Near-Optimal Regret Bounds for Multi-batch Reinforcement Learning

Zihan Zhang* **Yuhang Jiang**[†] **Yuan Zhou**[‡] **Xiangyang Ji** [§]

Abstract

In this paper, we study the episodic reinforcement learning (RL) problem modeled by finite-horizon Markov Decision Processes (MDPs) with constraint on the number of batches. The multi-batch reinforcement learning framework, where the agent is required to provide a time schedule to update policy before everything, which is particularly suitable for the scenarios where the agent suffers extensively from changing the policy adaptively. Given a finite-horizon MDP with S states, A actions and planning horizon H, we design a computational efficient algorithm to achieve near-optimal regret of $\tilde{O}(\sqrt{SAH^3K\ln(1/\delta)})$[5] in K episodes using $O\left(H+\log_2\log_2(K)\right)$ batches with confidence parameter δ. To our best of knowledge, it is the first $\tilde{O}(\sqrt{SAH^3K})$ regret bound with $O(H+\log_2\log_2(K))$ batch complexity. Meanwhile, we show that to achieve $\tilde{O}(\mathrm{poly}(S,A,H)\sqrt{K})$ regret, the number of batches is at least $\Omega\left(H/\log_A(K)+\log_2\log_2(K)\right)$, which matches our upper bound up to logarithmic terms.

Our technical contribution are two-fold: 1) a near-optimal design scheme to explore over the unlearned states; 2) an computational efficient algorithm to explore certain directions with an approximated transition model.

1 Introduction

In reinforcement learning (RL), the learning agent interacts with the environment to maximize the total reward by making sequential decisions. The agent typically has to achieve two seemingly very different goals: to try as many actions and reach as many states as possible so as to learn more information about the environment (a.k.a. *exploration*) and to follow the policy that collects the high rewards according to the learned information (a.k.a. *exploitation*). To address this exploration-exploitation dilemma and achieve the near-optimal regret bounds, the agent usually needs to adjust his/her strategies *adaptively* based on the historical trajectories and make frequent policy changes [Azar et al., 2017, Zanette and Brunskill, 2019, Zhang et al., 2020].

On the other hand, however, too much adaptivity requirement usually leads to lower level of parallelism, impeding the large-scale deployment of the RL algorithms (which is often in a distributed manner). Frequent policy updates also suffer the cost of re-deploying policies in many practical applications. For example, in medical domains, it often requires complete discussion among many experts to change the treatment plans, which is not affordable in terms of both time and monetary cost [Lei et al., 2012, Almirall et al., 2012, 2014]; in RL for hardware placement [Mirhoseini et al., 2017], rewriting the program into the hardware for too many times is strongly discouraged. Similar

*Department of Automation, Tsinghua University, zihan-zh17@mails.tsinghua.edu.cn

[†]Department of Automation, Tsinghua University, jiangyh19@mails.tsinghua.edu.cn

[‡]Yau Mathematical Sciences Center & Department of Mathematical Sciences, Tsinghua University, yuan-zhou@tsinghua.edu.cn

[§]Department of Automation, Tsinghua University, xyji@tsinghua.edu.cn

[5] $\tilde{O}(\cdot)$ hides logarithmic terms of (S,A,H,K)

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

challenges also arise in applying RL to personalized recommendation system [Yu et al., 2019] and database optimization [Krishnan et al., 2018].

In such cases, the learning agent should minimize the number of policy switches while keeping the regret affordable. Bai et al. [2019] first proposed the provably efficient RL algorithms with low switching costs under the Q-learning algorithmic framework together with the lazy update techniques. However, their method needs to actively monitor the data in real time to determine whether a policy change is to be initiated. In other words, although the number of policy switches by [Bai et al., 2019] is low, the (usually long) time periods when the same policy is used still cannot be parallelized due to the policy-change trigger in their algorithms which is intrinsically sequential.

In order to address this problem, we propose and study under the framework of *multi-batch RL*, where the learning agent has to determine the number of batches and length of each batch before the learning process starts,[6] and uses as few batches as possible to achieve a low regret. Multi-batch RL algorithms can be easily deployed in a distributed fashion as the episodes during the same batch can be easily and fully parallelized. The idea of batch learning is also being widely practiced. For example, in medical trials, the medical center usually collects the data during a fixed time period among a batch of patients and then designs the experiment for the next phase based on the learned information in previous phases [Lei et al., 2012, Almirall et al., 2012, 2014].

Formally, we define multi-batch RL and *batch complexity* as below.

Definition 1 (Multi-Batch RL with complexity M)**.** *The agent determines a group of lengths $\{t_m\}_{m=1}^{M}$ such that $\sum_{m=1}^{M} t_m = K$ before the learning process starts. For $m = 1, 2, \ldots, M$, the agent sets a policy π^m and then follows π^m for t_m episodes.*

We highlight that an upper bound for batch complexity implies the same upper bound for global switching cost, since each policy switch means a new batch. It is also worth noting that the proposed batch RL framework is fully parallelizable during each batch for the applications where dataset comes in batch (e.g., clinical trial). Like other RL settings, we have the natural and interesting question:

Question 1. *Is it possible to achieve near optimal batch complexity, while keeping the regret $\tilde{O}(\sqrt{SAH^3K})$.*

We provide a positive answer for Question 1, which we state as below.

Theorem 1. *Let[7] $\iota = \ln(2/\delta)$. For any episodic MDP, with probability $1-\delta$, under Algorithm 1 the regret in T episodes is bounded by*

$$\mathrm{Regret}(T) \leqslant \tilde{O}\left(\sqrt{SAH^3K\iota^2} + S^{\frac{15}{4}}A^{\frac{9}{8}}H^{\frac{17}{8}}\iota^{\frac{5}{8}}K^{\frac{3}{8}} + S^{\frac{19}{4}}A^{\frac{13}{4}}H^{\frac{33}{4}}\iota K^{\frac{1}{4}} + S^{\frac{11}{2}}A^{\frac{9}{2}}H^{\frac{17}{2}}\iota\right),$$

and the batch complexity is bounded by $O(H + \log_2\log_2(K))$. Moreover, the computational cost of Algorithm 1 is $\tilde{O}(S^4AHK^3 + S^3A^2H^2K^3)$.

On the other hand, we show a lower bound of batch complexity as below.

Theorem 2. *For any algorithm with $O(\mathrm{poly}(S, A, H)\sqrt{K})$ regret bound, the batch complexity is at least $\Omega(H/\log_A(K) + \log_2\log_2(K))$.*

Compared to the lower bound of $\Omega(\log_2\log_2(K))$ in [Gao et al., 2019] for multi-armed bandit problem, additional $\Omega(H/\log_A(K))$ batches are required to explore the structure of the MDP.

Due to space limitation, we defer the full proofs of Theorem 1 and Theorem 2 to Appendix D and Appendix B respectively.

Our contribution. We propose the framework of multi-batch RL, and first achieve $O(H + \log_2\log_2(K))$ sample complexity bound with the near-optimal $\tilde{O}(\sqrt{SAH^3K\iota})$ regret bound with an efficient algorithm. We also prove that for any algorithm with $O(\mathrm{poly}(S, A, H)\sqrt{K})$ regret, the global switching cost is at least $\Omega(H/\log_A(K) + \log_2\log_2(K))$, which implies a nearly matching lower bound of $\Omega(H/\log_2(K) + \log_2\log_2(K))$ for the batch complexity. We also note that the $O(H + \log_2\log_2(K))$ batch complexity implies an $O(H + \log_2\log_2(K))$ bound for the global switching cost, which is also a near optimal upper bound.

[6] In contrast, Bai et al. [2019] can update the policy at any time.

[7] Throughout the paper we use ι to denote $\ln(2/\delta)$.

2 Related Works

Bandit algorithms with limited adaptivity. Bandit problem with low switching cost is widely studied in past decades [Cesa-Bianchi et al., 2013, Perchet et al., 2016, Gao et al., 2019, Simchi-Levi and Xu, 2019]. Cesa-Bianchi et al. [2013] showed an $\tilde{\Theta}(K^{\frac{2}{3}})$ regret bound under adaptive adversaries and bounded memories. Perchet et al. [2016] proved a regret bound of $\tilde{\Theta}(K^{\frac{1}{1-2^{1-M}}})$ for the two-armed bandit problem within M batches, and later Gao et al. [2019] extended their result to the general A-armed case. Besides the setting of classical multi-armed bandit problem, other settings has also been studied, e.g., multinomial bandit problem [Dong et al., 2020] and linear bandit problem [Ruan et al., 2020].

Episodic reinforcement learning with low switching cost. For model-based algorithms, by doubling updates, the global switching cost is $O(SAH\log_2(K))$ while keeping the regret $\tilde{O}(\sqrt{SAKH^3})$Azar et al. [2017]. For model-free algorithms, Bai et al. [2019] first studied RL with low switching cost. They proposed a Q-learning algorithm with lazy update to achieve $\tilde{O}(\sqrt{SAKH^4})$ regret bound and $O(SAH^3\log(K/A))$ local switching cost. Recently Zhang et al. [2020] established a better regret bound of $\tilde{O}(\sqrt{SAKH^3})$ and $O(SAH^2\log(K/A))$ local switching cost. Besides, Gao et al. [2021] generalized the problem to Linear RL, and established a regret bound of $\tilde{O}(\sqrt{d^3H^4K})$ with $O(dH\log(K))$ global switching cost. Recent work Qiao et al. [2022] achieved $O(HSA\log_2\log_2(K))$ switching cost and $\tilde{O}(\mathrm{poly}(S,A,H)\sqrt{K})$ regret with a computational inefficient algorithm.

Regret minimization for reinforcement learning. There is a long line of works devoting to regret minimization for RL problem [Kakade, 2003, Jaksch et al., 2010, Bartlett and Tewari, 2009, Dann et al., 2019, Azar et al., 2017, Jin et al., 2018, Zanette and Brunskill, 2019, Zhang and Ji, 2019, Zhang et al., 2020, Li et al., 2020, Zhang et al., 2021]. For tabular setting, near optimal regret bound of $\tilde{O}(\sqrt{SAH^3T})$ has been established by [Azar et al., 2017, Zanette and Brunskill, 2019, Zhang et al., 2020] for both model-based and model-free algorithms. However, fewer algorithms focused on the setting of multi-batch RL.

3 Preliminaries

Episodic reinforcement learning. $M = \langle \mathcal{S}, \mathcal{A}, r, P, s_1 \rangle$, where $\mathcal{S}\times\mathcal{A}$ is the discrete state-action space, $r = \{r_h(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A},h\in[H]}$ is the known[8] reward function, $P = \{P_h(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A},h\in[H]}$ is the unknown transition model and s_1 is the fixed initial state[9]. We assume that the reward function $r_h(s,a)\in[0,1]$ for any (h,s,a). In each episode, the agent starts at s_1, then takes actions and transits to the next state step by step, and finally conducts the trajectory $\{(s_h,a_h,s_{h+1})\}_{h=1}^H$. The target of the agent is to maximize the accumulative reward function $\sum_{h=1}^H r_h(s_h,a_h)$.

A policy π can be viewed as a series of mappings $\{\pi_h\}_{h=1}^H$ where $\pi_h : \mathcal{S}\to\Delta^{\mathcal{A}}$ maps s_h to a distribution over the action space at the h-th step, where $\pi_h(a|s)$ is the probability taking action a at state s of the h-th horizon.

Given a policy π, the (optimal) Q-function and value function are given by

$$Q_h^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'}(s_{h'},a_{h'})\Big|(s_h,a_h)=(s,a)\right]; \qquad Q_h^*(s,a) = \sup_{\pi\in\Pi} Q_h^\pi(s,a);$$

$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'}(s_{h'},a_{h'})\Big|s_h=s\right]; \qquad V_h^*(s) = \max_a Q_h^*(s,a).$$

[8] This is a common assumption since the uncertainty of reward function is dominated by that of the transition model.

[9] The more general case, where the agent starts from a fixed initial distribution, could be reduced to our setting by increasing H by 1

Let $\pi^{(k)}$ denote the policy in the k-th episode. Then the regret is given by

$$\text{Regret}(K) := \sum_{k=1}^{K} (V_1^*(s_1) - V_1^{\pi^{(k)}}(s_1)). \tag{1}$$

Notations In this paper, we use $\mathbb{E}_{\pi,p}[\cdot]$ ($\mathbb{P}_{\pi,p}[\cdot]$) to denote the expectation (probability) following policy π under transition model p. In particular, $\mathbb{E}_{\pi}[\cdot](\mathbb{P}_{\pi}[\cdot])$ denotes the expectation (probability) following π under the true transition model P. We define the general value function

$$W^{\pi}(r', p) = \mathbb{E}_{\pi,p}\left[\sum_{h=1}^{H} r'_h(s_h, a_h)\right].$$

We use $\mathbf{1}$ to denote the S-dimensional vector $[1, 1, \ldots, 1]^{\top}$ and $\mathbf{1}_{h,s,a}$ to denote the reward function r' such that $r'_{h'}(s', a') = \mathbb{I}[(h, s, a) = (h', s', a')]$. We also define $\{d_h^{\pi}(s, a)\}_{(s,a,h)}$ be the occupancy distribution of π. That is, $d_h^{\pi}(s, a) = \mathbb{E}_{\pi}[\mathbb{I}[(s_h, a_h) = (s, a)]]$. Δ^d is used to denote the d-dimensional simplex. For two vector x, y with the same dimension, we write $x^{\top}y$ as xy for convenience. For $p \in \Delta^S$ and $v \in \mathbb{R}^S$, we define $\mathbb{V}(p, v) = pv^2 - (pv)^2$. For $N \geqslant 1$, we use $[N]$ to denote the set $[1, 2, \ldots, N]$.

4 Technique Overview

In this section, we first introduce the policy elimination framework, which enjoys the near-optimal batch complexity. Then we summarize the technical challenges to achieve the near-optimal regret bound efficiently under this framework. At last, we introduce our major technical contributions.

4.1 Policy Elimination Framework

Following the methods in multi-batch bandit learning Perchet et al. [2016], Gao et al. [2019], we construct our main algorithm using policy elimination. Like most model-based reinforcement learning methods, we maintain a confidence region $\mathcal{P}$ for the transition model, where the true transition model $P \in \mathcal{P}$ with high probability. Before each batch starts, for a policy π and a reward function u, by extended value iteration (See Algorithm 5 in Appendix C.2), we are able to compute the confidence interval $[L^{\pi}(u, \mathcal{P}), U^{\pi}(u, \mathcal{P})]$ for the value function of π, where

$$U^{\pi}(u, \mathcal{P}) := \max_{p' \in \mathcal{P}} W^{\pi}(u + \mathbf{1}_z, p'); \qquad L^{\pi}(u, \mathcal{P}) := \min_{p' \in \mathcal{P}} W^{\pi}(u, p'). \tag{2}$$

Here z is a virtual state for the *infrequent* state-action-state triples (See Function `clip` in Algorithm 2). The reason why we give reward 1 for z in computing the upper confidence bound is to encourage exploration to these *infrequent* state-action-state triples.

By policy elimination we get $\Pi(r, \mathcal{P}) = \left\{\pi \middle| U^{\pi}(r, \mathcal{P}) \geqslant \sup_{\pi'} L^{\pi'}(r, \mathcal{P})\right\}$ as the set of survived policies. The next step is to choose a policy $\pi \in \Pi(r, \mathcal{P})$ and execute π in the current batch. Defining $\mathcal{P}^m$ to be the confidence region for the transition model after the m-th batch and $\text{gap}^{m+1} = \max_{\pi \in \Pi(r, \mathcal{P}^m)}(U^{\pi}(r, \mathcal{P}^m) - L^{\pi}(r, \mathcal{P}^m))$, the regret in the $m + 1$-th batch could be bounded by $t^{m+1}\text{gap}^{m+1}$. Therefore, the main task is to design efficient exploration policy to reduce gap^m for each $1 \leqslant m \leqslant M$.

4.2 Technical Challenges

Following the policy elimination framework above, we have two major challenges to achieve the near-optimal regret bound with an efficient algorithm.

Difficulty in exploration Fix the reward function r and confidence region $\mathcal{P}$. To construct tight confidence interval for every policy $\pi \in \Pi(r, \mathcal{P})$, we need to find a policy $\pi \in \Pi(r, \mathcal{P})$ to collect enough samples for each (h, s, a). To address the problem, Qiao et al. [2022] proposed an algorithm named APEVE, which learns each (h, s, a) triple independently. More precisely, for each $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the algorithm searches for a policy $\pi^{h,s,a}$ to maximize the probability of visiting (h, s, a)

over $\Pi(r, \mathcal{P})$, and then execute $\pi^{h,s,a}$ to collect samples for (h, s, a). However, this algorithm might be inefficient in sampling, since different horizon-state-action triples may match along with the same exploration policy. As shown in Qiao et al. [2022], the regret bound might be sub-optimal with this algorithm. Therefore, to achieve the near-optimal regret bound, we need to design a new exploration strategy to utilize the correlationship among different horizon-state-action triples.

