \le \beta_{\hat{p}}\|\phi(x,a)\|_{\Lambda^{k_e}_h{}^{-1}}\Big\}; \tag{31}$$

$$E_6 = \left\{\sum_{k\in[K]} \mathbb{E}_{P,\pi^k}[Y_k] \le \sum_{k\in[K]} Y_k + ((1+\sqrt{2}\beta_\zeta)\beta_r + \sqrt{2}(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}})H)\sqrt{2K\log\frac{7}{\delta}}\right\}. \tag{32}$$

$$E_7 = \left\{\sum_{k\in[K]}\sum_{i\in[m]} p^k(i)\hat{V}^{k,i}_1(x_1) \le \sum_{k\in[K]} \hat{V}^{k,i_k}_1(x_1) + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}\right\}; \tag{33}$$

where $Y_k = \beta_r(1+\sqrt{2}\beta_\zeta)\|\phi^k\|_{(\Lambda^k)^{-1}} + \sqrt{2}(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}})\sum_{h\in[H]}\|\phi^k_h\|_{(\Lambda^k_h)^{-1}}$.

Lemma 21 (Good event). *Consider the following parameter setting:*

$$\lambda_p = 1, \lambda_r = H, m = 9\log(7K/\delta), \beta_r = 2H\sqrt{2dH\log(14K/\delta)}, \beta_\zeta = \sqrt{11dH\log(14mK/\delta)},$$
$$\beta_p = 4H\sqrt{3d\log(14KH/\delta)}, \beta_{\hat{p}} = 2\sqrt{5|\mathcal{X}|\log(20KH|\mathcal{X}||\mathcal{A}|/\delta)}, \beta_Q = 2H\beta_p\beta_\zeta.$$

Then $\Pr[E_g] \ge 1 - \delta$.

Proof in Appendix C.3.

Optimism and its cost.

Lemma 22 (Optimism). *Suppose that the good event E_g holds (Eqs. (27) to (33)) and define $\hat{j}_e = \arg\max_{i\in[m]}(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,i}$, then*

$$\sum_{h\in[H]} \mathbb{E}_{\widehat{P}^k,\pi^\star}\Big[Q^\star_h(x_h,a_h) - \phi(x_h,a_h)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k_e,\hat{j}_e} + \widehat{\psi}^k_h V^\star_{h+1})\Big] \le 0 \quad ,\forall e \in [E], k \in K_e.$$

Proof. We have that

$$\begin{aligned}
&\sum_{h\in[H]} \mathbb{E}_{\widehat{P}^k,\pi^\star}\left[Q^\star_h(x_h,a_h) - \phi(x_h,a_h)^\mathsf{T}(\widehat{\theta}^k_h + \zeta^{k_e,\hat{j}_e}_h + \widehat{\psi}^k_h V^\star_{h+1})\right] \\
&\quad = \sum_{h\in[H]} \mathbb{E}_{\widehat{P}^k,\pi^\star}\left[\phi(x_h,a_h)^\mathsf{T}(\theta_h - \widehat{\theta}^k_h - \zeta^{k_e,\hat{j}_e}_h + (\psi_h - \widehat{\psi}^k_h)V^\star_{h+1})\right] \\
&\quad = \sum_{h\in[H]} (\hat{\phi}^{k_e,\pi^\star}_h)^\mathsf{T}(\theta_h - \widehat{\theta}^k_h - \zeta^{k_e,\hat{j}_e}_h + (\psi_h - \widehat{\psi}^k_h)V^\star_{h+1}) \\
&\quad = -(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} + (\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}(\theta - \widehat{\theta}^k) + \sum_{h\in[H]} (\hat{\phi}^{k_e,\pi^\star}_h)^\mathsf{T}(\psi_h - \widehat{\psi}^k_h)V^\star_{h+1} \\
&\quad \le -(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} + \beta_r\|\hat{\phi}^{k_e,\pi^\star}\|_{(\Lambda^k)^{-1}} + \beta_p \sum_{h\in[H]} \|\hat{\phi}^{k_e,\pi^\star}_h\|_{(\Lambda^{k_e}_h)^{-1}} && \text{(Eqs. (27) and (28))} \\
&\quad \le -(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} + \beta_r\|\hat{\phi}^{k_e,\pi^\star}\|_{(\Lambda^{k_e})^{-1}} + \beta_p \sum_{h\in[H]} \|\hat{\phi}^{k_e,\pi^\star}_h\|_{(\Lambda^{k_e}_h)^{-1}} && (\Lambda^{k_e} \preceq \Lambda^k) \\
&\quad \le -(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,\hat{j}_e} + \|\hat{\phi}^{k_e,\pi^\star}\|_{2\beta_r^2(\Lambda^{k_e}_h)^{-1} + 2H\beta_p^2 \mathrm{diag}(\Lambda^{k_e}_1,\ldots,\Lambda^{k_e}_H)^{-1}} && \text{(Cauchy-Schwarz)} \\
&\quad = \|\hat{\phi}^{k_e,\pi^\star}\|_{\Sigma^{k_e}_\zeta} - \max_{i\in[m]}(\hat{\phi}^{k_e,\pi^\star})^\mathsf{T}\zeta^{k_e,i} && \text{(Definitions of } \hat{j}_e, \Sigma^{k_e}_\zeta) \\
&\quad \le 0. && \text{(Eq. (30))}
\end{aligned}$$