Difficulty in efficient implementation Because the policy set $\Pi(r, \mathcal{P})$ might have exponential size, naive enumeration is not applicable to searching for a good exploration policy. As a consequence, it requires additional efforts to study the structure of $\Pi(r, \mathcal{P})$. For example, when $r = 0$, $\Pi(r, \mathcal{P})$ is the set of all possible policies. In this case, we can use extended value iteration (See Algorithm 5) to find the policy which visits (h, s, a) most frequently.

4.3 Key Techniques

Near-optimal design scheme Unlike RL algorithm with limited switching cost, in multi-batch reinforcement learning, the agent can not change the policy adaptively. As a result, we need to design a policy with proper coverage ratio for all the survived policies. That is, using the data collected following this policy, the length of the confidence interval for any survived policy is bounded by a uniform threshold.

Recall that $d_h^\pi(s,a) = \mathbb{E}_\pi[\mathbb{I}[(s_h, a_h) = (s,a)]$. Using classical regret analysis for tabular RL [Azar et al., 2013, Zanette and Brunskill, 2019], for a fixed policy π, the length of confidence interval for π could be roughly bounded by

$$\tilde{O}\left(\sum_{s,a,h} d_h^\pi(s,a)\sqrt{\frac{\mathrm{Var}_h(s,a)}{N_h(s,a)}}\right) \underset{\text{Cauchy's ineq.}}{\leqslant} \tilde{O}\left(\sqrt{\sum_{s,a,h}\frac{d_h^\pi(s,a)}{N_h(s,a)}}\cdot\sqrt{\sum_{s,a,h} d_h^\pi(s,a)\mathrm{Var}_h(s,a)}\right), \tag{3}$$

where $\mathrm{Var}_h(s,a)$ is the variance term with respect to $P_{h,s,a}$ and $V *_{h+1} (\cdot)$, and $N_h(s,a) \geqslant 1$ is the count of (h, s, a).

Because $\sum_{s,a,h} d_h^\pi(s,a)\mathrm{Var}_h(s,a)$ could be uniformly bounded by $O(H^2)$ using classical analysis, we focus on bounding the term $\sum_{s,a,h}\frac{d_h^\pi(s,a)}{N_h(s,a)}$. Suppose the policy for current batch is $\tilde{\pi}$. After this batch, we roughly have that $N_h(s,a) \propto d_h^{\tilde{\pi}}(s,a)$. So it corresponds to find a policy $\tilde{\pi} \in \Pi(r, \mathcal{P})$ to minimize the *worst-case coverage number* $\max_{\pi\in\Pi(r,\mathcal{P})}\sum_{h,s,a}\frac{d_h^\pi(s,a)}{d_h^{\tilde{\pi}}(s,a)}$. For this problem, we have the lemma below, and the proof is deferred to Appendix E.1.

Lemma 1. *Let $d > 0$ be an integer. Let $\mathcal{X} \subset (\Delta^d)^m$. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X}$, such that*

$$\max_{x=\{x_i\}_{i=1}^{dm}\in\mathcal{X}} \sum_{i=1}^{dm}\frac{x_i}{y_i} = md,$$

where $y = \{y_i\}_{i=1}^{dm} = \mathbb{E}_{x\sim\mathcal{D}}[x]$. Moreover, if $\mathcal{X}$ has a boundary set $\partial\mathcal{X}$ with finite cardinality, we can find an approximation solution for $\mathcal{D}$ in $\mathrm{poly}(|\partial\mathcal{X}|)$ *time.*

Plugging $\mathcal{X} = \left\{\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H | \pi \in \Pi(r,\mathcal{P})\right\}$, $d = SA$ and $m = H$ into Lemma 1, there exists a policy $\tilde{\pi}$ being a mixture of policies in $\Pi(r, \mathcal{P})$, such that $\max_{\pi\in\Pi(r,\mathcal{P})}\sum_{s,a,h}\frac{d_h^\pi(s,a)}{d_h^{\tilde{\pi}}(s,a)} = SAH$. In this way, we can find the desired exploration policy $\tilde{\pi}$ by assuming the knowledge of $\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H$ for all $\pi \in \Pi(r, \mathcal{P})$.

Given the design scheme above, it remains two problems, for which we present solutions below: 1) $\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H$ *is unknown;* 2) *even assuming* $\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H$ *is known, it is hard to find* $\tilde{\pi}$ *since the cardinality of* $\left\{\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H | \pi \in \Pi(r,\mathcal{P})\right\}$ *might be exponential in* SH.

Constructing tight confidence region To estimate $\{d_h^\pi(\cdot,\cdot)\}_{h=1}^H$, we consider to construct a tight confidence region for the transition model to estimate the occupancy distribution up to a constant ratio.

Definition 2. *We say a confidence transition region $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$ is* tight *with respect to p' iff (i)$p' \in \mathcal{P}$; (ii) $e^{-\frac{1}{H}}p'_{h,s,a,s'} \leqslant p_{h,s,a,s'} \leqslant e^{\frac{1}{H}}p'_{h,s,a,s'}$ for any (h,s,a,s') and any $p_{h,s,a} \in \mathcal{P}_{h,s,a}$; (iii) $\mathcal{P}_{h,s,a}$ has the form $\mathcal{P}_{h,s,a} = \{p \in \Delta^S | a_i^\top p \leqslant b_i, i = 1,2,...,m\}$ where $m \leqslant \mathrm{poly}(SM)$.*

In model-based reinforcement learning, these conditions are natural and it is easy to construct a *tight* confidence region with acceptable error.

Once we have a confidence region which is *tight* w.r.t. the true transition model P, for any policy π and (h,s,a), we can estimate the expected visit count $W^\pi(\mathbf{1}_{h,s,a})$ by $W^\pi(\mathbf{1}_{h,s,a},p)$ for any $p \in \mathcal{P}$ because

$$e^{-1}W^\pi(\mathbf{1}_{h,s,a},p) \leqslant W^\pi(\mathbf{1}_{h,s,a}) = d_h^\pi(s,a) \leqslant eW^\pi(\mathbf{1}_{h,s,a},p).$$

With $W^\pi(\mathbf{1}_{h,s,a},p)$ as approximation of $d_h^\pi(s,a)$, we can continue the analysis above by paying a constant factor.

To learn such a confidence region, by Bennet's inequality (Lemma 3), it suffices to visit (h,s,a,s')[10] for $C_1 H^2 \iota$ for each (h,s,a,s'), where C_1 is an universal constant. By this idea, we try to visit each (h,s,a,s') as much as possible. In the meantime, it is very possible that some (h,s,a,s') tuples are extremely hard to visit. Fortunately, with proper exploration scheme, we can show that the maximal probability to visit such tuples is well-bounded, so that these tuples could be ignored by suffering regret $O(\sqrt{T})$.

Computational efficient design scheme Assume the confidence region $\mathcal{P}$ is *tight* w.r.t. P. We invoke reward-zero exploration to learn a sub-optimal solution for the problem $\min_{\tilde\pi\in\Pi(r,\mathcal{P})}\max_{\pi\in\Pi(r,\mathcal{P})}\sum_{h,s,a}\frac{d_h^\pi(s,a)}{d_h^{\tilde\pi}(s,a)}$. Let $p \in \mathcal{P}$ be fixed and define $\tilde d_h^\pi(s,a) = W^\pi(\mathbf{1}_{h,s,a},p)$ be the approximation for $d_h^\pi(s,a)$. We define $\tilde\pi^i = \arg\max_{\pi\in\Pi(r,\mathcal{P})} W^\pi(r^i,p)$ for $1 \leqslant i \leqslant k = K^3$, where $r_h^i(s,a) = \min\left\{\frac{1}{\sum_{j=1}^{i-1}\tilde d_h^{\tilde\pi^j}(s,a)},1\right\}$. Let $\tilde\pi$ be the mixture of $\{\tilde\pi^i\}_{i=1}^k$. For any policy π, we have that

$$\begin{aligned}
\sum_{s,a,h} d_h^\pi(s,a)\cdot\min\left\{\frac{1}{d_h^{\tilde\pi}(s,a)},k\right\} &\leqslant O\left(\sum_{s,a,h}\tilde d_h^\pi(s,a)\cdot\min\left\{\frac{1}{\tilde d_h^{\tilde\pi}(s,a)},k\right\}\right) &(4)\\
&\leqslant O\left(\sum_{i=1}^k W^\pi(r^i,p)\right) &(5)\\
&\leqslant O\left(\sum_{i=1}^k W^{\tilde\pi^i}(r^i,p)\right) &(6)\\
&\leqslant O\left(\sum_{s,a,h}\sum_{i=1}^k d_h^{\tilde\pi^i}(s,a)\cdot\min\left\{\frac{1}{\sum_{j=1}^{i-1}d_h^{\tilde\pi^j}(s,a)},1\right\}\right)\\
&\leqslant O\left(\sum_{s,a,h}\sum_{i=1}^k\log\left(\frac{\max\{\sum_{j=1}^{i}d_h^{\tilde\pi^j}(s,a),1\}}{\max\{\sum_{j=1}^{i-1}d_h^{\tilde\pi^j}(s,a),1\}}\right)\right)\\
&\leqslant O(SAH\log(k)). &(7)
\end{aligned}$$

Here (4) holds by the tightness of $\mathcal{P}$, (5) holds by the fact that $r_h^i(s,a) \geqslant r_h^{k+1}(s,a) = \min\left\{\frac{1}{\sum_{j=1}^k \tilde d_h^{\tilde\pi^j}(s,a)},1\right\} = \frac{1}{k}\min\left\{\frac{1}{\tilde d_h^{\tilde\pi}(s,a)},k\right\}$ for any (h,s,a), and (6) holds by the optimality of $\tilde\pi^i$ for $1 \leqslant i \leqslant k$. With (7) in hand, $\max_{\pi\in\Pi(r,\mathcal{P})}\sum_{h,s,a}\frac{d_h^\pi(s,a)}{d_h^{\tilde\pi}(s,a)}$ is roughly bounded by $O(SAH\log(K))$[11], which nearly matches the best *worst-case coverage number* number of SAH.

[10] A tuple (h,s,a,s') is visited means $(s_h,a_h,s_{h+1}) = (s,a,s')$.

[11] We remark the there is still a gap between $\max_{\pi\in\Pi(r,\mathcal{P})}\sum_{h,s,a}\frac{d_h^\pi(s,a)}{d_h^{\tilde\pi}(s,a)}$ and $\sum_{s,a,h}d_h^\pi(s,a)\cdot\min\left\{\frac{1}{d_h^{\tilde\pi}(s,a)},K^3\right\}$. Actually (7) is sufficient for further regret analysis.

Algorithm 1 Main Algorithm

```
Input: state-action space $\mathcal{S} \times \mathcal{A}$, number of episodes $K$, confidence parameter $\delta$;
Initialize: $\iota \leftarrow \ln(2/\delta)$, $k_1 \leftarrow 144\sqrt{SAKH\iota}$, $k_2 \leftarrow 288S^3A^2H^4\sqrt{K\iota}$;
$\{\mathcal{D}_1\} \leftarrow$ Raw Exploration$(0, \varnothing, k_1)$;
$\{\mathcal{D}_2\} \leftarrow$ Raw Exploration$(r, \mathcal{D}_1, k_2)$;
Policy Elimination$(\mathcal{D}_2, K - Hk_1 - Hk_2)$.
```

Computational efficient constrained exploration Let u, u' be two reward functions and $\mathcal{P}$ be a set of transition models. As stated before, for general $\Pi(u, \mathcal{P})$, it might be non-trivial to solve the problem $\tilde{\pi} = \arg\max_{\pi \in \Pi(u,\mathcal{P})} W^{\pi}(u', p)$ for fixed $p \in \mathcal{P}$. As a trade-off, we turn to find some policy $\tilde{\pi} \in \Pi(u, \mathcal{P})$ such that $W^{\tilde{\pi}}(u', p) \geqslant c \max_{\pi \in \Pi(u,\mathcal{P})} W^{\pi}(u', p)$, where $c > 0$ is some universal constant. The problem turns out to be a RL problem with a soft constraint. For general $\Pi(u, \mathcal{P})$, the problem might be hard to solve. Fortunately, on the benefit of the *tight* property of $\mathcal{P}$, we can find such $\tilde{\pi}$ efficiently.

5 Algorithms

In this section we present our algorithms. The main algorithm (Algorithm 1) consists of three stages.

In the first two stages, we conduct naive exploration to identify the tuples which are hard to visit, which we called *infrequent* tuples. In particular, the length of the second stage is slightly larger than that of the first stage, where we use the dataset in the first stage to reduce the regret in the second stage. In this way, we can bound the regret in the first two stages by $\tilde{O}(\sqrt{SAH^3K})$, while the probability of visiting the *infrequent* tuples is small enough.

After ignoring the *infrequent* tuples, we could obtain a *tight* confidence region. Given the *tight* confidence region, we compute the confidence region for each policy and conduct policy elimination in the third stage. The first and second stages contains $O(H)$ batches, and the third stage contains $O(\log_2 \log_2(K))$ batches. So the batch complexity of Algorithm 1 is $O(H + \log_2 \log_2(K))$. Below we describe Raw Exploration (Algorithm 2) and Policy Elimination (Algorithm 3) in detail.

5.1 Raw Exploration

Given a dataset $\mathcal{D}$ with counts $\{N_h(s, a, s')\}$, we define the set of *known* tuples as $\{(h, s, a, s') : N_h(s, a, s') \geqslant C_1H^2\iota\}$ and the left tuples are regarded as *infrequent* tuples.

In Algorithm 2, we are given a dataset. Then we compute the corresponding confidence region $\mathcal{P}$ in Line 20, where $\alpha(n, n') = \sqrt{\frac{4n'\iota}{n^2} + \frac{5\iota}{n}}$.

We conduct exploration layer by layer over policies in the set of survived policies $\Pi(r, \mathcal{P})$. By visiting each (h, s, a) as much as possible, we can judge whether a tuple (h, s, a, s') is hard to visit using policies in $\Pi(r, \mathcal{P})$.

Given the set of *known* tuples $\mathcal{W}$, we redirect all tuples not in $\mathcal{W}$ to an additional absorbed state z using clip$(\cdot, \cdot)$. Once we prove that the probability of reaching z is small enough for the any optimal policy, we can directly learn under the clipped transition model.

In Line 6 Algorithm 2, the algorithm Policy Search is invoked. Given any reward u, u', any confidence region $\mathcal{P}$ and threshold $\epsilon > 0$, this algorithm returns a policy $\tilde{\pi} \in \Pi(u, \mathcal{P})$ such that $W^{\tilde{\pi}}(u', p) \geqslant c \max_{\pi \in \Pi(u,\mathcal{P})} W^{\pi}(u', p) - \epsilon$ with some universal constant $c > 0$. Moreover, when $\mathcal{P}$ is *tight* w.r.t. the true transition model P after clipping, the time complexity of the algorithm is $O(\text{poly}(SAHK) \log(1/\epsilon))$. The algorithm and corresponding analysis is postponed to Appendix C.

It is also worth noting that executing each $\pi_{h,s,a}$ with probability $\frac{1}{SA}$ can not be regarded as a (history-independent) policy because the agent need to keep in mind which policy is chosen in current episode. In contrast, the agent only needs to observe current state to take actions following a policy. To address this problem, we define an operator Sum to take sum over policies under some transition model. Formally, we have the lemma below and postpone the proof to Appendix E.2.

Algorithm 2 Raw Exploration$(u, \mathcal{D}, k)$

```
**Input**: reward function $u$, dataset $\mathcal{D}$, length $k$;
**Initialize:** $C_1 \leftarrow 200$;
**for** $h = 1, 2, \ldots, H$ **do**
    $\mathcal{P} \leftarrow \mathtt{CR}(\mathcal{D})$;
    **for** $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
        $\pi^{h,s,a} \leftarrow \mathtt{Policy\,Search}(u, \mathbf{1}_{h,s,a}, \mathcal{P})$;
    **end for**
    $p \leftarrow$ arbitrary element in $\mathcal{P}$;
    $\{\tilde{\pi}^h, p\} \leftarrow \mathtt{Sum}\left(\left\{\frac{1}{SA}, \pi^{h,s,a}, p\right\}_{(h,s,a)}\right)$;
    $\pi^h$ be the policy which is the same as $\tilde{\pi}^h$ in the first $h-1$ steps, and be the uniformly random
    policy in the left $H - h + 1$ steps;
    Execute $\pi^h$ for $k$ episodes, and collect the samples as $\mathcal{D}_h$;
    $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_h$;
**end for**
**return:** $\{\mathcal{D}\}$;
**Function**: $\mathtt{CR}(\mathcal{D})$:
    $N_h(s, a, s') \leftarrow$ count of $(h, s, a, s')$ in $\mathcal{D}$, for all $(s, a, s')$;
    $N_h(s, a) \leftarrow \max\{\sum_{s'} N_h(s, a, s'), 1\}$ for all $(s, a)$;
    $\hat{p}_{h,s,a,s'} \leftarrow \frac{N_h(s,a,s')}{N_h(s,a)}, \forall(h, s, a, s')$;
    $\mathcal{W} \leftarrow \{(h, s, a, s') : N_h(s, a, s') \geqslant C_1 H^2 \iota\}$;
    $\tilde{\mathcal{P}}_{h,s,a} \leftarrow \left\{p \in \Delta^S \,|\, |p_{s'} - \hat{p}_{h,s,a,s'}| \leqslant \alpha(N_h(s, a), N_h(s, a, s')), \forall s' \in \mathcal{S}\right\}, \forall(h, s, a)$;
    $\mathcal{P}_{h,s,a} \leftarrow \{\mathtt{clip}(p, \mathcal{W}) : p \in \tilde{\mathcal{P}}_{h,s,a}\}, \forall(h, s, a)$;
    **Return**: $\otimes_{h,s,a}\mathcal{P}_{h,s,a}$.
**Function**: $\mathtt{clip}(p, \mathcal{W})$
    $p'_{h,s,a,s'} \leftarrow p_{h,s,a,s'}, \forall(h, s, a, s) \in \mathcal{W}$;
    $p'_{h,s,a,s'} \leftarrow 0, \forall(h, s, a, s') \notin \mathcal{W}$;
    $p'_{h,s,a,z} \leftarrow \sum_{s':(h,s,a,s') \notin \mathcal{W}} p_{h,s,a,s'}, \forall(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$;
    $p'_{h,z,a} \leftarrow \mathbf{1}_z, \forall(h, a) \in [H] \times \mathcal{A}$;
    **Return**: $p$.
```

Lemma 2. *Let $\mathcal{P} = \otimes_{(h,s,a)}\mathcal{P}_{h,s,a}$ be a set of transition models such that $\mathcal{P}_{h,s,a} \subset \Delta^S$ is convex for any (h, s, a). Let $\{(\pi^i, P^i)\}_{i=1}^n$ be a sequence of policy-transition pairs such that $P^i \in \mathcal{P}$. For any $\{\lambda_i\}_{i=1}^n$ such that $\lambda_i \geqslant 0$ for $i \geqslant 1$ and $\sum_i \lambda_i = 1$, there exists a policy π and $P \in \mathcal{P}$, satisfying that*

$$W^{\pi}(\boldsymbol{1}_{h,s,a}, P) = \sum_i \lambda_i W^{\pi^i}(\boldsymbol{1}_{h,s,a}, P^i) \tag{8}$$

for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Furthermore, the time complexity to find $\{\pi, P\}$ could be bounded by $O(nS^3A^2H^2)$.

Therefore, for any $\{\lambda_i, \pi^i, P^i\}_{i=1}^n$ satisfying $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i \geqslant 0$ for $i \geqslant 1$ as input, there exists $\{\pi, P\}$ such that $W^{\pi}(\mathbf{1}_{h,s,a}, P) = \sum_i \lambda_i W^{\pi^i}(\mathbf{1}_{h,s,a}, P^i)$ and $P_{h,s,a} \in \mathrm{Convex}(\{P^i_{h,s,a}\}_{i=1}^n)$ for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, where $\mathrm{Convex}(\mathcal{U})$ denotes the convex hull of the set $\mathcal{U}$. Then Sum is defined as $\mathtt{Sum}(\{\lambda_i, \pi^i, P^i\}_{i=1}^n) = \{\pi, P\}$.

5.2 Policy Elimination

Given the dataset collected in the first two stages, we first compute the *known* set $\mathcal{W}$. Unlike Algorithm 2, we do not update $\mathcal{W}$ in the rest time because the first two stages can ensure that the probability of visiting $\mathcal{W}^C$ is $O(1/\sqrt{K})$.

As mentioned in Section 4, for each batch, we invoke reward-zero exploration to search for the policy with near-optimal coverage. Based on such a policy, we can provide uniform bound for the length

Algorithm 3 `Policy Elimination`

```
1: Input: dataset $\mathcal{D}$, length $k$;
2: Initialize: $\mathcal{D}^0 \leftarrow \mathcal{D}$, $\mathcal{P}^{-1} \leftarrow (\Delta^S)^{SA}$ $C_1 \leftarrow 100$, $v_h^{-1}(s) \leftarrow H - h + 1, \forall (h,s) \in [H] \times \mathcal{S}$;
   $K_m \leftarrow \left\lceil K^{1-\frac{1}{2^m}} \right\rceil$ for $m = 1, 2, \ldots, M = \lceil \log_2 \log_2(K) \rceil$;
3: $N_h(s,a,s') \leftarrow$ count of $(h,s,a,s')$ in $\mathcal{D}$;
4: $\mathcal{W} \leftarrow \{(h,s,a,s') : N_h(s,a,s') \geqslant C_1 H^2 \iota\}$;
5: for $m = 0, 1, 2, \ldots, M-1$ do
6:     $\mathcal{P}^m \leftarrow \mathcal{P}^{m-1} \cap \texttt{CR*}\left(\mathcal{D}^m, \overline{\mathcal{D}}^m, \mathcal{W}, \{v_h^{m-1}(s)\}_{(h,s)}\right)$;
7:     $\pi^{m+1} \leftarrow \texttt{Design}((\mathcal{P}^m)$;
8:     if $\sum_{m'=1}^{m} K_{m'} \leqslant k$ then
9:         Execute $\pi^{m+1}$ for $K_{m+1}$ episodes;
10:    else
11:        Execute $\pi^{m+1}$ for $k - (\sum_{m'=1}^{m} K_{m'})$ episodes;
12:    end if
13:    $\overline{D}^{m+1} \leftarrow$ the dataset in the $(m+1)$-th batch;
14:    Update the dataset $\mathcal{D}^{m+1} \leftarrow \mathcal{D}^m \cup \overline{\mathcal{D}}^{m+1}$;
15:    $v_h^m(s) \leftarrow \max_{\pi, p \in \mathcal{P}^m} \mathbb{E}_{\pi,p}\left[\sum_{h'=h}^{H} r_h(s_h, a_h) | s_h = s\right]$ for all $(h,s) \in [H] \times \mathcal{S}$;
16: end for
17: Function: $\texttt{CR*}(\mathcal{D}, \mathcal{D}', \mathcal{W}, v)$:
18:    $\{N_h(s,a,s')\} \leftarrow$counts in $\mathcal{D}$, $N_h(s,a) \leftarrow \max\{\sum_{s'} N_h(s,a,s'), 1\}$ for all $(h,s,a,s')$;
19:    $\hat{p}_{h,s,a,s'} \leftarrow \frac{N_h(s,a,s')}{N_h(s,a)}, \forall (h,s,a,s')$;
20:    $\{\check{N}_h(s,a,s')\} \leftarrow$ counts in $\mathcal{D}'$, $\check{N}_h(s,a) \leftarrow \max\{\sum_{s'} \check{N}_h(s,a,s'), 1\}$ for all $(h,s,a,s')$;
21:    $\check{p}_{h,s,a,s'} \leftarrow \frac{\check{N}_h(s,a,s')}{\check{N}_h(s,a)}, \forall (h,s,a,s')$;
22:    $\tilde{\mathcal{P}}_{h,s,a} \leftarrow \Big\{p \in \Delta^S \,|\, |p_{s'} - \hat{p}_{h,s,a,s'}| \leqslant \alpha(N_h(s,a), N_h(s,a,s')), \forall s' \in \mathcal{S}$,
                    $|(p - \check{p}_{h,s,a})v| \leqslant \alpha^*(\check{N}_h(s,a), \check{p}_{h,s,a}, v)\Big\}, \forall (h,s,a)$;
23:    $\mathcal{P}_{h,s,a} \leftarrow \{\texttt{clip}(p, \mathcal{W}) : p \in \tilde{\mathcal{P}}_{h,s,a}\}, \forall (h,s,a)$;
24:    Return: $\otimes_{h,s,a} \mathcal{P}_{h,s,a}$.
25: Function: $\texttt{Design}(\mathcal{P})$:
26:    $p \leftarrow$ arbitrary element in $\mathcal{P}$;
27:    for $i = 1, 2, ..., K^3$ do
28:    $\tilde{d}_h^{\tilde{\pi}^j}(s,a) \leftarrow W^{\tilde{\pi}^j}(\mathbf{1}_{h,s,a}, p)$ for $1 \leqslant j \leqslant i-1$ and any $(h,s,a)$;
29:    $r_h^i(s,a) \leftarrow \min\left\{\frac{1}{\sum_{j=1}^{i-1} \tilde{d}_h^{\tilde{\pi}^j}(s,a)}, 1\right\}, \forall (h,s,a)$;
30:    $\tilde{\pi}^i \leftarrow \texttt{Policy Search}(r, r^i, \mathcal{P})$;
31:    end for
32:    $\{\pi, p\} \leftarrow \texttt{Sum}\left(\left\{\frac{1}{K^3}, \tilde{\pi}^i, p\right\}_{i=1}^{K^3}\right)$;
33:    Return: $\pi$.
```

of confidence intervals for all survived policies, which enables us to using the batch sizes in bandit algorithms [Perchet et al., 2016, Gao et al., 2019].

Besides, to obtain a better regret bound, we estimate the optimal value function at the end of each batch, and use it to build a tighter confidence region. As presented in Line 22 Algorithm 3, we use two empirical transition probabilities to construct the confidence region. Noting that the samples in the m-th batch is independent of v^{m-1}, we could add a Bernstein-style constraint, where $\alpha^*(n,p,v) = 5\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} + \frac{3\iota}{n}$..

6 Conclusion

In this paper, we study multi-batch reinforcement learning, and provide an efficient algorithm to achieve the near-optimal regret bound and batch complexity. It would be an interesting problem to generalize our results to reinforcement learning with function approximation case, e.g., linear MDP. Another important direction is to study the exact batch-regret trade-off for multi-batch reinforcement learning.

Broader Impact This work focus on the theory of multi-batch reinforcement learning, and the broader impact is not applicable.

References

Daniel Almirall, Scott N Compton, Meredith Gunlicks-Stoessel, Naihua Duan, and Susan A Murphy. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in medicine*, 31(17):1887–1902, 2012.

Daniel Almirall, Inbal Nahum-Shani, Nancy E Sherwood, and Susan A Murphy. Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine*, 4(3):260–274, 2014.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8004–8013, 2019.

Peter L Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009))*, 2009.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pages 1160–1168, 2013.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1507–1516, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Kefan Dong, Yingkai Li, Qin Zhang, and Yuan Zhou. Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615. PMLR, 2020.

Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Sham M Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Sanjay Krishnan, Zongheng Yang, Ken Goldberg, Joseph Hellerstein, and Ion Stoica. Learning to optimize join queries with deep reinforcement learning. *arXiv preprint arXiv:1808.03196*, 2018.

Huitan Lei, Inbal Nahum-Shani, Kevin Lynch, David Oslin, and Susan A Murphy. A" smart" design for building individualized treatment sequences. *Annual review of clinical psychology*, 8:21–48, 2012.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

Azalia Mirhoseini, Hieu Pham, Quoc V Le, Benoit Steiner, Rasmus Larsen, Yuefeng Zhou, Naveen Kumar, Mohammad Norouzi, Samy Bengio, and Jeff Dean. Device placement optimization with reinforcement learning. In *International Conference on Machine Learning*, pages 2430–2439. PMLR, 2017.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog (t) switching cost. *arXiv preprint arXiv:2202.06385*, 2022.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. *arXiv preprint arXiv:2007.01980*, 2020.

David Simchi-Levi and Yunzong Xu. Phase transitions and cyclic phenomena in bandits with switching constraints. *Available at SSRN 3380783*, 2019.

Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1910.12156*, 2019.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2823–2832, 2019.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [N/A] The paper is theoretical and there is no numerical experiments.
- Did you include the license to the code and datasets? [N/A] The paper is theoretical and there is no numerical experiments.
- Did you include the license to the code and datasets?[N/A] The paper is theoretical and there is no numerical experiments.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We provide a near-optimal regret bound for multi-batch RL
   (b) Did you describe the limitations of your work? [Yes] We focus on studying the tabular case. More efforts are required to extend the results to RL with function approximation
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The paper is theoretical and there is no possible negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] .
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We sketch the proof in the main body. The details are postpone to the appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] The paper is theoretical and there is no numerical experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] The paper is theoretical and there is no numerical experiments.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] This paper is irrelevant to crowdsourcing or human projects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Technical Lemmas

Lemma 3. *Let $Z, Z_1, ..., Z_n$ be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Define $\mathbb{V}Z = \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right]$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}\left[Z\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\mathbb{V}Z\ln(2/\delta)}{n}} + \frac{\ln(2/\delta)}{n}\right] \leqslant \delta.$$

Lemma 4. *Let $X_1, X_2, \ldots$ be a sequence of random variables taking value in $[0, l]$. Define $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k|\mathcal{F}_k]$ for $k \geqslant 1$. For any $\delta > 0$, we have that*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \leqslant 3\sum_{k=1}^{n} Y_k + l\ln(1/\delta)\right] \leqslant \delta$$

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geqslant 3\sum_{k=1}^{n} X_k + l\ln(1/\delta)\right] \leqslant \delta.$$

Proof. Let $t \in [0, 1/l]$ be fixed. Consider to bound $Z_k := \mathbb{E}[\exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))]$. By definition, we have that

$$\begin{aligned}
\mathbb{E}[Z_k|\mathcal{F}_k] &= \exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))\mathbb{E}\left[t(X_k - 3Y_k)\right] \\
&\leqslant \exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))\exp(3Y_k)\cdot\mathbb{E}[1 + tX_k + 2t^2X_k^2] \\
&\leqslant \exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))\exp(3Y_k)\cdot\mathbb{E}[1 + 3tX_k] \\
&= \exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))\exp(3Y_k)\cdot(1 + 3tY_k) \\
&\leqslant \exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'})) \\
&= Z_{k-1},
\end{aligned}$$

where the second line is by the fact that $e^x \leqslant 1 + x + 2x^2$ for $x \in [0,1]$. Define $Z_0 = 1$ Then $\{Z_k\}_{k\geqslant 0}$ is a super-martingale with respect to $\{\mathcal{F}_k\}_{k\geqslant 1}$. Let τ be the smallest n such that $\sum_{k=1}^{n} X_k - 3\sum_{k=1}^{n} Y_k > l\ln(1/\delta)$. It is easy to verify that $Z_{\min\{\tau,n\}} \leqslant \exp(tl\ln(1/\delta) + tl) < \infty$. Choose $t = 1/l$. By the optimal stopping time theorem, we have that

$$\begin{aligned}
&\mathbb{P}\left[\exists n \leqslant N, \sum_{k=1}^{n} X_k \geqslant 3\sum_{k=1}^{n} Y_k + l\ln(1/\delta)\right] \\
&= \mathbb{P}\left[\tau \leqslant N\right] \\
&\leqslant \mathbb{P}\left[Z_{\min\{\tau,N\}} \geqslant \exp(tl\ln(1/\delta))\right] \\
&\leqslant \frac{\mathbb{E}[Z_{\min\{\tau,N\}}]}{\exp(tl\ln(1/\delta))} \\
&\leqslant \delta.
\end{aligned}$$

Letting $N \to \infty$, we have that

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \leqslant 3\sum_{k=1}^{n} Y_k + l\ln(1/\delta)\right] \leqslant \delta.$$

Considering $W_k = \mathbb{E}[\exp(t\sum_{k'=1}^{k}(Y_k/3 - X_k))]$, using similar arguments and choosing $t = 1/(3l)$, we have that

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geqslant 3\sum_{k=1}^{n} X_k + l\ln(1/\delta)\right] \leqslant \delta.$$

The proof is completed. □

Lemma 5. *Let the policy π and reward r be fixed. Let p and p' be two transition model, it holds that*

$$W^{\pi}(r,p) - W^{\pi}(r,p') = \sum_{h,s,a} W^{\pi}(\boldsymbol{1}_{h,s,a}, p)(p'_{h,s,a} - p_{h,s,a})V'_{h+1}, \tag{9}$$

where $\{V'_h(s)\}_{(h,s)\in[H]\times\mathcal{S}}$ is the value function under p' following π.