■

Lemma 23 (Cost of optimism). *Suppose that the good event E_g holds (Eqs. (27) to (33)), then for all $e \in [E], k \in K_e$*

$$\hat{V}^{k,i_k}_1(x_1) - V^{\pi^k}_1(x_1) \le \mathbb{E}_{P,\pi^k}\left[\beta_r(1+\sqrt{2}\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \sqrt{2}(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}}) \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda^k_h)^{-1}}\right].$$

Proof. By Lemma 25, a value difference lemma by Shani et al. (2020),

$$\begin{aligned}
\hat{V}^{k,i_k}_1(x_1) - V^{\pi^k}_1(x_1) &= \mathbb{E}_{P,\pi^k}\left[\sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}\left(\widehat{\theta}^k_h - \theta_h + \zeta^{k_e,i_k}_{r,h} + \zeta^{k_e,i_k}_{p,h} + (\widehat{\psi}^k_h - \psi_h)\hat{V}^{k,i_k}_{h+1}\right)\right] \\
&= \mathbb{E}_{P,\pi^k}\left[\phi(x_{1:H},a_{1:H})^\mathsf{T}(\widehat{\theta}^k - \theta + \zeta^{k_e,i_k}_r) + \sum_{h\in[H]} \phi(x_h,a_h)^\mathsf{T}\left(\zeta^{k_e,i_k}_{p,h} + (\widehat{\psi}^k_h - \psi_h)\hat{V}^{k,i_k}_{h+1}\right)\right] \\
&\le \mathbb{E}_{P,\pi^k}\Bigg[\beta_r\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \beta_r\beta_\zeta\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^{k_e})^{-1}} \qquad \text{(Cauchy-Schwarz, Eqs. (27) and (29))} \\
&\qquad + \sum_{h\in[H]} \beta_p\beta_\zeta\|\phi(x_h,a_h)\|_{(\Lambda^{k_e}_h)^{-1}} + \|(\phi(x_h,a_h))^\mathsf{T}(\widehat{\psi}^k_h - \psi_h)\|_1\|\hat{V}^{k,i_k}_{h+1}\|_\infty\Bigg] \\
&\le \mathbb{E}_{P,\pi^k}\left[\beta_r(1+\sqrt{2}\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \sqrt{2}(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}}) \sum_{h\in[H]} \|\phi(x_h,a_h)\|_{(\Lambda^k_h)^{-1}}\right]. \\
&\qquad \text{(Lemma 27 and Eqs. (28) and (31))}
\end{aligned}$$

■

Regret bound.

Proof of Theorem 20. Suppose that the good event E_g holds (Eqs. (27) to (33)). By Lemma 21, this holds with probability at least $1-\delta$. For every epoch $e \in [E]$, let $\hat{j}_e = \arg\max_{i\in[m]} \sum_{k\in e} \hat{\phi}^{k,\pi^\star}(x_1)^\mathsf{T}\zeta^{k,i}$. Additionally, for every $e \in [E], k \in$

$K_e, i \in [m]$ let $\hat{V}_h^{k,i,\pi^\star}$ be the value function of the true optimal policy $\pi^\star$ in the empirical MDP-like structure whose rewards and dynamics are defined as $\hat{r}_h^{k,i}(x,a) = \phi(x,a)^\mathsf{T}(\widehat{\theta}_h^k + \zeta_h^{k_e,i})$ and $\widehat{P}_h^k(\cdot|x,a) = \phi(x,a)^\mathsf{T}\widehat{\psi}_h^k$. Now decompose the regret as follows.

$$\begin{aligned}
\text{Regret} = \sum_{k\in[K]} V_1^{\pi^\star}(x_1) - V_1^{\pi^k}(x_1) &= \underbrace{\sum_{k\in[K]} \hat{V}_1^{k,i_k}(x_1) - V_1^{\pi^k}(x_1)}_{(i)} + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,\hat{j}_e}(x_1) - \hat{V}_1^{k,i_k}(x_1)}_{(ii)} \\
&+ \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} \hat{V}_1^{k,\hat{j}_e,\pi^\star}(x_1) - \hat{V}_1^{k,\hat{j}_e}(x_1)}_{(iii)} + \underbrace{\sum_{e\in[E]}\sum_{k\in K_e} V_1^\star(x_1) - \hat{V}_1^{k,\hat{j}_e,\pi^\star}(x_1)}_{(iv)}.
\end{aligned}$$

For term (i), we use Lemma 23 as follows.