B Lower Bound (Proof of Theorem 2)

Firstly, by the lower bound on batched bandit (Theorem 3 in [Gao et al., 2019]), to achieve $O(\text{poly}(S,A,H)\sqrt{K})$ regret, the number of batches is at least $\Omega(\log_2\log_2(K))$. To show a lower bound of $\Omega(H/\log_A(K))$, we have the lemma below by considering an MDP with 2 states and A actions.

Lemma 6. *Let $\mathcal{S} = \{s^{(0)}, s^{(1)}\}$, $\mathcal{A} = \{a_0, a_1, \ldots, a_A\}$ and $s_1 = s^{(0)}$. Let $d = \lfloor 2\log_A(K)\rfloor + 2$. For $v = [v_1, v_2, \ldots, v_d]^{\top} \in A^d$, we define the transition model P^v by setting $P^v_{h,s^{(0)},a_x} = [1,0]^{\top}, \forall x \neq v_h$, $P^v_{h,s^{(0)},a_{v_h}} = [0,1]^{\top}$ and $P^v_{h,s^{(1)},a_x} = [0,1]^{\top}, \forall 1 \leqslant x \leqslant A$ for $1 \leqslant h \leqslant d$. Let π be a stochastic policy, Then there exists v such that with probability $1 - \frac{1}{K}$, $(h, s^{(0)})$ is never visited in K episodes following π.*

Proof. Denote the distribution of π as $\mathtt{D}$, we define $x = [x_1, x_2, \ldots, x_d]^{\top}$ as below. Let $x_1 = \arg\min_i \mathbb{E}_{\pi\sim D}\left[\pi_1(a_i|s^{(0)})\right]$. For $2 \leqslant h \leqslant d$, we define

$$x_h = \arg\max_i \frac{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h-1} = s^{(0)}|P^{x,h-1}]\pi_h(a_i|s_h)\right]}{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h-1} = s^{(0)}|P^{x,h-1}]\right]},$$

where $P^{x,h-1}$ denote the first $(h-1)$-layers of the transition model P^x. Because $\mathbb{P}_{\pi}[s_h = s^{(0)}|P^x]$ is determined by the first $(h-1)$-layers of P^x, x_h is well-defined. By definition we have that

$$\frac{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h-1} = s^{(0)}|P^{x,h-1}]\pi_h(a_{x_h}|s_h)\right]}{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h-1} = s^{(0)}|P^{x,h-1}]\right]} \leqslant \frac{1}{A}. \tag{10}$$

Recall that $x = [x_1, x_2, \ldots, x_d]^{\top}$. For $1 \leqslant h' \leqslant d$, by (10) we have that

$$\begin{aligned}
&\mathbb{E}_{\pi\sim\mathtt{D}}\mathbb{P}_{\pi}\left[s_{h'} = s^{(0)}|P^x\right] \\
&= \mathbb{E}_{\pi\sim\mathtt{D}}\Pi_{h=1}^{h'-1}\pi_h(a_{x_h}|s^{(0)}) \\
&= \mathbb{E}_{\pi\sim\mathtt{D}}\mathbb{P}_{\pi}\left[s_{h'-1} = s^{(0)}|P^x\right] \cdot \frac{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h'-1} = s^{(0)}|P^{x,h-1}]\pi_h(a_{x_h}|s_h)\right]}{\mathbb{E}_{\pi\sim\mathtt{D}}\left[\mathbb{I}_{\pi}[s_{h'-1} = s^{(0)}|P^{x,h-1}]\right]} \\
&\leqslant \frac{1}{A}\mathbb{E}_{\pi\sim\mathtt{D}}\mathbb{P}_{\pi}\left[s_{h'-1} = s^{(0)}|P^x\right].
\end{aligned} \tag{11}$$

Therefore, $\mathbb{E}_{\pi\sim\mathtt{D}}\mathbb{P}_{\pi}\left[s_d = s^{(0)}|P^x\right] \leqslant \frac{1}{A^{d-1}} \leqslant \frac{1}{K}$. Then the probability of visiting $(h, s^{(0)})$ in K episodes is at most $\frac{1}{K}$, where the conclusion follows. □

We name the MDP in Lemma 6 as a basic MDP. Now we construct our counter-example by concatenating $\Theta(H/\log_A(K))$ basic MDPs and a tail MDP with large rewards. Let $\mathcal{S} = \{s^{(0)}, s^{(1)}\}$ and $\mathcal{A} = \{a_1, a_2, \ldots, a_A\}$. Let $d = \lfloor 2\log_A(K)\rfloor + 2$ and $c = \left\lfloor \frac{H}{2d}\right\rfloor$. Then $c = C'H/\log_A(K)$ for some constant C'. For $v = [v_1, v_2, \ldots, v_{cd}]^{\top} \in \{0,1\}^{cd}$, we define the transition model P^v as below: $P^v_{id+j,s^{(0)},a_{v_{id+j}}} = [1,0]^{\top}$, $P^v_{id+j,s^{(0)},a_l} = [0,1]^{T}$ for $l \neq v_{id+j}$ and $P^v_{id+j,s^{(1)},a_l} = [0,1]^{\top}$

for $1 \leqslant l \leqslant A$ for any $0 \leqslant i \leqslant c-1$ and $1 \leqslant j \leqslant d$; $P_{h,s^{(0)},a_l} = [1,0]^\top$ and $P_{h,s^{(1)},a_l} = [0,1]^\top$ for any $1 \leqslant l \leqslant A$ and $cd+1 \leqslant h \leqslant H$. The reward function r is given by $r_{h,s^{(0)},a_l}$ for $1 \leqslant l \leqslant A$ $cd+1 \leqslant h \leqslant H$ and 0 for other (h,s,a) triples.

To achieve sub-linear regret, the agent needs to visit $(cd+1, s^{(0)})$ for at least one time. Then the proof is completed by the lemma below.

Lemma 7. *If the number of batches $M \leqslant c-2$, for any algorithm $\mathcal{G}$ there exists v such that with probability $1-\frac{c}{K} \geqslant \frac{1}{2}$, $(cd+1, s^{(0)})$ is never visited.*

Proof. Let m_i denote the number of batches used at the time when $(id+1, s^{(0)})$ is visited for the first time. Besides, we let $\pi(i)$ denote the policy at time m_i. Because $\pi(i)$ is determined before visiting $(id+1, s^{(0)})$, given the algorithm $\mathcal{G}$, π^i could be viewed as a stochastic function of $\{v_1, v_2, \ldots, v_{id}\}$. By Lemma 6, when $\{v_1, v_2, \ldots, v_{id}\}$ is fixed, we can choose $\{v_{id+1}, \ldots, v_{id+d}\}$ properly, so that with probability $1-\frac{1}{K}$, $((i+1)d+1, s^{(0)})$ is never visited in K episodes following $\pi(i)$. Therefore, with probability $1-\frac{1}{K}$, $\pi(i+1) \neq \pi(i)$, which implies that $m_{i+1} \geqslant m_i + 1$. By choosing $\{v_{id+1}, v_{id+1}, \ldots, v_{id+d}\}$ recursively following the way in Lemma 6 for $0 \leqslant i \leqslant c-1$, we have that with probability $1-\frac{c}{K}$, $m_{i+1} \geqslant m_i + 1$ for $1 \leqslant i \leqslant c$, where $m_c \geqslant c-1$ follows. Then the conclusion follows by the equation below.

$$\mathbb{P}\left[M \leqslant c-2,\ (cd+1, s^{(0)}) \text{ is visited}\right] = \mathbb{P}\left[m_c \leqslant c-2\right] \leqslant \frac{c}{M}.$$

□

C Efficient Implementation of the Proposed Algorithm

In this section, we analyze the computational cost of Algorithm 1. In particular, we first introduce the algorithm `PolicySearch` to show that it can help find the desired exploration policy efficiently.

C.1 The Algorithm

`Policy Search` is presented in Algorithm 4. The algorithms takes two reward functions u, u' and a confidence region $\mathcal{P}$ as input, and output a policy π and $p \in \mathcal{P}$ such that $W^\pi(u', P)$ is large enough compared to $\max_{\pi' \in \Pi(u,\mathcal{P}),} W^{\pi'}(u', P)$.

In the algorithm, we first compute $a := \max_\pi U^\pi(u + \mathbf{1}_z, \mathcal{P})$ and $b := \max_\pi L^\pi(u, \mathcal{P})$. Then we set the target reward as $u + \mathbf{1}_z + \eta u'$ for different η and learn the corresponding optimal policy and transition model $\{\pi^\eta, P^\eta\}$. In intuition, the larger η is, the larger $W^{\pi^\eta}(u, P^\eta)$ is. In this way, we aim to find the maximal η such that π^η is not eliminated, i.e., $\pi^\eta \in \Pi(u, \mathcal{P})$. To find such η, we play the naive dichotomy method as presented in Algorithm 4.

When $u = r$, we assume that $a - b \geqslant \frac{1}{K^3}$ without loss of generality. Note that when $a - b \leqslant \frac{1}{K^3}$, any policy π in $\Pi(r, \mathcal{P})$ is $\frac{1}{K^3}$ optimal and we can follow π in the rest episodes.

In Algorithm 4, we invoke extended value iteration (EVI, see Algorithm 5) as a sub-routine. Algorithm 5 targets compute $(\pi, p) \leftarrow \arg\max_{\pi, p \in \mathcal{P}} W^\pi(u, p)$ for some reward function u and confidence region $\mathcal{P}$. In finite-horizon MDP, this step could be implemented by back induction. So it suffices to solve $\arg\max_{a, p \in \mathcal{P}} p_{h,s,a} V_{h+1}$ where V_{h+1} is the value function computed by back induction. Note that in this paper, the confidence region could be described by at most $O(S^2AK)$ linear constraints, which enables us to find an approximate solution in polynomial time. Besides, given u and $\mathcal{P}$, $\max \pi U^\pi(u, \mathcal{P})$ and $\max_\pi L^\pi(u, \mathcal{P})$ could be computed in a similar way, for which we present Algorithm 6. As a conclusion, Algorithm 4 is computationally efficient.

C.2 Theoretical Results and Proofs for Algorithm 4

Lemma 8. *Let u, u' be two reward functions and $\mathcal{P}$ be a set of transition models. Assume $\mathcal{P} = \otimes_{h,s,a} \mathcal{P}_{h,s,a}$ is* tight *w.r.t. a transition model P. Then by Algorithm 4 we can find π such that*

$$W^\pi(u', P) \geqslant \frac{1}{18} \max_{\pi' \in \Pi(u,\mathcal{P}),} W^{\pi'}(u', P) - \frac{2}{9}\epsilon$$

Algorithm 4 Policy Search

```
Input: reward $u$, $u'$, confidence region $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$;
Initialization: threshold $\epsilon = \frac{1}{(SAHK)^{1}0}$, $b \leftarrow \max_\pi L^\pi(u, \mathcal{P})$, $a \leftarrow \max_\pi U^\pi(u + \mathbf{1}_z, \mathcal{P})$;
$\eta_0 \leftarrow (a - b)/2$;
for $i = 0, 1, 2, \ldots,$ do
    $\{\pi^{(i)}, P^{(i)}\} \leftarrow \texttt{EVI}(u + \mathbf{1}_z + \eta_i u', \mathcal{P})$;
    if $\frac{1}{\epsilon} \leqslant \eta_i < \frac{2}{\epsilon}$; then
        Return: $\pi^{(i)}$;
    else if $W^{\pi^{(i)}}(u, P^{(i)}) \leqslant b$ then
        $\xi = \frac{b - W^{\pi^{(i)}}(u, P^{(i)})}{W^{\pi^{(i-1)}}(u, P^{(i-1)}) - W^{\pi^{(i)}}(u, P^{(i)})}$;
        $(\check{\pi}, \check{P}) \leftarrow \texttt{Sum}(\{\xi, \pi^{(i-1)}, P^{(i-1)}\}, \{1 - \xi, \pi^{(i)}, P^{(i)}\})$
        Return: $\check{\pi}$ ;
    else
        $\eta_{i+1} = 2\eta_i$;
    end if
end for
```

Algorithm 5 Extended Value Iteration (EVI)

```
Input: reward function $u$, confidence region $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$
Initialize: $Q_h(s, a) \leftarrow 0, V_h(s) \leftarrow 0, \forall(h, s, a) \in [H + 1] \times \mathcal{S} \times \mathcal{A}$
for $h = H, H - 1, \ldots, 1$ do
    $Q_h(s, a) \leftarrow \max_{q \in \mathcal{P}_{h,s,a}} (u(s, a) + qV_{h+1}), \forall(s, a) \in \mathcal{S} \times \mathcal{A}$;
    $p_{h,s,a} \leftarrow \arg\max_{q \in \mathcal{P}_{h,s,a}} (u(s, a) + qV_{h+1})$;
    $V_h(s) \leftarrow \max_a Q_h(s, a), \forall s \in \mathcal{S}$;
    $\pi_h(a|s) \leftarrow \mathbb{I}[a = \arg\max_{a'} Q_h(s, a')], \forall(s, a)$;
end for
Return: $\{\pi, p\}$.
```

Algorithm 6 Upper&Lower Confidence Bound

```
Input: reward function $u$, confidence region $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$;
Initialize: $\overline{Q}_h(s, a), \overline{V}_h(s), \underline{Q}_h(s, a), \underline{V}_h(s) \leftarrow 0, \forall(h, s, a) \in [H + 1] \times \mathcal{S} \times \mathcal{A}$;
for $h = H, H - 1, \ldots, 1$ do
    $\overline{Q}_h(s, a) \leftarrow \max_{q \in \mathcal{P}_{h,s,a}}(u(s, a) + q\overline{V}_{h+1}), \forall(s, a) \in \mathcal{S} \times \mathcal{A}$;
    $\overline{V}_h(s) \leftarrow \max_a \overline{Q}_h(s, a), \forall s \in \mathcal{S}$;
    $\underline{Q}_h(s, a) \leftarrow \min_{q \in \mathcal{P}_{h,s,a}}(u(s, a) + q\underline{V}_{h+1}), \forall(s, a) \in \mathcal{S} \times \mathcal{A}$;
    $\underline{V}_h(s) \leftarrow \max_a \underline{Q}_h(s, a), \forall s \in \mathcal{S}$;
end for
Return: $\max_\pi U^\pi(u, \mathcal{P}) := \overline{V}_1(s_1), \max_\pi L^\pi(u, \mathcal{P}) := \underline{V}_1(s_1)$;
```

in time $O(S^4AHM^3\log(SAHK)\log(SAHK/(a-b)))$, where $a = \max_\pi U^\pi(u + \boldsymbol{1}_z, \mathcal{P})$ and $b = \max_\pi L^\pi(u, \mathcal{P})$.

Proof. Let $\tilde{u} = u + \mathbf{1}_z$. For any $\eta \geqslant 0$, we define (π^η, p^η) be the policy-transition pair such that

$$(\pi^\eta, P^\eta) = \arg\max_{\pi, p \in \mathcal{P}} W^\pi(\tilde{u} + \eta u', p).$$

By Lemma 10, with Algorithm 6, we can compute a and b within time $\tilde{O}\left(S^4AHM^3\log(SAHK)\right)$. In the same way, with Algorithm 5 we can find (π^η, p^η) within time $\tilde{O}\left(S^4AHM^3\log(SAHK)\right)$ for any $\eta > 0$. Note that in Algorithm 4, the value of i is at most $\log(1/(\eta_0\epsilon)) = O(\log(\frac{1}{\epsilon(a-b)})) = O(\log(SAHK))$. As a result, the computational cost is at most $O(S^4AHM^3\log(SAHK)\log(SAHK/(a-b)))$.

We continue with an useful property of (π^η, P^η).

Lemma 9. *Let $0 < \eta < \eta'$ be fixed. Let (π^η, p^η), $(\pi^{\eta'}, P^{\eta'})$ be such that*

$$(\pi^\eta, P^\eta) = \arg\max_{\pi, p \in \mathcal{P}} W^\pi(\tilde{u} + \eta u', p)$$
$$(\pi^{\eta'}, P^{\eta'}) = \arg\max_{\pi, p \in \mathcal{P}} W^\pi(\tilde{u} + \eta' u', p).$$

Then we have that

$$W^{\pi^\eta}(\tilde{u}, P^\eta) \geqslant W^{\pi^{\eta'}}(\tilde{u}, P^{\eta'}).$$

Proof. Let $x_1 = W^{\pi^\eta}(\tilde{u}, P^\eta)$, $x_2 = W^{\pi^{\eta'}}(\tilde{u}, P^{\eta'})$, $y_1 = W^{\pi^\eta}(u', P^\eta)$ and $y_2 = W^{\pi^{\eta'}}(u', P^{\eta'})$. It suffices to show that $x_1 \geqslant x_2$. By the optimality of (π^η, P^η) and $(\pi^{\eta'}, P^{\eta'})$, we have that

$$x_1 + \eta y_1 \geqslant x_2 + \eta y_2;$$
$$x_2 + \eta' y_2 \geqslant x_1 + \eta' y_1.$$

If $x_1 < x_2$, then we have that $y_1 > y_2$. It then follows that $x_2 + \eta' y_2 = x_2 + \eta y_2 + (\eta' - \eta)y_2 < x_1 + \eta y_1 + (\eta' - \eta)y_1 = x_1 + \eta' y_1$, which leads to contradiction. □

In Algorithm 4, there are two breaking conditions.

Case 1 Recall that $\{\pi^{(i)}, P^{(i)}\} = \arg\max_{\pi, p \in \mathcal{P}} W^\pi(u + \mathbf{1}_z + \eta_i \mu', p) = \arg\max_{\pi, p \in \mathcal{P}} W^\pi(\tilde{u} + \eta_i \mu', p)$ In the first case, we end with obtaining some i satisfying that

$$W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) \leqslant b.$$

Because $W^{\pi^{(0)}}(\tilde{u}, P^{(0)}) \geqslant a - \eta_0 > b$, it holds that $\eta_i > \eta_0$ for any $i \geqslant 1$. By Lemma 9 and the stopping condition, we have that $W^{\pi^{(i-1)}}(\tilde{u}, P^{(i-1)}) \geqslant b$. By Lemma 2, we can find a policy $\check{\pi}$ and $\check{P} \in \mathcal{P}$ such that

$$W^{\check{\pi}}(v, \check{P}) = \xi W^{\pi^{(i)}}(v, P^{(i)}) + (1 - \xi) W^{\pi^{(i-1)}}(v, P^{(i-1)}) \tag{12}$$

for any reward function v.

Noting that $\xi = \frac{b - W^{\pi^{(i)}}(\tilde{u}, P^{(i)})}{W^{\pi^{(i-1)}}(\tilde{u}, P^{(i-1)}) - W^{\pi^{(i)}}(\tilde{u}, P^{(i)})}$, we have that $U^{\check{\pi}}(u, \mathcal{P}) \geqslant W^{\check{\pi}}(\tilde{u}, \check{P}) = \xi W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) + (1 - \xi) W^{\pi^{(i-1)}}(\tilde{u}, P^{(i-1)}) = b$, which implies that $\check{\pi} \in \Pi(u, \mathcal{P})$.

Note that $W^\pi(v, p)$ is linear in v for fixed π and p. For any policy $\pi \in \Pi(r, \mathcal{P})$ and $p' \in \mathcal{P}$, we have that

$$W^\pi(\tilde{u}, p') + \eta_i W^\pi(u', p') \leqslant W^{\pi^{(i)}}(\tilde{u}, p^{(i)}) + \eta_i W^{\pi^{(i)}}(u', P^{(i)}), \tag{13}$$
$$W^\pi(\tilde{u}, p') + \eta_{i-1} W^\pi(u', p') \leqslant W^{\pi^{(i-1)}}(\tilde{u}, P^{(i-1)}) + \eta_{i-1} W^{\pi^{(i-1)}}(u', P^{(i-1)}). \tag{14}$$

It then follows that

$$\begin{aligned}
&W^{\pi}(\tilde{u}, p') + \eta_{i-1} W^{\pi}(u', p') \\
&\leqslant \xi \left(W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) + \eta_i W^{\pi^{(i)}}(u', P^{(i)}) \right) + (1-\xi) \left(W^{\pi^{(i-1)}}(\tilde{u}, P^{(i-1)}) + \eta_{i-1} W^{\pi^{(i-1)}}(u', P^{(i-1)}) \right) \\
&\leqslant b + \eta_i W^{\check{\pi}}(u', \check{P}). \qquad (15)
\end{aligned}$$

For any $\pi \in \Pi(u, \mathcal{P})$, there exists $p' \in \Pi(u, \mathcal{P})$ such that $W^{\pi}(\tilde{u}, p') \geqslant b$. By (15) and noting that $\eta_i = 2\eta_{i-1}$, we have

$$W^{\pi}(u', p') \leqslant \frac{\eta_i}{\eta_{i-1}} W^{\check{\pi}}(u', \check{P}) \leqslant 2W^{\check{\pi}}(u', \check{P}). \qquad (16)$$

On the other hand, by Lemma 17, for any π it holds that

$$W^{\pi}(u', p) \leqslant 3W^{\pi}(u', p') \leqslant 9W^{\pi}(u', p), \qquad (17)$$

for any $p' \in \bar{\mathcal{P}}$, which implies that

$$W^{\check{\pi}}(u', p) \geqslant \frac{1}{6} \max_{\pi \in \Pi(u, \mathcal{P})} W^{\pi}(u', p).$$

Case 2 In the second case, we end with some i such that $\frac{1}{\epsilon} \leqslant \eta_i < \frac{2}{\epsilon}$.

In this case, because $W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) \geqslant b$, we have that $\pi^{(i)} \in \Pi(u, \mathcal{P})$. For any $\pi \in \Pi(u, \mathcal{P})$ such that

$$W^{\pi}(u', p) \geqslant 18W^{\pi^{(i)}}(u', p), \qquad (18)$$

by the *tightness* of $\mathcal{P}$ (w.r.t. p) it holds that

$$\eta_i W^{\pi}(u', p') \geqslant \frac{\eta_i}{3} W^{\pi}(u', p) \geqslant 6\eta_i W^{\pi^{(i)}}(u', p) \geqslant 2\eta_i W^{\pi^{(i)}}(u', P^{(i)}) \qquad (19)$$

for any $p' \in \mathcal{P}$. On the other hand, by optimality of $(\pi^{(i)}, P^{(i)})$, we have that

$$\eta_i W^{\pi}(u', p') \leqslant W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) + \eta_i W^{\pi^{(i)}}(u', P^{(i)}). \qquad (20)$$

Combine (19) with (20), we have that

$$\eta_i W^{\pi^{(i)}}(u', P^{(i)}) \leqslant W^{\pi^{(i)}}(\tilde{u}, P^{(i)}) \leqslant 2. \qquad (21)$$

Combining (20) with (21), for any $p' \in \mathcal{P}$, using the optimality of $(\pi^{(i)}, P^{(i)})$ and (21), we have that

$$\eta_i W^{\pi}(u', p') \leqslant W^{\pi^{(i)}}(u, P^{(i)}) + \eta_i W^{\pi^{(i)}}(u', P^{(i)}) \leqslant 4. \qquad (22)$$

It then follows $W^{\pi}(u', p) \leqslant 4\epsilon$. Therefore, for any $\pi \in \Pi(u, \mathcal{P})$, it holds either $W^{\pi}(u', p) \leqslant 18W^{\pi^{(i)}}(u', p)$ or $W^{\pi}(u', p) \leqslant 4\epsilon$. We then have that

$$W^{\pi^{(i)}}(u', p) \geqslant \frac{1}{18} \max_{\pi \in \Pi(u, \mathcal{P})} W^{\pi}(u', p) - \frac{2}{9}\epsilon. \qquad (23)$$

The proof is completed.

□

Lemma 10. *The computational cost of Algorithm 5 and Algorithm 6 is bounded by* $O(S^3 AHM^3 \log(SAKH))$.

Proof. To implement the two algorithm, we need to solve SAH linear optimization problem, which has the form $\max_{q \in \mathcal{P}_{h,s,a}}(r + qv)$ or $\min_{q \in \mathcal{P}_{h,s,a}}(r + qv)$. Note that $\mathcal{P}_{h,s,a}$ has the form $\{p \in \Delta^{\mathcal{S}} : a_i^{\top}(p - p') \leqslant b_i, i \geqslant 1\}$, and the number of linear constraints is increased for at most $O(S)$ in each batch. As a result, the total number of linear constraints in $\mathcal{P}_{h,s,a}$ is bounded by $O(SM)$. By the results in Cohen et al. [2021], the time cost to solve the linear program problem above is bounded by $O(S^3 M^3 \log(SAHK))$. Therefore, the total computational cost is bounded by $O(S^3 AHM^3 \log(SAKH))$. □

D Proof of Theorem 1

Additional Notations In this section, we use $N_h^m(s,a,s')$ to denote the visit count of (s,a,h,s') after the m-th batch. We also define $N_h^m(s,a) = \max\{\sum_{s'} N_h^m(s,a,s'),1\}$. We use $\{\check{N}_h^m(s,a,s')\}$ to denote the counts of the m-th batch. Similarly we define $\check{N}_h^m(s,a) = \max\{\sum_{s'} \check{N}_h^m(s,a,s'),1\}$. Let W^* be the *known* set after the first two stages. Let $\hat{P}^m_{h,s,a,s'} = \frac{N_h^m(s,a,s')}{N_h^m(s,a)}$ be the empirical transition model for $1 \leqslant m \leqslant 2H+M$. For $2H+1 \leqslant m \leqslant 2H+M$, define $\{\check{P}^m_{h,s,a}\}$ be the clipped transition model, i.e., $\{\check{P}^m_{h,s,a}\}_{h,s,a} = \texttt{clip}\left(\left\{\left[\frac{\check{N}_h^m(s,a,s')}{\check{N}_h^m(s,a)}\right]_{s'\in\mathcal{S}}\right\}_{h,s,a}, \mathcal{W}^*\right)$.

Note that the m-batch in Algorithm 3 indicates the $2H+m$-th batch in the main algorithm. To align the indices, with a slight abuse of notations we use $\mathcal{P}^m$ and v^m to denote respectively the value of $\mathcal{P}^{m-2H}$ and v^{m-2H} in Algorithm 3 for $m \geqslant 2H$.

Table 1: Explanation of the notations

$W^\pi(u,p)$	the general value function: $W^\pi(u,p) = \mathbb{E}_{p,\pi,s_1\sim\mu_1}[\sum_{h=1}^H u_h(s_h,a_h)]$
$U^\pi(u,\mathcal{P})$	the upper confidence bound w.r.t. policy π, reward u and confidence region $\mathcal{P}$;
$L^\pi(u,\mathcal{P})$	the lower confidence bound w.r.t. policy π, reward u and confidence region $\mathcal{P}$;
$N_h^m(s,a,s')$	the visit count of (s,a,h,s') after the m-th batch
$N_h^m(s,a)$	$N_h^m(s,a) = \max\{\sum_{s'} N_h^m(s,a,s'),1\}$;
$\check{N}_h^m(s,a,s')$	the count of $\left(h,s,a,s'\right)$ in the m-th batch;
$\check{N}_h^m(s,a)$	$\check{N}_h^m(s,a) = \max\left\{\sum_{s'}\check{N}_h^m(s,a,s'),1\right\}$
W^*	the *known* set after the first two stages
$\hat{P}^m_{h,s,a,s'}$	$\hat{P}^m_{h,s,a,s'} = \frac{N_h^m(s,a,s')}{N_h^m(s,a)}$, the empirical transition probability;
$\check{P}^m_{h,s,a}$	$\{\check{P}^m_{h,s,a}\}_{h,s,a} = \texttt{clip}\left(\left\{\left[\frac{\check{N}_h^m(s,a,s')}{\check{N}_h^m(s,a)}\right]_{s'\in\mathcal{S}}\right\}_{h,s,a}, \mathcal{W}^*\right)$;
$\bar{P}$	$\bar{P} = \texttt{clip}\left(P,W^*\right)$, the clipped true transition model;
$\mathcal{P}^m$	the confidence region after the m-th batch;
$\{v_h^m(s)\}$	the extended optimal value function after the m-th batch;
$V^*\left(\bar{V}^*\right)$	the optimal value function for the (clipped) true transition model;
$\alpha(n,n')$	$\alpha(n,n') = \sqrt{\frac{4n'\iota}{n^2}} + \frac{5\iota}{n}$;
$\alpha^*(n,p,v)$	$\alpha^*(n,p,v) = 5\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} + \frac{3\iota}{n}$;

The good event For $1 \leqslant m \leqslant 2H+M$, define $\mathcal{G}^m_{h,s,a,s'}$ be the event where it holds

$$\left|\hat{P}^m_{h,s,a,s'} - P_{h,s,a,s'}\right| \leqslant \beta^m_{h,s,a,s'} := \min\left\{\sqrt{\frac{2P_{h,s,a,s'}\iota}{N_h^m(s,a)}} + \frac{\iota}{3\cdot N_h^m(s,a)}, \sqrt{\frac{4\check{P}^m_{h,s,a,s'}\iota}{N_h^m(s,a)}} + \frac{5\iota}{N_h^m(s,a)}\right\}. \tag{24}$$

By Lemma 3 and Bernstein inequality, we have that $\mathbb{P}[\mathcal{G}^m_{h,s,a,s'}] \geqslant 1-2\delta$.

For $1 \leqslant m \leqslant 2H$, we set $\check{\mathcal{G}}^m_{h,s,a}$ to be the whole event. For $2H+1 \leqslant m \leqslant M$, we define $\check{\mathcal{G}}^m_{h,s,a}$ be the event where it holds

(25)

$$\left|(\check{P}_{h,s,a} - P)v^{m-1}\right| \leqslant \lambda^m_{h,s,a} := \min\left\{5\sqrt{\frac{\mathbb{V}(\check{P}^m_{h,s,a}, v^{m-1})\iota}{\check{N}_h^m(s,a)}}\right\} \tag{26}$$

$$\left|(\check{P}_{h,s,a} - P)\bar{V}^*\right| \leqslant \lambda^{m,*}_{h,s,a} := \min\left\{5\sqrt{\frac{\mathbb{V}(\check{P}^m_{h,s,a}, \bar{V}^*)\iota}{\check{N}_h^m(s,a)}}\right\}. \tag{27}$$

Noting that $\check{P}_{h,s,a}$ is independent with both $\bar{V}^*$ and v^{m-1}, by Bernstein's inequality, we have that $\mathbb{P}[\check{\mathcal{G}}^m_{h,s,a,s'}] \geqslant 1 - 4\delta$

The good event $\mathcal{G}$ is defined as $\mathcal{G} = \bigcap_{h,s,a,s'} \bigcap_{m=1}^{M} \left(\mathcal{G}^m_{h,s,a,s'} \cap \check{\mathcal{G}}^m_{h,s,a} \right)$ Then $\mathbb{P}[\mathcal{G}] \geqslant 1 - 6S^2AHM\delta$. Throughout the analysis, we always assume $\mathcal{G}$ holds.

Lemma 11. *Conditioned on $\mathcal{G}$, we have $\bar{P} \in \mathcal{P}^m$ for $2H \leqslant m \leqslant 2H + M$.*

Noting that the batch complexity is bounded by $2H + M = O(H + \log_2 \log_2(K))$, it suffices to prove the regret bound. We start with counting the regret in the first two stages. The regret in the first batch is bounded by $O(H^2 k_1)$ trivially. As for the second batch, we have that

Lemma 12. *Conditioned on $\mathcal{G}$, with probability $1 - 4SAH\delta$ the regret bound in the second batch is bounded by $O\left(\frac{k_2\sqrt{S^4A^3H^8\iota}}{\sqrt{k_1}} + \frac{k_2 S^3A^3H^4\iota}{k_1} \right)$.*

To count the regret in the third stage, we first show that the difference between the clipped model and the original model could be ignored.