$$\begin{aligned}
(i) &\le \sum_{e\in[E]}\sum_{k\in K_e} \mathbb{E}_{P,\pi^k}\left[\beta_r(1+\sqrt{2}\beta_\zeta)\|\phi(x_{1:H},a_{1:H})\|_{(\Lambda^k)^{-1}} + \sqrt{2}(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}})\sum_{h\in[H]}\|\phi(x_h,a_h)\|_{(\Lambda_h^k)^{-1}}\right] \\
&\le \beta_r(1+\sqrt{2}\beta_\zeta)\sum_{k\in[K]}\|\phi^k\|_{(\Lambda^k)^{-1}} + \sqrt{2}(\beta_p\beta_\zeta+\beta_Q\beta_{\hat{p}})\sum_{k\in[K]}\sum_{h\in[H]}\|\phi_h^k\|_{(\Lambda_h^k)^{-1}} && \text{(Eq. (32))} \\
&\quad + ((1+\sqrt{2}\beta_\zeta)\beta_r + \sqrt{2}(\beta_p\beta_\zeta+\beta_Q\beta_{\hat{p}})H)\sqrt{2K\log\frac{7}{\delta}} \\
&\le 3\beta_p\beta_\zeta H\sqrt{Kd\log(2K)} + 5H^2\beta_p\beta_\zeta\beta_{\hat{p}}\sqrt{Kd\log(2K)} + 6H^2\beta_p\beta_\zeta\beta_{\hat{p}}\sqrt{K\log\frac{7}{\delta}} \\
& && \text{(Lemma 26, } \beta_\zeta,\beta_{\hat{p}}\ge 3, \beta_p\sqrt{H}\ge\beta_r, \beta_Q = 2H\beta_p\beta_\zeta\text{)} \\
&\le 12H^2\beta_p\beta_\zeta\beta_{\hat{p}}\sqrt{Kd\log\frac{7K}{\delta}}.
\end{aligned}$$

Next, by Lemmas 12 and 14 (with our choice of η_x), we have that

$$\begin{aligned}
(ii) \le 2\beta_Q\sqrt{KE\log m} + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}} &\le 2\beta_Q\sqrt{3KdH\log(2K)\log m} + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}} \\
&\le 7\beta_Q\sqrt{KdH}\log\left(\frac{7Km}{\delta}\right).
\end{aligned}$$

Next, we decompose term (iii) using Lemma 25 as follows.

$$\begin{aligned}
(iii) &= \sum_{e\in[E]}\sum_{k\in K_e}\mathbb{E}_{\widehat{P}^k,\pi^\star}\left[\sum_{h\in[H]}\sum_{a\in\mathcal{A}}\hat{Q}_h^{k,\hat{j}_e}(x_h,a)(\pi_h^\star(a\mid x_h) - \pi_h^{k,\hat{j}_e}(a\mid x_h))\right] \\
&= \sum_{h\in[H]}\sum_{e\in[E]}\mathbb{E}_{\widehat{P}^k,\pi^\star}\left[\sum_{k\in K_e}\sum_{a\in\mathcal{A}}\hat{Q}_h^{k,\hat{j}_e}(x_h,a)(\pi_h^\star(a\mid x_h) - \pi_h^{k,\hat{j}_e}(a\mid x_h))\right] && (\widehat{P}^k \text{ fixed within epoch}) \\
&\le \sum_{h\in[H]}\sum_{e\in[E]}\left(\frac{\log|\mathcal{A}|}{\eta_o} + \eta_o\sum_{k\in K_e}\beta_Q^2\right) && \text{(Lemma 13, } \widehat{P}^k \text{ is a sub-probability measure)} \\
&\le \frac{HE\log|\mathcal{A}|}{\eta_o} + \eta_o HK\beta_Q^2 \\
&\le 4H\beta_Q\sqrt{KdH}\log(2K|\mathcal{A}|). && \text{(Lemma 14 and choice of } \eta_o\text{)}
\end{aligned}$$

Next, we decompose term (iv) using Lemma 25 as follows.

$$(iv) = \sum_{e\in[E]}\sum_{k\in K_e}\sum_{h\in[H]}\mathbb{E}_{\widehat{P}^k,\pi^\star}\left[Q_h^\star(x_h,a_h) - \phi(x_h,a_h)^\mathsf{T}(\widehat{\theta}_h^k + \zeta^{k_e,\hat{j}_e} + \widehat{\psi}_h^k V_{h+1}^\star)\right] \le 0,$$

where the second inequality is by Lemma 22. Putting everything together, we conclude that

$$
\begin{aligned}
\text{Regret} &\le 12H^2\beta_p\beta_\zeta\beta_{\hat{p}}\sqrt{Kd\log\frac{7K}{\delta}} + 7\beta_Q\sqrt{KdH}\log\left(\frac{7Km}{\delta}\right) + 4H\beta_Q\sqrt{KdH}\log(2K|\mathcal{A}|) \\
&\le 12H^2\beta_p\beta_\zeta\beta_{\hat{p}}\sqrt{Kd\log\frac{7K}{\delta}} + 22H^2\beta_p\beta_\zeta\sqrt{KdH}\log\left(\frac{7Km|\mathcal{A}|}{\delta}\right) \\
&\le 24H^2\beta_p\beta_\zeta\sqrt{5Kd|\mathcal{X}|}\log\frac{20KH|\mathcal{X}||\mathcal{A}|}{\delta} + 22H^2\beta_p\beta_\zeta\sqrt{KdH}\log\left(\frac{7Km|\mathcal{A}|}{\delta}\right) \\
&\le 76H^2\beta_p\beta_\zeta\sqrt{Kd\max\{H,|\mathcal{X}|\}}\log\frac{20KH|\mathcal{X}||\mathcal{A}|m}{\delta} \\
&\le 1747\sqrt{Kd^3H^7\max\{H,|\mathcal{X}|\}}\log^2\frac{20KH|\mathcal{X}||\mathcal{A}|m}{\delta} \\
&= 1747\sqrt{K|\mathcal{X}|^3|\mathcal{A}|^3H^7\max\{H,|\mathcal{X}|\}}\log^2\frac{20KH|\mathcal{X}||\mathcal{A}|m}{\delta}.
\end{aligned}
$$

■

C.3. Proofs of good event (REPO in Tabular MDPs)

Lemma (restatement of Lemma 21). *Consider the following parameter setting:*

$$
\begin{gathered}
\lambda_p = 1, \lambda_r = H, m = 9\log(7K/\delta), \beta_r = 2H\sqrt{2dH\log(14K/\delta)}, \beta_\zeta = \sqrt{11dH\log(14mK/\delta)}, \\
\beta_p = 4H\sqrt{3d\log(14KH/\delta)}, \beta_{\hat{p}} = 2\sqrt{5|\mathcal{X}|\log(20KH|\mathcal{X}||\mathcal{A}|/\delta)}, \beta_Q = 2H\beta_p\beta_\zeta.
\end{gathered}
$$

Then $\Pr[E_g] \ge 1-\delta$.