Lemma 13. *Conditioned on $\mathcal{G}$, with probability $1 - 4S^2AH^2\delta$, for any optimal policy π^*, it holds that $\Pr_{\pi^*}[\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^*] \leqslant O\left(\frac{S^3A^2H^3\iota}{k_2} \right)$*

Based on Lemma 13, we further have that

Lemma 14. *Recall that $\bar{V}^*$ be the optimal value function with respect to the transition model $\bar{P}$ and reward function r. It then holds that $\bar{V}_1^*(s_1) \leqslant V_1^*(s_1) \leqslant \bar{V}_1^*(s_1) + O\left(\frac{S^3A^2H^4\iota}{k_2} \right)$.*

Proof. The left side is obvious since the reward at z is always 0. On the other hand, letting π^* be an optimal policy and E be the event where $\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^*$. Then we have that

$$
\begin{aligned}
V_1^{\pi^*}(s_1) &\leqslant \mathbb{E}_{\pi^*}\left[\left(\sum_{h=1}^{H} r_h(s_h, a_h) \right) \mathbb{I}[E] \right] + H\Pr_{\pi^*}[E] \\
&\leqslant \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \mathbb{I}[\forall h' < h, (h', s_{h'}, a_{h'}, s_{h'+1}) \in \mathcal{W}^*] \right] + O\left(\frac{S^3A^2H^4\iota}{k_2} \right) \\
&= \bar{V}_1^{\pi^*}(s_1) + O\left(\frac{S^3A^2H^4\iota}{k_2} \right).
\end{aligned}
$$

□

Recall that $\mathrm{gap}^{m+1} := \max_{\pi \in \Pi(r, \mathcal{P}^m)} (U^\pi(\mathcal{P}^m) - L^\pi(\mathcal{P}^m))$. For $m \geqslant 2H + 1$, we have that

Lemma 15. *Conditioned on $\mathcal{G}$, with probability $1 - 4SAHKM\delta$, it holds that*

$$
\begin{aligned}
&\mathrm{gap}^{m+1} \\
&\leqslant O\left(\sqrt{\frac{SAH^3\ln(K)\iota^2}{K_{m-2H}}} + \frac{SAH^2\ln(K)\iota}{K_{m-2H}} + \sqrt{\frac{S^{\frac{11}{2}}A^4H^7\ln(K)\iota^{\frac{5}{2}}}{K_{m-2H}k_1}} + \sqrt{\frac{S^4A^{\frac{5}{2}}H^4\ln(K)\iota^{\frac{3}{2}}}{K_{m-2H}\sqrt{k_1}}} \right).
\end{aligned}
\tag{28}
$$

By Lemma 11, 14 and 15, for any $2H \leqslant m \leqslant 2H + K$ and any $\pi \in \Pi(\mathcal{P}^m)$, we have that

$$
V_1^\pi(s_1) \geqslant L^\pi(\mathcal{P}^m) \geqslant U^\pi(\mathcal{P}^m) - \mathrm{gap}^{m+1} \geqslant \bar{V}_1^*(s_1) - \mathrm{gap}^{m+1} - O\left(\frac{S^3A^2H^4\iota}{k_2} \right).
$$

Recall that $k_1 = 144\sqrt{SAK\iota/H}$, $k_2 = 288S^3A^2H^4\sqrt{K\iota}$ and $K_m = \left\lceil K^{1-\frac{1}{2^m}} \right\rceil$ for $1 \leqslant m \leqslant M$. It then holds that $\frac{K_{m-2H+1}}{\sqrt{K_{m-2H}}} = \sqrt{K}$ for any $2H + 1 \leqslant m \leqslant 2H + K$. Noting that the regret in the

$m+1$-th batch is bounded by $K_{m+1-2H} \cdot \mathrm{gap}^{m+1}$, and the regret in the $2H+1$-th batch is bounded by $K_1 = O(\sqrt{K})$, the total regret is bounded by

$$\mathrm{Regret}(K) = M \cdot O\left(\sqrt{SAH^3 K \ln(K) \iota^2} + S^{\frac{15}{4}} A^{\frac{9}{8}} H^{\frac{17}{8}} \iota^{\frac{5}{8}} K^{\frac{3}{8}} + S^{\frac{19}{4}} A^{\frac{13}{4}} H^{\frac{33}{4}} \ln(K) \iota K^{\frac{1}{4}} + S^{\frac{11}{2}} A^{\frac{9}{2}} H^{\frac{17}{2}} \iota\right).$$

By replacing δ by $\frac{\delta}{20S^2AHK}$, we get the desired regret bound.

Below we analyze the computational cost of Algorithm 1. By Lemma 2 the computational costs of `Sum` is $O(nS^3A^2H^2)$, where n is the number of inputs for `Sum`.

Below we analyze the computational cost of `PolicySearch`. By Lemma 8, for input $(u, u', \mathcal{P})$, the computational cost of `PolicySearch` is bounded by $O(S^4AHM^3 \log(SAHK) \log(SAHK/(a-b)))$ with $a = \max_\pi U^\pi(u + \mathbf{1}_z, \mathcal{P})$ and $b = \max_\pi L^\pi(u, \mathcal{P})$.

In the first stage, we invoke `PolicySearch` with $u = 0$, which implies $b = 0$ and $W^\pi(u, p) = 0$ for any π and $p \in \mathcal{P}$. Then the condition in Line 7 Algorithm 4 is satisfied and the loop would break. Therefore, by Lemma 10, the computational cost of `PolicySearch` in the first stage is bounded by $O(S^4AHM^3 \log(SAKH))$.

In the second and the third stage, we invoke `PolicySearch` with $u = r$. In this case, if $a - b \leqslant 1/K$, then we can learn an $1/K$-optimal policy by solving $\pi' = \arg\max_\pi L^\pi(r, \mathcal{P})$. Then we can simply run this policy in the left episodes. Without loss of generality, we then assume that $a - b > 1/K$, which implies the time cost of `PolicySearch` is bounded by $O(S^4AHM^3 \log^2(SAKH))$.

Now we count the number of callings to `Sum` and `PolicySearch`. In the first and second stage, `Sum` is called for $2H$ times with $n = SAH$ inputs, and `PolicySearch` is called for $2H$ times. In the third stage, `Sum` is called for M times with $n = K^3$ inputs, and `PolicySearch` is called for K^3M times. So the total time cost due to `Sum` and `PolicySearch` is bounded by $\tilde{O}(S^4AHK^3 + S^3A^2H^2K^3)$. On the other hand, to compute $\{v_h^m(s)\}_{h \in [H], s \in \mathcal{S}}$ in Line 15 Algorithm 4, we need to invoke `EVI` (see Algorithm 5) for M times, which needs additional $O(S^4AHM^4 \log(SAHK))$ time by Lemma 10. Finally, to observe the samples and compute the confidence region, we need $O(S^2AHK)$ time.

Putting all together, the computational cost of Algorithm 1 is bounded by $\tilde{O}(S^4AHK^3 + S^3A^2H^2K^3)$. The proof is completed.

D.1 Proof of Lemma 11

Lemma 11 (restated) *Conditioned on $\mathcal{G}$, we have $\bar{P} \in \mathcal{P}^m$ for $2H \leqslant m \leqslant 2H + M$.*

Proof. with a slight abuse of notation, we use v^m to denote the value of v^{m-2H} in Algorithm 3.

Recall the definition of $\mathcal{P}^m$. It suffices to show that $\bar{P} \in \mathtt{CR}^*(\mathcal{D}^m, \mathcal{D}^m, W^*, \{v_h^{\overline{m-1}}(s)\}_{(h,s)})$ for each $m \geqslant 2H$.

Note that after the m-th batch $\hat{p}_{h,s,a,s'} = \hat{P}^m_{h,s,a,s'}$ and $\check{p}_{h,s,a} = \check{P}^m_{h,s,a}$. By the definition of $\mathcal{G}$, and recalling the definition of $\beta^m_{h,s,a,s'}$ and $\lambda^m_{h,s,a}$ in (24) and (26), we have that

$$\left|\bar{P}_{h,s,a,s'} - \hat{P}^m_{h,s,a,s'}\right| \leqslant \beta^m_{h,s,a,s'} \leqslant \alpha(N_h^m(s,a), N_h^m(s,a,s'))$$
$$\left|(\bar{P}_{h,s,a} - \check{P}^m_{h,s,a})v^{m-1}\right| \leqslant \lambda^m_{h,s,a} \leqslant \alpha^*(\check{N}_h^m(s,a), \check{P}^m_{h,s,a}, v^{m-1}).$$

The proof is completed. □

D.2 Proof of Lemma 12

Lemma 12 (restated) *Conditioned on $\mathcal{G}$, with probability $1 - 4SAH\delta$ the regret bound in the second stage is bounded by $O\left(\frac{k_2\sqrt{S^4A^3H^8\iota}}{\sqrt{k_1}} + \frac{k_2S^3A^3H^4\iota}{k_1}\right)$.*

Proof. Let $\mathcal{D}^1$ and $\mathcal{D}^2$ be respectively the dataset after the first and second stage. Let $\{\bar{N}_h^1(s,a,s')\}$ and $\{\bar{N}_h^2(s,a,s')\}$ be the corresponding counts. Let $\bar{\mathcal{W}}^1$ and $\bar{\mathcal{W}}^2$ be the corresponding *known* set.

Note that $\mathcal{W}^* = \bar{\mathcal{W}}^2$. By Lemma 16, with probability $1 - 8S^2AH^2\delta$, it holds that

$$\max_{\pi} \mathbb{P}_{\pi}\left[\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \bar{\mathcal{W}}^1\right] \leqslant \frac{36C_1S^2A^2H^3\iota}{k_1}$$

$$\bar{N}_h^1(s,a) \geqslant \frac{ck}{27SA}\max_{\pi} W^{\pi}(\mathbf{1}_{h,s,a}, P) - 4\iota - \frac{36C_1SAH^3\iota}{27}. \tag{29}$$

For any policy π in $\Pi(\mathtt{CR}(\mathcal{D}^1))$, using policy difference lemma we have that

$$\begin{aligned}
&U^{\pi}(\mathtt{CR}(\mathcal{D}^1)) - L^{\pi}(\mathtt{CR}(\mathcal{D}^1))\\
&= U^{\pi}(\mathtt{CR}(\mathcal{D}^1)) - W^{\pi}(r, \mathtt{clip}(P, \bar{\mathcal{W}}^1)) + W^{\pi}(r, \mathtt{clip}(P, \bar{\mathcal{W}}^1)) - L^{\pi}(\mathtt{CR}(\mathcal{D}^1)) && (30)\\
&\leqslant \max_{\pi} \mathbb{P}_{\pi}\left[\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \bar{\mathcal{W}}^1\right] + O\left(\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a}, \mathtt{clip}(P, \bar{\mathcal{W}}^1))\sqrt{\frac{S\iota}{\bar{N}_h^1(s,a)}} \cdot H\right)\\
&\leqslant \frac{36C_1S^2A^2H^3\iota}{k_1} + O\left(\sum_{h,s,a}\left(\frac{SA(\bar{N}_h^1(s,a) + SAH^3\iota)}{k}\right)\sqrt{\frac{SH^2\iota}{\bar{N}_h^1(s,a)}}\right) && (31)\\
&\leqslant \frac{36C_1S^2A^2H^3\iota}{k_1} + O\left(\sqrt{\frac{S^4A^3H^8\iota}{k_1}} + \frac{S^3A^3H^4\iota}{k_1}\right),
\end{aligned}$$

where the third line is by (29) and the last line is by Cauchy's inequality and the fact that $\bar{N}_h^1(s,a) \geqslant 1$. Conditioned on $\mathcal{G}$, we have that $= \mathtt{clip}(P, \bar{\mathcal{W}}^1) \in \mathtt{CR}(\mathcal{D}^1)$. As a result, we have that $\max_{\pi} U^{\pi}(\mathtt{CR}(\mathcal{D}^1)) \geqslant V_1^*(s_1) - \frac{36C_1S^3A^2H^4\iota}{k_1}$. To conclude, the regret in the second stage is bounded by $O\left(\frac{k_2\sqrt{S^4A^3H^8\iota}}{\sqrt{k_1}} + \frac{k_2S^3A^3H^4\iota}{k_1}\right)$. □

D.3 Proof of Lemma 13

Lemma 13 (restated) *Conditioned on $\mathcal{G}$, with probability $1 - 4S^2AH^2\delta$, for any optimal policy π^*, it holds that* $\Pr_{\pi^*}[\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^*] \leqslant O\left(\frac{S^3A^2H^3\iota}{k_2}\right)$.

Proof. By Lemma 16, with probability $1 - 4S^2AH^2\delta$, it holds that

$$\max_{\pi \in \Pi^*} \Pr_{\pi}\left[\exists h \in [H], (h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^*\right] \leqslant \frac{36C_1S^2A^2H^3\iota}{k_2}.$$

The proof is completed. □

D.4 Proof of Lemma 15

Lemma 15 (restated) *Conditioned on $\mathcal{G}$, with probability $1 - 4SAHKM\delta$, it holds that*

$$\mathrm{gap}^m \leqslant O\left(\sqrt{\frac{SAH^3\ln(K)\iota^2}{K_{m-2H}}} + \frac{SAH^2\ln(K)\iota}{K_{m-2H}} + \sqrt{\frac{S^{\frac{11}{2}}A^4H^7\ln(K)\iota^{\frac{5}{2}}}{K_{m-2H}k_1}} + \sqrt{\frac{S^4A^{\frac{5}{2}}H^4\ln(K)\iota^{\frac{3}{2}}}{K_{m-2H}\sqrt{k_1}}}\right)$$

for $2H + 1 \leqslant m \leqslant M$.

Proof. Let $m \in [2H+1, M]$ be fixed. Conditioned on $\mathcal{G}$, we have that for any $p \in \mathcal{P}^{m-1}$, for any $(h, s, a, s') \in \mathcal{W}^*$ it holds that

$$\begin{aligned}
\left|\hat{P}_{h,s,a,s'}^{m-1} - \bar{P}_{h,s,a,s'}\right| &\leqslant \sqrt{\frac{4\hat{P}_{h,s,a,s'}^{m-1}\iota}{N_h^{m-1}(s,a)}} + \frac{\iota}{3N_h^{m-1}(s,a)}\\
&= \frac{1}{N_h^{m-1}(s,a)} \cdot \left(\sqrt{4N_h^{m-1}(s,a,s')\iota} + 1/3\right)\\
&\leqslant 3\hat{P}_{h,s,a,s'}^{m-1} \cdot \sqrt{\frac{\iota}{N_h^{m-1}(s,a,s')}}\\
&\leqslant \frac{1}{3H}\hat{P}_{h,s,a,s'}^{m-1}.
\end{aligned}$$

On the other hand, noting that for any $p \in \mathcal{P}^{m-1}$ and $(h,s,a,s') \in \mathcal{W}^*$, with similar computation it holds that

$$\begin{aligned}\left|p_{h,s,a,s'} - \bar{P}_{h,s,a,s'}\right| &\leqslant \left|p_{h,s,a,s'} - \hat{P}^{m-1}_{h,s,a,s'}\right| + \left|\hat{p}_{h,s,a,s'} - \bar{P}^{h'}_{h,s,a,s'}\right| \\ &\leqslant \frac{1}{3H}\bar{P}_{h,s,a,s'} + \frac{1}{3H}\hat{P}^{m-1}_{h,s,a,s'} \\ &\leqslant \left(\frac{2}{3H} + \frac{1}{9H^2}\right)\bar{P}_{h,s,a,s'}\end{aligned}$$

Therefore $\mathcal{P}^{m-1}$ is *tight* with respect to $\bar{P}$. Let $p^{m-1} \in \mathcal{P}^{m-1}$ be the value of p in Line 26 Algorithm 3. Let $r^{i,m-1}$ be the value of r^i defined in Line 29 Algorithm 3. Let $\{\tilde{\pi}^{i,m-1}\}$ be the value of $\tilde{\pi}(i)$ in Line 30 Algorithm 3.