Proof. First, by Lemma 31 and our choice of parameters, E_1 (Eq. (27)) holds with probability at least $1-\delta/7$. Next, E_3, E_4 (Eqs. (29) and (30)) follow exactly as Eqs. (18) and (19) proved in Lemma 18. Specifically, the argument for E_4 is unchanged since $\widehat{P}^{k_e}$ and therefore $\hat{\phi}^{k_e,\pi^\star}$ are determined before drawing $(\zeta^{k_e,i})_{i\in[m]}$.

Now, to prove that E_5 (Eq. (31)) holds with probability at least $1-\delta/7$, we use standard arguments for tabular MDPs. Let $\bar{P}_h^k(x' \mid x,a) = n_h^k(x,a,x')/\max\{1, n_h^k(x,a)\}$ be the empirical transition function where $n_h^k(x,a,x')$ is the number of times (until the beginning of episode k) that the agent visited state x at step h, took action a and transitioned to state x', and $n_h^k(x,a) = \sum_{x'} n_h^k(x,a,x')$. By Jaksch et al. (2010, Lemma 17), with probability at least $1-\delta/7$ for all $x\in\mathcal{X}, a\in\mathcal{A}, h\in[H], k\in[K]$ simultaneously

$$
\|\bar{P}_h^k(\cdot \mid x,a) - P_h(\cdot \mid x,a)\|_1 \le \sqrt{\frac{5|\mathcal{X}|\log(20KH|\mathcal{X}||\mathcal{A}|/\delta)}{\max\{1, n_h^k(x,a)\}}}.
$$

Now notice that $\phi(x,a)^\mathsf{T}\psi_h = P_h(\cdot \mid x,a)$ and that $\|\phi(x,a)\|_{(\Lambda_h^{k_e})^{-1}} = 1/\sqrt{1+n_h^{k_e}(x,a)}$. In addition, $\phi(x,a)^\mathsf{T}\widehat{\psi}_h^k = \widehat{P}_h^{k_e}(\cdot \mid x,a)$ when defining $\widehat{P}_h^{k_e}(x' \mid x,a) = n_h^{k_e}(x,a,x')/(1+n_h^{k_e}(x,a))$. Thus, E_5 (Eq. (31)) is given by the triangle inequality together with

$$
\|\widehat{P}_h^{k_e}(\cdot \mid x,a) - \bar{P}_h^{k_e}(\cdot \mid x,a)\|_1 \le \frac{1}{n_h^{k_e}(x,a)+1}.
$$

Next, to prove that E_2 (Eq. (28)) holds, first use Lemma 32 with a set of value functions that contains only a single value function $V^\star$ to get that with probability at least $1-\delta/7$

$$
\|(\psi_h - \widehat{\psi}_h^k)V_{h+1}^\star\|_{\Lambda_h^{k_e}} \le \beta_p, \quad \forall e\in[E], k\in K_e, h\in[H].
$$

Now, suppose that E_1, E_3 Eqs. (27) and (29) hold. Then for all $e\in[E], k\in K_e, h\in[H], i\in[m], a\in\mathcal{A}, x\in\mathcal{X}$:

$$
\begin{aligned}
|\hat{r}_h^{k,i}(x,a)| = \|\widehat{\theta}_h^k + \zeta_h^{k_e,i}\| &\le |\phi(x,a)^\mathsf{T}\theta_h| + \|\widehat{\theta}^k - \theta\| + \|\zeta_r^{k_e,i}\| + \|\zeta_{p,h}^{k_e,i}\| && \text{(triangle inequality, } \|\phi(x,a)\| \le 1) \\
&\le 1 + \frac{\|\widehat{\theta}^k - \theta\|_{\Lambda^k} + \|\zeta_r^{k_e,i}\|_{\Lambda^{k_e}}}{\sqrt{H}} + \|\zeta_p^{k_e,i}\|_{\Lambda_h^{k_e}} && (r_h(x,a)\in[0,1], \Lambda^k, \Lambda^{k_e} \succeq HI, \Lambda_h^{k_e} \succeq I) \\
&\le 1 + \beta_r(1+\beta_\zeta)/\sqrt{H} + \beta_p\beta_\zeta && \text{(Eqs. (27) and (29))} \\
&\le 2\beta_p\beta_\zeta. && (\beta_p \ge 2\beta_r/\sqrt{H}, \beta_\zeta \ge 3, \beta_p \ge 4)
\end{aligned}
$$

Notice that $\hat{Q}_h^{k,i}(x,a) = \mathbb{E}_{\widehat{P}^k,\pi^{k,i}}\left[\sum_{h'=h}^{H} \hat{r}_{h'}^{k,i}(x_{h'},a_{h'}) \mid x_h = x, a_h = a\right]$. Since $\widehat{P}^k$ is a sub-probability measure, we conclude that

$$\|\hat{Q}_h^{k,i}\|_\infty \leq \sum_{h'=h}^{H} \max_{x\in\mathcal{X},a\in\mathcal{A}} |\hat{r}_h^{k,i}(x,a)| \leq 2H\beta_p\beta_\zeta = \beta_Q,$$

establishing E_2 (Eq. (28)).

Next, notice that $\|\phi^k\|_{\Lambda^{k-1}}, \|\phi_h^k\|_{\Lambda_h^{k_e-1}} \leq 1$, thus $0 \leq Y_k \leq \beta_r(1+\sqrt{2}\beta_\zeta) + \sqrt{2}H(\beta_p\beta_\zeta + \beta_Q\beta_{\hat{p}})$. Using Azuma's inequality, we conclude that E_6 (Eq. (32)) holds with probability at least $1-\delta/7$. Finally, we use Azuma's inequality to get that with probability at least $1-\delta/7$

$$\sum_{k\in[K]} \sum_{i\in[m]} p^k(i)\mathrm{clip}_{\beta_Q}\left[\hat{V}_1^{k,i}(x_1)\right] \leq \sum_{k\in[K]} \mathrm{clip}_{\beta_Q}\left[\hat{V}_1^{k,i_k}(x_1)\right] + 2\beta_Q\sqrt{2K\log\frac{7}{\delta}}.$$

Since $\|\hat{Q}_h^{k,i}\|_\infty \leq \beta_Q$, we conclude that $\mathrm{clip}_{\beta_Q}\left[\hat{V}_1^{k,i}(x_1)\right] = \hat{V}_1^{k,i}(x_1)$, establishing E_7 (Eq. (33)).

Taking a union bound, all of the events so far hold with probability at least $1-\delta$. ■

D. Technical tools

D.1. Online Mirror Descent

We begin with a standard regret bound for entropy regularized online mirror descent (hedge). See Sherman et al. (2023, Lemma 25).

Lemma 24. *Let $y_1, \dots, y_T \in \mathbb{R}^A$ be any sequence of vectors, and $\eta > 0$ such that $\eta y_t(a) \geq -1$ for all $t \in [T], a \in [A]$. Then if $x_t \in \Delta_A$ is given by $x_1(a) = 1/A \; \forall a$, and for $t \geq 1$:*

$$x_{t+1}(a) = \frac{x_t(a) e^{-\eta y_t(a)}}{\sum_{a' \in [A]} x_t(a') e^{-\eta y_t(a')}},$$

then,

$$\max_{x \in \Delta_A} \sum_{t=1}^{T} \sum_{a \in [A]} y_t(a)(x_t(a) - x(a)) \leq \frac{\log A}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in [A]} x_t(a) y_t(a)^2.$$

D.2. Value difference lemma

We use the following extended value difference lemma by Shani et al. (2020). We note that the lemma holds unchanged even for MDP-like structures where the transition kernel P is a sub-stochastic transition kernel, i.e., one with non-negative values that sum to at most one (instead of exactly one).

Lemma 25 (Extended Value difference Lemma 1 in (Shani et al., 2020)). *Let $\mathcal{M}$ be an MDP, $\pi, \hat{\pi} \in \Pi_M$ be two policies, $\hat{Q}_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}, h \in [H]$ be arbitrary function, and $\hat{V}_h : \mathcal{X} \to \mathbb{R}$ be defined as $\hat{V}_h(x) = \sum_{a \in \mathcal{A}} \hat{\pi}_h(a \mid x) \hat{Q}_h(x, a)$. Then*

$$\begin{aligned} V_1^{\pi}(x_1) - \hat{V}_1(x_1) &= \mathbb{E}_{P,\pi}\left[\sum_{h \in [H]} \sum_{a \in \mathcal{A}} \hat{Q}_h(x_h, a)(\pi(a \mid x_h) - \hat{\pi}(a \mid x_h))\right] \\ &\quad + \mathbb{E}_{P,\pi}\left[\sum_{h \in [H]} r_h(x_h, a_h) + \sum_{x' \in \mathcal{X}} P(x' \mid x_h, a_h) \hat{V}_{h+1}(x') - \hat{Q}_h(x_h, a_h)\right]. \end{aligned}$$

We note that, in the context of linear MDP $r_h(x_h, a_h) + \sum_{x' \in \mathcal{X}} P(x' \mid x_h, a_h) \hat{V}_{h+1}(x') = \phi(x_h, a_h)^{\mathsf{T}}(\theta_h + \psi_h \hat{V}_{h+1})$.

D.3. Algebraic lemmas

Next, is a well-known bound on harmonic sums (see, e.g., Cohen et al., 2019, Lemma 13). This is used to show that the optimistic and true losses are close on the realized predictions. See proof below for completeness.