As a result, by Lemma 8, Lemma 2 and Lemma 17

$$\begin{aligned}W^{\tilde{\pi}^{i,m-1}}(r^{i,m-1},\bar{P}) &\geqslant \frac{c}{9}\max_{\pi\in\Pi(\mathcal{P}^{m-1})} W^{\pi}(r^{i,m-1},\bar{P}) \\ W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P}) &\geqslant \frac{1}{9K^3}\sum_{i=1}^{K^3} W^{\tilde{\pi}^{i,m-1}}(\mathbf{1}_{h,s,a},\bar{P}), \forall (h,s,a). \end{aligned} \tag{32}$$

Consequently, for any $\pi \in \Pi(\mathcal{P}^{m-1})$ and (h,s,a), it holds that

$$\begin{aligned}W^{\pi}(r^{K^3+1,m-1},\bar{P}) \leqslant& \frac{81}{cK^3}\sum_{i=1}^{K^3} W^{\tilde{\pi}^{i,m-1}}(r^{i,m-1},\bar{P}) \\ &= \frac{81}{cK^3}\sum_{i=1}^{K^3}\sum_{h,s,a} W^{\tilde{\pi}^{i,m-1}}(\mathbf{1}_{h,s,a},\bar{P})\cdot\min\left\{\frac{1}{\sum_{j=1}^{i-1}W^{\tilde{\pi}^{j,m-1}}(\mathbf{1}_{h,s,a},p^{m-1})},1\right\} \\ &= \frac{81}{cK^3}\sum_{h,s,a}\sum_{i=1}^{K^3}\sum_{h,s,a} W^{\tilde{\pi}^{i,m-1}}(\mathbf{1}_{h,s,a},\bar{P})\cdot\min\left\{\frac{1}{\sum_{j=1}^{i-1}W^{\tilde{\pi}^{j,m-1}}(\mathbf{1}_{h,s,a},p^{m-1})},1\right\} \\ &\leqslant \frac{243}{cK^3}\sum_{h,s,a}\sum_{i=1}^{K^3}\sum_{h,s,a} W^{\tilde{\pi}^{i,m-1}}(\mathbf{1}_{h,s,a},\bar{P})\cdot\min\left\{\frac{1}{\sum_{j=1}^{i-1}W^{\tilde{\pi}^{j,m-1}}(\mathbf{1}_{h,s,a},\bar{P})},1\right\} \\ &\leqslant \frac{243SAH\ln(K)}{cK^3} \end{aligned}\tag{33}$$

where the second line is by the *tightness* (w.r.t. $\bar{P}$) of $\mathcal{P}^{m-1}$, and the last line is by the fact that for any non-negative $\{x_i\}_{i=1}^n$

$$\begin{aligned}\sum_{i=1}^n x_i\cdot\min\left\{\frac{1}{\sum_{j=1}^{i-1}x_j},1\right\} &\leqslant 2 + 2\sum_{i=1}^n\left(\ln\left(\sum_{j=1}^i x_i\right) - \ln\left(\sum_{j=1}^{i-1}x_j\right)\right)\mathbb{I}\left[\left(\sum_{j=1}^{i-1}x_j\right)\geqslant 1\right] \\ &\leqslant 2 + 2\ln\left(\sum_{i=1}^n x_i\right).\end{aligned}$$

By definition of $r^{K^3,m-1}$, we have that for any (h,s,a)

$$\begin{aligned}r_h^{K^3+1,m-1}(s,a) &= \min\left\{\frac{1}{\sum_{j=1}^{K^3}W^{\tilde{\pi}^{j,m-1}}(\mathbf{1}_{h,s,a},p^{m-1})},1\right\} \\ &\geqslant \frac{1}{3}\min\left\{\frac{1}{\sum_{j=1}^{K^3}W^{\tilde{\pi}^{j,m-1}}(\mathbf{1}_{h,s,a},\bar{P})},1\right\} = \frac{1}{3}\min\left\{\frac{1}{K^3W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P})},1\right\}.\end{aligned}\tag{34}$$

By (33) and (34), for any $\pi \in \Pi(\mathcal{P}^{m-1})$ it holds that

$$\sum_{h,s,a}W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\cdot\min\left\{\frac{1}{K^3W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P})},1\right\} \leqslant 3\sum_{h,s,a}W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})r_h^{K^3+1,m-1}(s,a) \leqslant \frac{729SAH\ln(K)}{cK^3}. \tag{35}$$

Note that π^m is executed for K_{m-2H} rounds. By Lemma 4, with probability $1 - 4SAH\delta$, it holds that

$$\check{N}_h^m(s,a) \geqslant \frac{1}{3}K_{m-2H}W^{\pi^m}(1_{h,s,a},\bar{P}) - \iota. \tag{36}$$

Fix $\pi \in \Pi(r, \mathcal{P}^{m-1})$. Let $\{f_h(\cdot)\}_{h=1}^S$ be the value function under π and $\bar{P}$. For any $P' \in \mathcal{P}^m$, by policy difference lemma, we have that

$$\begin{aligned}
&\left|W^\pi(r,P') - W^\pi(r,\bar{P})\right| \\
&= \left|\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},P')\cdot(P'_{h,s,a} - \bar{P}_{h,s,a})f_{h+1}\right| \\
&\leqslant \underbrace{\left|\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},P')(P'_{h,s,a} - \bar{P}_{h,s,a})v_{h+1}^{m-1}\right|}_{\mathbf{Term.1}} + \underbrace{\left|\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},P')(P'_{h,s,a} - \bar{P}_{h,s,a})(f_{h+1} - v_{h+1}^{m-1})\right|}_{\mathbf{Term.2}}.
\end{aligned} \tag{37}$$

By the definition of $\mathcal{P}^m$ and $\mathcal{G}$, we have that

$$\begin{aligned}
\mathbf{Term.1} &= \left|\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},P')(P'_{h,s,a} - \check{P}^m_{h,s,a} + \check{P}^m_{h,s,a} - P_{h,s,a})v_{h+1}^{m-1}\right| \\
&\leqslant \sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},P')\cdot\left(5\sqrt{\frac{\mathbb{V}(\check{P}^m_{h,s,a},v_{h+1}^{m-1})\iota}{\check{N}_h^m(s,a)}} + 5\sqrt{\frac{\mathbb{V}(\bar{P}_{h,s,a},v_{h+1}^{m-1})\iota}{\check{N}_h^m(s,a)}} + \frac{8\iota}{\check{N}_h^m(s,a)}\right) \\
&\leqslant O\left(\sqrt{\sum_{h,s,a}\frac{W^\pi(\mathbf{1}_{h,s,a},\bar{P})\iota}{\check{N}_h^m(s,a)}}\cdot\sqrt{\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\cdot\left(\mathbb{V}(\check{P}^m_{h,s,a},v_{h+1}^{m-1}) + \mathbb{V}(\bar{P}_{h,s,a},v_{h+1}^{m-1})\right)}\right) \\
&\qquad + O\left(\sum_{h,s,a}\frac{W^\pi(\mathbf{1}_{h,s,a},\bar{P})\iota}{\check{N}_h^m(s,a)}\right)
\end{aligned} \tag{38}$$

Define $T_1 = \sum_{h,s,a}\frac{W^\pi(\mathbf{1}_{h,s,a},\bar{P})\iota}{\check{N}_h^m(s,a)}$, $T_2 = \sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\cdot\mathbb{V}(\bar{P}_{h,s,a},v_{h+1}^{m-1})$ and $T_2 = \sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\cdot\mathbb{V}(\check{P}^m_{h,s,a},v_{h+1}^{m-1})$.

Bound of T_1 By (35) and (36), we have that

$$\begin{aligned}
T_1 &\leqslant 3\sum_{h,s,a}\frac{W^\pi(\mathbf{1}_{h,s,a},\bar{P})}{\max\{K_{m-2H}W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P}) - 3\iota, 1\}} \\
&= \frac{3K^3}{K_{m-2H}}\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\cdot\min\left\{\frac{1}{K^3W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P}) - 3K^3\iota/K_{m-2H}}, \frac{K_{m-2H}}{K^3}\right\} \\
&\leqslant \frac{3K^3}{K_{m-2H}}\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\cdot\left(\min\left\{\frac{2}{K^3W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P})}, 1\right\}\cdot\mathbb{I}\left[K_{m-2H}W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P}) \geqslant 6\iota\right]\right) \\
&\qquad + 3\sum_{h,s,a} W^\pi(\mathbf{1}_{h,s,a},\bar{P})\mathbb{I}\left[K_{m-2H}W^{\pi^m}(\mathbf{1}_{h,s,a},\bar{P}) < 6\iota\right] \\
&\leqslant \frac{3K^3}{K_{m-2H}}\cdot\frac{729SAH\ln(K)}{cK^3} + \frac{18SAH\iota}{K_{m-2H}} \\
&= O\left(\frac{SAH\ln(K)\iota}{K_{m-2H}}\right).
\end{aligned} \tag{39}$$

Bound of T_2

$$
\begin{align}
T_2 &= \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \mathbb{V}(\bar{P}_{h,s,a}, v_{h+1}^{m-1}) \nonumber\\
&= \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \left(\bar{P}_{h,s,a}(v_{h+1}^{m-1})^2 - (\bar{P}_{h,s,a}v_{h+1}^{m-1})^2\right) \nonumber\\
&\leqslant \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \left((v_{h}^{m-1}(s))^2 - (\bar{P}_{h,s,a}v_{h+1}^{m-1})^2\right) + H^2 \nonumber\\
&\leqslant H \sum_{h=1}^{H} \mathbb{E}_{\pi,\bar{P}}\left[|v_h^{m-1}(s_h) - \bar{P}_{h,s_h,a_h}v_{h+1}^{m-1}|\right] + H^2 \nonumber\\
&= H \sum_{h=1}^{H} \mathbb{E}_{\pi,\bar{P}}\left[v_h^{m-1}(s_h) - \bar{P}_{h,s_h,a_h}v_{h+1}^{m-1}\right] + H^2 \tag{40}\\
&\leqslant H \sum_{h=1}^{H} \mathbb{E}_{\pi,\bar{P}}[r_h(s_h,a_h)] + 2H^2 \nonumber\\
&\leqslant 4H^2. \tag{41}
\end{align}
$$

Here (40) is by the fact that v_h^{m-1} is the optimal value function with respect to $\mathcal{P}^{m-1}$ and $\bar{P} \in \mathcal{P}^{m-1}$.

Bound of T_3 By Lemma 3, with probability $1 - 4S^2AH\delta$, it holds that

$$
\left|\check{P}^{m}_{h,s,a,s'} - \bar{P}_{h,s,a,s'}\right| \leqslant 4\sqrt{\frac{\bar{P}_{h,s,a,s'}\iota}{\check{N}_h^m(s,a)}} + \frac{3\iota}{\check{N}_h^m(s,a)} \leqslant 2\bar{P}_{h,s,a,s'} + \frac{5\iota}{\check{N}_h^m(s,a)}.
$$

As a result, we have that

$$
\begin{align}
T_3 &= \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \mathbb{V}(\check{P}^{m}_{h,s,a}, v_{h+1}^{m-1}) \nonumber\\
&\leqslant \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \sum_{s'} \check{P}^{m}_{h,s,a,s'}\left(v_{h+1}^{m-1}(s') - \bar{P}_{h,s,a,s'}v_{h+1}^{m-1}\right)^2 \nonumber\\
&\leqslant \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \sum_{s'} \bar{P}^{m}_{h,s,a,s'}\left(v_{h+1}^{m-1}(s') - \bar{P}_{h,s,a,s'}v_{h+1}^{m-1}\right)^2 + \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P}) \cdot \frac{5H^2\iota}{\check{N}_h^m(s,a)} \nonumber\\
&= 4T_2 + 5H^2\iota T_1 \nonumber\\
&\leqslant O\left(H^2 + \frac{SAH^2\ln(K)\iota^2}{K_{2m-H}}\right). \tag{42}
\end{align}
$$

By (39), (41) and (42), $\mathbf{Term.1}$ is bounded by

$$
\mathbf{Term.1} \leqslant O\left(\sqrt{\frac{SAH^3\ln(K)\iota^2}{K_{m-2H}}} + \frac{SAH^2\ln(K)\iota}{K_{m-2H}}\right). \tag{43}
$$

To bound $\mathbf{Term.2}$, by definition of $\mathcal{P}^m$ and $\mathcal{G}$, we have

$$\begin{aligned}\mathbf{Term.2} &= \left|\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},P')(P'_{h,s,a}-\bar{P}_{h,s,a})(f_{h+1}-v_{h+1}^{m-1})\right| \\ &\leqslant \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},P')\sum_{s'}\left(10\sqrt{\frac{\bar{P}_{h,s,a,s'}\iota}{N_h^m(s,a)}}+\frac{6\iota}{N_h^m(s,a)}\right)\cdot|f_{h+1}(s')-v_{h+1}^{m-1}(s')-l| \\ &\leqslant O\left(\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\sum_{s'}\sqrt{\frac{\bar{P}_{h,s,a,s'}\iota}{N_h^m(s,a)}}|f_{h+1}(s')-v_{h+1}^{m-1}(s')-l_h(s,a)|\right) \\ &\qquad\qquad\qquad\qquad + O\left(\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\frac{SH\iota}{N_h^m(s,a)}\right),\end{aligned} \tag{44}$$

where $l_h(s,a) = \bar{P}_{h,s,a}(f_{h+1}-v_{h+1}^m)$. By (39), the second term in (44) is bounded by $O\left(\frac{SAH\ln(K)\iota}{K_{m-2H}}\right)$. To bound the the first term in (44), by Cauchy's inequality, we have that

$$\begin{aligned} & O\left(\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\sqrt{\frac{S\mathbb{V}(\bar{P}_{h,s,a},f_{h+1}-v_{h+1}^{m-1})\iota}{N_h^m(s,a)}}\right) \\ &\leqslant O\left(\sqrt{\frac{SW^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\iota}{N_h^m(s,a)}}\cdot\sqrt{\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\mathbb{V}(\bar{P}_{h,s,a},f_{h+1}-v_{h+1}^{m-1})}\right) \\ &\leqslant O\left(\sqrt{\frac{S^2AH\ln(K)\iota^2}{K_{m-2H}}}\cdot\sqrt{\sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\mathbb{V}(\bar{P}_{h,s,a},f_{h+1}-v_{h+1}^{m-1})}\right),\end{aligned}$$

where the last line is by (39). Continuing the computation:

$$\begin{aligned} & \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\mathbb{V}(\bar{P}_{h,s,a},f_{h+1}-v_{h+1}^{m-1}) \\ &= \sum_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})\left(\bar{P}_{h,s,a}(f_{h+1}-v_{h+1^{m-1}})^2-(\bar{P}_{h,s,a}f_{h+1}-\bar{P}_{h,s,a}v_{h+1}^{m-1})^2\right) \\ &\leqslant \mathbb{E}_{\pi,\bar{P}}\left[\sum_{h=1}^{H}\left((f_{h+1}(s_{h+1})-v_{h+1}^{m-1}(s_{h+1}))^2-(\sum_a \pi_h(a|s)\bar{P}_{h,s,a}(f_{h+1}-v_{h+1}))^2\right)\right] \qquad (45) \\ &\leqslant (v_1^{m-1}(s_1)-f_1(s_1))^2+H\mathbb{E}_{\pi,\bar{P}}\left[\sum_{h=1}^{H}\left|f_h(s_h)-v_h^{m-1}(s_h)-\sum_a \pi_h(a|s_h)\bar{P}_{h,s_h,a}(f_{h+1}-v_{h+1}^{m-1})\right|\right] \\ &\leqslant (v_1^{m-1}(s_1)-f_1(s_1))^2+H\mathbb{E}_{\pi,\bar{P}}\left[\sum_{h=1}^{H} v_h^{m-1}(s_h)-\sum_a \pi_h(a|s_h)\left(r_h(s_h,a)+\bar{P}_{h,s_h,a}v_{h+1}^{m-1})\right)\right] \\ &\qquad (46) \\ &\leqslant (v_1^{m-1}(s_1)-f_1(s_1))^2+H(v_1^{m-1}(s_1)-f_1(s_1)) \\ &\leqslant 2H(v_1^{m-1}(s_1)-f_1(s_1)).\end{aligned}$$

Here (45) holds by the fact that $\mathrm{Var}(X)\geqslant\mathbb{E}_Y[\mathrm{Var}(X|Y)]$ for any random variables X and Y (recalling that $\mathrm{Var}(X)$ denotes the variance of X), and (46) is by the fact that $v_h^{m-1}(s_h)\geqslant\sum_a \pi_h(a|s_h)(r_h(s_h,a)+\bar{P}_{h,s,a}v_{h+1}^{m-1})$ and $f_h(s_h)=\sum_a \pi_h(a|s_h)(r_h(s_h,a)+\bar{P}_{h,s,a}f_{h+1})$ for any $1\leqslant h\leqslant H$.