Lemma 26. *Let $z_t \in \mathbb{R}^{d'}$ be a sequence such that $\|z_t\|^2 \leq \lambda$, and define $V_t = \lambda I + \sum_{s=1}^{t-1} z_s z_s^{\mathsf{T}}$. Then*

$$\sum_{t=1}^{T} \|z_t\|_{V_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \|z_t\|_{V_t^{-1}}^2} \leq \sqrt{2Td' \log(T+1)}.$$

Proof. Notice that $0 \leq z_t^{\mathsf{T}} V_t^{-1} z_t \leq \|z_t\|^2 / \lambda \leq 1$. Next, notice that

$$\det(V_{t+1}) = \det(V_t + z_t z_t^{\mathsf{T}}) = \det(V_t) \det(I + V_t^{-1/2} z_t z_t^{\mathsf{T}} V_t^{-1/2}) = \det(V_t)(1 + z_t^{\mathsf{T}} V_t^{-1} z_t),$$

which follows from the matrix determinant lemma. We thus have that

$$z_t^{\mathsf{T}} V_t^{-1} z_t \leq \log(1 + z_t^{\mathsf{T}} V_t^{-1} z_t) = \log \frac{\det(V_{t+1})}{\det(V_t)} \qquad (x \leq 2\log(1+x), \forall x \in [0,1])$$

We conclude that

$$\begin{aligned}
\sum_{t=1}^{T} z_t^{\mathsf{T}} V_t^{-1} z_t &\leq 2 \sum_{t=1}^{T} \log(\det(V_{t+1}) / \det(V_t)) \\
&= 2 \log(\det(V_{T+1}) / \det(\lambda I)) && \text{(telescoping sum)} \\
&\leq 2d' \log(\|V_{T+1}\| / \lambda) && (\det(V) \leq \|V\|^{d'}) \\
&\leq 2d' \log\left(1 + \sum_{t=1}^{T} \|z_t\|^2 / \lambda\right) && \text{(triangle inequality)} \\
&\leq 2d' \log(T+1).
\end{aligned}$$

The proof is concluded by applying the Cauchy-Schwarz inequality to get that $\sum_{t=1}^{T} \|z_t\|_{V_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \|z_t\|_{V_t^{-1}}^2}$. ■

Next, we need the following well-known matrix inequality.

Lemma 27 ((Cohen et al., 2019), Lemma 27). *If $N \succeq M \succ 0$ then for any vector v*

$$\|v\|_N^2 \leq \frac{\det N}{\det M} \|v\|_M^2$$

D.4. Concentration and anti-concentration bounds

Next, we have an anti-concentration for standard Gaussian random variables.

Lemma 28. *Let $\sigma, m \geq 0$. Suppose that $g_i \sim \mathcal{N}(0, \sigma^2)$, $i \in [m]$ are i.i.d Gaussian random variables. With probability at least $1 - e^{-m/9}$*

$$\max_{i \in [m]} g_i \geq \sigma.$$

Proof. Recall that for a standard Gaussian random variable $G \sim \mathcal{N}(0,1)$ we have that $\Pr[G \geq 1] \geq 1/9$. Since g_i are independent, we conclude that

$$\Pr[\max_{i \in [m]} g_i \leq \sigma] = \Pr[G \leq 1]^m \leq (8/9)^m \leq e^{-m/9},$$

and taking the complement of this event concludes the proof. ■

Next, a tail inequality on the norm of a standard Gaussian random vector

Lemma 29. *Let $g \sim \mathcal{N}(0, I)$ be a d dimensional Gaussian random vector. With probability at least $1 - \delta$*

$$\|g\| \leq \sqrt{\frac{3d}{2} + 4 \log \frac{1}{\delta}} \leq \sqrt{\frac{11d}{2} \log \frac{1}{\delta}}.$$

Next, we state a standard concentration inequality for a self-normalized processes.

Lemma 30 (Concentration of Self-Normalized Processes (Abbasi-Yadkori et al., 2011)). *Let η_t ($t \geq 1$) be a real-valued stochastic process with corresponding filtration $\mathcal{F}_t$. Suppose that $\eta_t \mid \mathcal{F}_{t-1}$ are zero-mean R-subGaussian, and let ϕ_t ($t \geq 1$) be an $\mathbb{R}^d$-valued, $\mathcal{F}_{t-1}$-measurable stochastic process. Assume that Λ_0 is a $d \times d$ positive definite matrix and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} \phi_s \phi_s^{\mathsf{T}}$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \geq 0$*

$$\left\| \sum_{s=1}^{t} \phi_s \eta_s \right\|_{\Lambda_t^{-1}}^2 \leq 2R^2 \log\left[\frac{\det(\Lambda_t)^{1/2} \det(\Lambda_0)^{-1/2}}{\delta}\right]$$

Additionally, if $\Lambda_0 = \lambda I$ and $\|\phi_t\|^2 \leq \lambda$, for all $t \geq 1$ then

$$\left\| \sum_{s=1}^{t} \phi_s \eta_s \right\|_{\Lambda_t^{-1}}^2 \leq R^2[d \log(t+1) + 2 \log(1/\delta)]$$

D.5. Reward and dynamics estimation bounds

We derive the standard guarantee for the rewards least-squares estimate.

Lemma 31 (reward error bound). *Let $\widehat{\theta}^k$ be as in Algorithm 2 and suppose that $\lambda_r = H$. With probability at least $1-\delta$, for all $k \geq 1$*

$$\|\theta - \widehat{\theta}^k\|_{\Lambda^k} \leq 2H\sqrt{2dH\log(2K/\delta)}$$

Proof. We write $v^\tau = (\phi^\tau)^\mathsf{T}\theta + \eta_\tau$ where $\eta_\tau = \sum_{h\in[H]} r_h^k - r_h(x_h^k, a_h^k)$. Notice that $|\eta_\tau| \leq H$, making it H-subGaussian, which satisfies the conditions of Lemma 30 with $R = H$. Next, notice that

$$\widehat{\theta}^k = {\Lambda^k}^{-1}\sum_{\tau=1}^{k-1}\phi^\tau v^\tau = {\Lambda^k}^{-1}\sum_{\tau=1}^{k-1}\phi^\tau[(\phi^\tau)^\mathsf{T}\theta + \eta_\tau] = \theta - \lambda_r{\Lambda^k}^{-1}\theta + {\Lambda^k}^{-1}\sum_{\tau=1}^{k-1}\phi^\tau\eta_\tau$$

We conclude that with probability at least $1-\delta/5$, for all $k \geq 1$

$$\begin{aligned}
\left\|\theta - \widehat{\theta}^k\right\|_{\Lambda^k} &\leq \|\lambda_r(\Lambda^k)^{-1}\theta\|_{\Lambda^k} + \left\|(\Lambda^k)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau\eta_\tau\right\|_{\Lambda^k} \\
&= \lambda_r\|\theta\|_{(\Lambda^k)^{-1}} + \left\|\sum_{\tau=1}^{k-1}\phi^\tau\eta_\tau\right\|_{(\Lambda^k)^{-1}} \\
&\leq \lambda_r\|\theta\|_{(\Lambda^k)^{-1}} + H\sqrt{dH\log(K+1) + 2\log(1/\delta)} && \text{(Lemma 30)} \\
&\leq \sqrt{\lambda_r}\|\theta\| + H\sqrt{dH\log(K+1) + 2\log(1/\delta)} && (\Lambda^k \succeq \lambda_r I) \\
&\leq H\sqrt{d} + H\sqrt{dH\log(K+1) + 2\log(1/\delta)} && (\|\theta\| \leq \sqrt{dH}, \lambda_r = H) \\
&\leq 2H\sqrt{dH\log(K+1) + 2\log(1/\delta)} \\
&\leq 2H\sqrt{2dH\log(2K/\delta)}. && \blacksquare
\end{aligned}$$

Next, we derive a standard error bound for the dynamics approximation.

Lemma 32 (dynamics error uniform convergence). *Suppose that $\lambda_p = 1$. For all $h \in [H]$, let $\mathcal{V}_h \subseteq \mathbb{R}^\mathcal{X}$ be a set of mappings $V : \mathcal{X} \to \mathbb{R}$ such that $\|V\|_\infty \leq \beta$ and $\beta \geq 1$. Let $\widehat{\psi}_h^k : \mathbb{R}^\mathcal{X} \to \mathbb{R}^d$ be the linear operator defined as*

$$\widehat{\psi}_h^k V = (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau V(x_{h+1}^\tau).$$

With probability at least $1-\delta$, for all $h \in [H]$, $V \in \mathcal{V}_{h+1}$ and $k \geq 1$

$$\|(\psi_h - \widehat{\psi}_h^k)V\|_{\Lambda_h^k} \leq 4\beta\sqrt{d\log(K+1) + 2\log(H\mathcal{N}_\epsilon/\delta)},$$

where $\epsilon \leq \beta\sqrt{d}/2K$, $\mathcal{N}_\epsilon = \sum_{h\in[H]}\mathcal{N}_{h,\epsilon}$, and $\mathcal{N}_{h,\epsilon}$ is the $\epsilon-$covering number of $\mathcal{V}_h$ with respect to the supremum distance.

Proof. For any $h \in [H]$ let $\mathcal{V}_{h,\epsilon}$ be a minimal $\epsilon-$cover for $\mathcal{V}_h$. Next, for any $h \in [H]$ and $V \in \mathcal{V}_{h+1}$ define the linear map $\eta_{\tau,h}(V) = V(x_{h+1}^\tau) - (\phi_h^\tau)^\mathsf{T}\psi_h V$. Notice that $|\eta_{\tau,h}(V)| \leq 2\|V\|_\infty \leq 2\beta$, thus $\eta_{\tau,h}(V)$ satisfies Lemma 30 with $R = 2\beta$. Taking a union bound, we conclude that with probability at least $1-\delta$

$$\left\|\sum_{\tau=1}^{k-1}\phi_h^\tau\eta_{\tau,h}(\tilde{V})\right\|_{(\Lambda_h^k)^{-1}} \leq 2\beta\sqrt{d\log(K+1) + 2\log(H\mathcal{N}_\epsilon/\delta)} \quad ,\forall k \in [K], h \in [H], \tilde{V} \in \mathcal{V}_{h+1,\epsilon}. \tag{34}$$

We assume that the above event holds for the remainder of the proof. Now, notice that

$$\widehat{\psi}_h^k V = (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau V(x_{h+1}^\tau) = \psi_h V - \lambda_p(\Lambda_h^k)^{-1}\psi_h V + (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau\eta_{\tau,h}(V).$$

We are now ready to finish the proof. Let $k \in [K], h \in [H], V \in \mathcal{V}_{h+1}$ and let $\tilde{V} \in \mathcal{V}_{h+1,\epsilon}$ be the point in the cover corresponding to V. Denoting their residual as $\Delta_V = V - \tilde{V}$ we have that

$$
\begin{aligned}
\left\|(\psi_h - \widehat{\psi}_h^k)V\right\|_{\Lambda_h^k} &\le \left\|\lambda_p(\Lambda_h^k)^{-1}\psi_h V\right\|_{\Lambda_h^k} + \left\|(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(V)\right\|_{\Lambda_h^k} \\
&= \lambda_p\|\psi_h V\|_{(\Lambda_h^k)^{-1}} + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(V)\right\|_{(\Lambda_h^k)^{-1}} \\
&\le \lambda_p\|\psi_h V\|_{(\Lambda_h^k)^{-1}} + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(\tilde{V})\right\|_{(\Lambda_h^k)^{-1}} + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(\Delta_V)\right\|_{(\Lambda_h^k)^{-1}} \\
&\le \lambda_p\|\psi_h V\|_{(\Lambda_h^k)^{-1}} + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(\tilde{V})\right\|_{(\Lambda_h^k)^{-1}} + 2k\epsilon \qquad (\eta_{\tau,h}(\Delta_V) \le 2\|\Delta_V\|_\infty \le 2\epsilon) \\
&\le \sqrt{\lambda_p}\||\psi_h|(\mathcal{X})\|\|V\|_\infty + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(\tilde{V})\right\|_{(\Lambda_h^k)^{-1}} + 2k\epsilon \qquad (\Lambda_h^k \succeq \lambda_p I, \text{Cauchy-Schwarz}) \\
&\le \beta\sqrt{d\lambda_p} + \left\|\sum_{\tau=1}^{k-1}\phi_h^\tau \eta_{\tau,h}(\tilde{V})\right\|_{(\Lambda_h^k)^{-1}} + 2k\epsilon \qquad (\|V\|_\infty \le \beta, \||\psi_h|(\mathcal{X})\| \le \sqrt{d}) \\
&\le \beta\sqrt{d\lambda_p} + 2\beta\sqrt{d\log(K+1) + 2\log(H\mathcal{N}_\epsilon/\delta)} + 2k\epsilon \qquad (\text{Eq. } (34)) \\
&\le 4\beta\sqrt{d\log(K+1) + 2\log(H\mathcal{N}_\epsilon/\delta)}, \qquad (\epsilon \le \beta\sqrt{d}/2k, \lambda_p = 1)
\end{aligned}
$$

thus concluding the proof. ■

D.6. Covering numbers

The following results are (mostly) standard bounds on the covering number of several function classes.

Lemma 33. *For any $\epsilon > 0$, the ϵ-covering of the Euclidean ball in $\mathbb{R}^d$ with radius $R \ge 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

Lemma 34 (Lemma D.6 of (Jin et al., 2020b)). *Let $\mathcal{V}$ denote a class of functions $V : \mathcal{X} \to \mathbb{R}$ with the following parametric form*

$$V(\cdot) = \mathrm{clip}_R\left[\max_a \theta^\mathsf{T}\phi(\cdot,a) + \beta\sqrt{\phi(\cdot,a)^\mathsf{T}\Lambda^{-1}\phi(\cdot,a)}\right],$$

where $R \ge 0$ is a constant and $(\theta, \beta, \Lambda, R)$ are parameters satisfying $\|\theta\| \le L, \beta \in [0, B], \lambda_{\min}(\Lambda) \ge \lambda$ where $\lambda_{\min}$ denotes the minimal eigenvalue. Assume $\|\phi(x,a)\| \le 1$ for all (x,a) pairs, and let $\mathcal{N}_\epsilon$ be the $\epsilon-$covering number of $\mathcal{V}$ with respect to the supremum distance. Then

$$\log\mathcal{N}_\epsilon \le d\log(1 + 4L/\epsilon) + d^2\log[1 + 8\sqrt{d}B^2/(\lambda\epsilon^2)].$$

Lemma 35. *Let $\mathcal{V} = \{V(\cdot;\theta) : \|\theta\| \le W\}$ denote a class of functions $V : \mathcal{X} \to \mathbb{R}$. Suppose that any $V \in \mathcal{V}$ is L-Lipschitz with respect to θ and supremum distance, i.e.,*

$$\|V(\cdot;\theta_1) - V(\cdot;\theta_2)\|_\infty \le L\|\theta_1 - \theta_2\|, \quad \|\theta_1\|, \|\theta_2\| \le W.$$

Let $\mathcal{N}_\epsilon$ be the $\epsilon-$covering number of $\mathcal{V}$ with respect to the supremum distance. Then

$$\log\mathcal{N}_\epsilon \le d\log(1 + 2WL/\epsilon)$$

Proof. Let $\Theta_{\epsilon/L}$ be an (ϵ/L)-covering of the Euclidean ball in $\mathbb{R}^d$ with radius W. Define $\mathcal{V}_\epsilon = \{V(\cdot;\theta) : \theta \in \Theta_{\epsilon/L}\}$. By Lemma 33 we have that $\log|\mathcal{V}_\epsilon| \le d\log(1 + 2WL/\epsilon)$. We show that $\mathcal{V}_\epsilon$ is an ϵ-cover of $\mathcal{V}_\epsilon$, thus concluding the proof. Let $V \in \mathcal{V}$ and θ be its associated parameter. Let $\theta' \in \Theta_{\epsilon/L}$ be the point in the cover nearest to θ and $V' \in \mathcal{V}$ its associated function. Then we have that

$$\|V(\cdot) - V'(\cdot)\|_\infty = \|V(\cdot;\theta) - V(\cdot;\theta')\|_\infty \le L\|\theta - \theta'\| \le L(\epsilon/L) = \epsilon.$$

■