Because $\pi\in\Pi(r,\mathcal{P}^{m-1})$, we learn that $v_1^{m-1}-f_1(s_1)\leqslant\mathrm{gap}^m$. By Lemma 16, we have that for any $m\geqslant H+1$, $N_h^{m-1}(s,a)\geqslant\frac{ck_1}{27SA}\max_{\pi}W^{\pi}(\mathbf{1}_{h,s,a},\bar{P})-4\iota-\frac{36C_1SAH^3\iota}{27}$. With similar analysis, and noting that $\|p'_{h,s,a} - p''_{h,s,a}\|_1 \leqslant O(\sqrt{S\iota/N_h^{m-1}(s,a)})$ for any $p', p'' \in \mathcal{P}^{m-1}$, we have

$$\begin{aligned}\mathrm{gap}^m &\leqslant O\left(\max_{\pi'} \sum_{h,s,a} W^{\pi'}(\mathbf{1}_{h,s,a}, \bar{P})\sqrt{\frac{SH^2\iota}{N_h^{m-1}(s,a)}}\right) \\ &\leqslant O\left(\sqrt{\frac{S^4A^3H^4\iota}{k_1}} + \frac{S^{\frac{7}{2}}A^3H^5\iota^{\frac{3}{2}}}{k_1}\right). \end{aligned} \tag{47}$$

As a result, we have that

$$\mathbf{Term.2} \leqslant O\left(\sqrt{\frac{S^{\frac{11}{2}}A^4H^7\ln(K)\iota^{\frac{5}{2}}}{K_{m-2H}k_1}} + \sqrt{\frac{S^4A^{\frac{5}{2}}H^4\ln(K)\iota^{\frac{3}{2}}}{K_{m-2H}\sqrt{k_1}}} + \frac{S^2AH\ln(K)\iota}{K_{m-2H}}\right). \tag{48}$$

Putting all together, for any $\pi \in \Pi(r, \mathcal{P}^{m-1})$ and any $P' \in \mathcal{P}^m$, we have

$$\begin{aligned}&|W^\pi(r, P') - W^\pi(r, \bar{P})| \\ &\leqslant O\left(\sqrt{\frac{SAH^3\ln(K)\iota^2}{K_{m-2H}}} + \frac{SAH^2\ln(K)\iota}{K_{m-2H}} + \sqrt{\frac{S^{\frac{11}{2}}A^4H^7\ln(K)\iota^{\frac{5}{2}}}{K_{m-2H}k_1}} + \sqrt{\frac{S^4A^{\frac{5}{2}}H^4\ln(K)\iota^{\frac{3}{2}}}{K_{m-2H}\sqrt{k_1}}}\right).\end{aligned}$$

By definition, there exists P', P'' such that $U^\pi(\mathcal{P}^m) = W^\pi(r, P')$ and $L^\pi(\mathcal{P}^m) = W^\pi(r, P'')$. Therefore,

$$\begin{aligned}&|U^\pi(\mathcal{P}^m) - L^\pi(\mathcal{P}^m)| \\ &\leqslant O\left(\sqrt{\frac{SAH^3\ln(K)\iota^2}{K_{m-2H}}} + \frac{SAH^2\ln(K)\iota}{K_{m-2H}} + \sqrt{\frac{S^{\frac{11}{2}}A^4H^7\ln(K)\iota^{\frac{5}{2}}}{K_{m-2H}k_1}} + \sqrt{\frac{S^4A^{\frac{5}{2}}H^4\ln(K)\iota^{\frac{3}{2}}}{K_{m-2H}\sqrt{k_1}}}\right).\end{aligned}$$

Taking maximization over $\pi \in \Pi(r, \mathcal{P}^{m-1})$ we finish the proof. □

D.4.1 Statement and Proof of Lemma 16

Lemma 16. *Given a dataset $\mathcal{D}$ and $k \geqslant 0$, let $\mathcal{D}'$ be the output by running Algorithm 2 with input $(r, \mathcal{D}, k)$. Let $\{N_h(s,a,s')\}(\{N'_h(s,a,s')\})$ be the counts with respect to $\mathcal{D}(\mathcal{D}')$. Let $\mathcal{W} = \{(h,s,a,s')|N_h(s,a,s') \geqslant C_1H^2\iota\}$ and $\mathcal{W}' = \{(h,s,a,s')|N'_h(s,a,s') \geqslant C_1H^2\iota\}$. Let $\bar{p} = \texttt{clip}(P, \mathcal{W})$. With probability $1 - 4S^2AH^2\delta$, it holds that*

$$\max_{\pi\in\Pi^*} \mathrm{Pr}_\pi\left[\exists h' \in [h], (h', s_{h'}, a_{h'}, s_{h'+1}) \notin \mathcal{W}'\right] \leqslant \frac{36C_1S^2A^2H^3\iota}{k}, \tag{49}$$

where Π^* is the set of optimal policies. Moreover, if $\mathcal{D} = \varnothing$ and $u = 0$, with probability $1 - 4S^2AH^2\delta$ it holds that

$$N'_{h,s,a} \geqslant \frac{ck}{27SA}\max_\pi W^\pi(\boldsymbol{1}_{h,s,a}, P) - 4\iota - \frac{36C_1SAH^3\iota}{27} \tag{50}$$

for any $1 \leqslant h \leqslant H$.

Proof. For $h' = 1, 2, ..., H$, we denote $\mathcal{D}^{h'}$ as the value of $\mathcal{D}$ after the h'-th batch in Algorithm 2. Similarly, we define $\{N_h^{h'}(s,a,s')\}$, $\{N_h^{h'}(s,a)\}$ and $\{\hat{p}_{h,s,a}^{h'}\}$ be respectively the value of $\{N_h(s,a,s')\}$, $\{N_h(s,a)\}$ and $\{\hat{p}_{h,s,a}\}$ after the h'-th batch. Note that $\mathcal{P}^{h'} = \texttt{CR}(\mathcal{D}^{h'})$ is the value of $\mathcal{P}$ after the h'-th batch.

Define $\mathcal{W}^{h'} := \{(h,s,a,s') : N_h^{h'}(s,a,s') \geqslant C_1H^2\iota\}$ and $P^{h'} = \texttt{clip}(P, \mathcal{W}^{h'})$. Let $p^{h'} \in \mathcal{P}^{h'-1}$ be the transition model chosen at line 8 Algorithm 2.

Using Lemma 3 and Lemma 4, with probability $1 - 4S^2AH^2\delta$, for any $(h, s, a, s') \in \mathcal{W}^{h'}$, it holds that

$$\left|P^{h'}_{h,s,a,s'} - \hat{p}^{h'}_{h,s,a,s'}\right| \leqslant \sqrt{\frac{4P^{h'}_{h,s,a,s'}\iota}{N^{h'}_h(s,a)}} + \frac{\iota}{3N^{h'}_h(s,a)}$$
$$\leqslant \frac{1}{3H}P^{h'}_{h,s,a,s'},$$

where in the last inequality, we use Lemma 4 to get that $N^{h'}_h P^{h'}_{h,s,a,s'} \geqslant \frac{1}{3}N^{h'}_h(s,a,s') - \iota \geqslant 64H^2\iota$ with probability $1 - \delta$.

It then holds that $P^{h'} \in \mathcal{P}^{h'}$ for each h'. Moreover, noting that for any $p \in \mathcal{P}^{h'}$ and $(h, s, a, s') \in \mathcal{W}^{h'}$, with similar computation it holds that

$$\left|p_{h,s,a,s'} - P^{h'}_{h,s,a,s'}\right| \leqslant \left|p_{h,s,a,s'} - \hat{p}^{h'}_{h,s,a,s'}\right| + \left|\hat{p}_{h,s,a,s'} - P^{h'}_{h,s,a,s'}\right|$$
$$\leqslant \frac{1}{3H}P^{h'}_{h,s,a,s'} + \frac{1}{3H}\hat{p}^{h'}_{h,s,a,s'}$$
$$\leqslant \left(\frac{2}{3H} + \frac{1}{9H^2}\right)P^{h'}_{h,s,a,s'}$$

As a result, $\mathcal{P}^{h'}$ is *tight* with respect to $P^{h'}$.

Fix $h \in [H]$. Recall that

$$\pi^{h,s,a} = \texttt{Policy Search}(\mathbf{1}_{h,s,a}, \mathcal{P}^{h-1});$$
$$\{\tilde{\pi}^h, p^h\} = \texttt{Sum}\left(\left\{\frac{1}{SA}, \pi_{h,s,a}, p^h\right\}_{h,s,a}\right). \tag{51}$$

Recall that, for the first $h-1$ steps π^h is the policy which is the same as $\tilde{\pi}^h$, and for the left $H-h+1$ steps, π^h is the uniformly random policy.

We first show that the h-th layer is well explored. By the property of `Policy Search` and `Sum` (see Lemma 8 and 2), there exists a constant $c > 0$ such that[12]

$$W^{\pi^{h,s,a}}(\mathbf{1}_{h,s,a}, P^{h-1}) \geqslant c \max_{\pi \in \Pi(r,\mathcal{P}^{h-1})} W^{\pi}(\mathbf{1}_{h,s,a}, P^{h-1})$$
$$W^{\pi^h}(\mathbf{1}_{h,s,a}, p^h) = \frac{1}{SA}W^{\pi_{h,s,a}}(\mathbf{1}_{h,s,a}, p^h), \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Noting that $p^h \in \mathcal{P}^{h-1}$ and $\mathcal{P}^{h-1}$ is *tight* with respect to P^{h-1}, by Lemma 17 we obtain that

$$W^{\pi^h}(\mathbf{1}_{h,s,a}, P^{h-1}) \geqslant \frac{c}{9SA} \max_{\pi \in \Pi(r,\mathcal{P}^{h-1})} W^{\pi}(\mathbf{1}_{h,s,a}, P^{h-1}) \tag{52}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Using Lemma 4, with probability $1 - 4SA\delta$, the count of (h, s, a) in the h-th batch is at least $\frac{ck}{27SA} \cdot \max_{\pi \in \Pi(r,\mathcal{P}^{h-1})} W^{\pi}(\mathbf{1}_{h,s,a}, P^{h-1}) - \iota$. As a result, we have that

$$N^h_h(s,a) \geqslant \frac{ck}{27SA} \cdot \max_{\pi \in \Pi(r,\mathcal{P}^{h-1})} W^{\pi}(\mathbf{1}_{h,s,a}, P^{h-1}) - 4\iota \tag{53}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

In the meantime, if $(h, s, a, s') \notin \mathcal{W}^h$, we have that $N^h_h(s,a,s') \leqslant C_1H^2\iota$. Using Lemma 4, with probability $1 - \delta$, we have that

$$kW^{\pi^h}(\mathbf{1}_{h,s,a}, P^{h-1})P_{h,s,a,s'} \leqslant 3C_1H^2\iota + 4\iota. \tag{54}$$

[12] We omit ϵ for convenience. By setting $\epsilon = 1/(SAHK)^{10}$, it is easy to verify the error only leads to a lower order term.

Combining (52) and (54), we have that

$$\begin{aligned}\max_{\pi \in \Pi(r, \mathcal{P}^{h-1})} \mathbb{P}_{\pi, P^{h-1}}\left[(s_h, a_h, s_{h+1}) = (s, a, s')\right] &= \max_{\pi \in \Pi(r, \mathcal{P}^{h-1})} W^{\pi}(\mathbf{1}_{h,s,a}, P^{h-1}) P_{h,s,a,s'} \\ &\leqslant \frac{9SA}{c} W^{\pi^{h}}(\mathbf{1}_{h'',s,a}, P^{h-1}) P_{h'',s,a,s'} \\ &\leqslant \frac{9SA}{c} \cdot \frac{3C_1 H^2 \iota + 4\iota}{k} \\ &\leqslant \frac{36 C_1 SAH^2 \iota}{k}. \end{aligned} \tag{55}$$

With an union bound over all $(h, s, a, s') \notin \mathcal{W}^h$, we have that

$$\max_{\pi \in \Pi(r, \mathcal{P}^{h-1})} \mathbb{P}_{\pi, P^{h-1}}\left[(h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^h\right] \leqslant \frac{36 C_1 S^2 A^2 H^2 \iota}{k}. \tag{56}$$

Note that $\mathcal{W}^h$ is non-decreasing in h. For any $\pi \in \cap_{h=1}^{h'} \Pi(r, \mathcal{P}^h)$, it holds that

$$\begin{aligned}&\mathbb{P}_{\pi, P}\left[\exists h' \leqslant H, (h', s_{h'}, a_{h'}, s_{h'+1}) \notin \mathcal{W}^H\right] \\ &= \mathbb{P}_{\pi, P^H}\left[\exists h' \leqslant H, (h', s_{h'}, a_{h'}, s_{h'+1}) \notin \mathcal{W}^H\right] \\ &= \sum_{h'=1}^{H} \mathbb{P}_{\pi, P^H}\left[(h', s_{h'}, a_{h'}, s_{h'+1}) \in \mathcal{W}^H, \forall 1 \leqslant h' < h, (h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^H\right] \\ &\leqslant \sum_{h=1}^{H} \max_{\pi \in \Pi(r, \mathcal{P}^{h-1})} \mathbb{P}_{\pi, P^{h-1}}\left[(h, s_h, a_h, s_{h+1}) \notin \mathcal{W}^h\right] \\ &\leqslant \frac{36 C_1 S^2 A^2 H^3 \iota}{k}. \end{aligned} \tag{57}$$

Recall that $\Pi(r, \mathcal{P}) := \{\pi | U^{\pi}(r + \mathbf{1}_z, \mathcal{P}) \geqslant \max_{\pi} L^{\pi}(r, \mathcal{P})\}$. Because $P^h \in \mathcal{P}^h$ for any h, for any optimal policy π^* and any policy π', we have that $U^{\pi^*}(r + \mathbf{1}_z, \mathcal{P}) \geqslant V_1^*(s_1) \geqslant W^{\pi'}(r, P^h) \geqslant L^{\pi'}(r, \mathcal{P})$. Therefore, $\pi^* \in \Pi(\mathcal{P}^h)$ for any $1 \leqslant h \leqslant H$. By (57), (49) is proven.

In the case $u = 0$, we have that $\Pi(u, \mathcal{P}^h) = \overline{\Pi}$ for $1 \leqslant h \leqslant H$, where $\overline{\Pi}$ is the set of all possible policies. By (53), we have that

$$\begin{aligned}&N_h^h(s,a) \qquad (58)\\ &\geqslant \frac{ck}{27SA} \max_{\pi} W^{\pi}(\mathbf{1}_{h,s,a}, P^h) - 4\iota \qquad (59)\\ &\geqslant \frac{ck}{27SA} \max_{\pi} W^{\pi}(\mathbf{1}_{h,s,a}, P^{)} - 4\iota - \frac{ck}{27SA} \max_{\pi} \mathbb{P}_{\pi, P}\left[\exists h' \in [H], (h', s_{h'}, a_{h'}, s_{h'+1}) \notin \mathcal{W}^H\right] \\ &\geqslant \frac{ck}{27SA} \max_{\pi} W^{\pi}(\mathbf{1}_{h,s,a}, P) - 4\iota - \frac{36 C_1 SAH^3 \iota}{27}. \end{aligned}$$

The proof is completed by noting that $N_h'(s,a) \geqslant N_h^h(s,a)$.

□

D.5 Statement and Proof of Lemma 17

Lemma 17. *Suppose $\mathcal{P}$ is* tight *with respect to p. Then we have that*

$$3W^{\pi}(\boldsymbol{1}_{h,s,a}, p) \geqslant W^{\pi}(\boldsymbol{1}_{h,s,a}, p') \geqslant \frac{1}{3} W^{\pi}(\boldsymbol{1}_{h,s,a}, p) \tag{60}$$

for any $p' \in \mathcal{P}$, policy π and (h, s, a).

Proof. For each trajectory $L = (s_1, a_1, ..., s_H, a_H, s_{H+1})$ such that $s_h \neq z$ for $1 \leqslant h \leqslant H+1$, we have that

$$\mathbb{P}_{\pi,p}[L] = \pi_{h=1}^{H} \pi_h(a_h | s_h) p_{h, s_h, a_h, s_{h+1}} \geqslant e^{-\frac{H}{H}} \mathbb{P}_{\pi,p}[L] = \pi_{h=1}^{H} \pi_h(a_h | s_h) p'_{h, s_h, a_h, s_{h+1}} \geqslant \frac{1}{3} \mathbb{P}_{\pi, p'}[L].$$

So the left side of (60) is proven. By reversing p and p' the right side follows.

□

E Other Missing Proofs

E.1 Proof of Lemma 1

Lemma 1 (restated) *Let $d > 0$ be an integer. Let $\mathcal{X} \subset (\Delta^d)^m$. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X}$, such that*

$$\max_{x=\{x_i\}_{i=1}^{dm} \in \mathcal{X}} \sum_{i=1}^{dm} \frac{x_i}{y_i} = md,$$

where $y = \{y_i\}_{i=1}^{dm} = \mathbb{E}_{x\sim\mathcal{D}}[x]$. Moreover, if $\mathcal{X}$ has a boundary set $\partial\mathcal{X}$ with finite cardinality, we can find $\mathcal{D}$ in $\mathrm{poly}(|\partial\mathcal{X}|)$ time.

Proof. Note that $\mathcal{X}$ is always bounded. Without loss of generality, we assume $\mathcal{X}$ is a discrete set with $\mathcal{X} = \{x^1, x^2, ..., x^L\}$ where $x^i = \{x^i_n\}_{n=1}^{dm}$ For $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_L\} \in \Delta^L$, we define $E(\lambda)$ by

$$E(\lambda) := \Pi_{i=1}^{dm} \left(\sum_{j=1}^{L} \lambda_j x_i^j \right).$$

Then $E(\lambda)$ is bounded and Δ^L is compact. Consider to maximize $\ln(E(\lambda))$ over $\lambda \in \Delta^L$. It's not hard to verify that $\ln(E(\lambda))$ is concave in λ, so it is efficient to maximize it by gradient ascent algorithms. Let λ^* be the optimal solution. By the KKT condition, we have that for any j', j'' such that $\lambda^*_{j'}, \lambda^*_{j''} \in (0,1)$, it holds that

$$w := \sum_{i=1}^{dm} \frac{x_i^{j'}}{\sum_{j=1} \lambda_j x_i^j} = \sum_{i=1}^{dm} \frac{x_i^{j''}}{\sum_{j=1} \lambda_j x_i^j}.$$

Therefore, if for any $\lambda_j^* \neq 1$ for any j, we have that

$$w = \sum_{j=1}^{L} \lambda_j w = \sum_{i=1}^{dm} \frac{\sum_{j=1} \lambda_j x_i^j}{\sum_{j=1} \lambda_j x_i^j} = dm.$$

Then λ^* is the desired solution. Otherwise, suppose $\lambda_1^* = 1$. Then we have that

$$dm = \sum_{i=1}^{dm} \frac{x_i^1}{x_i^1} \geqslant \sum_{i=1}^{dm} \frac{x_i^{j'}}{x_i^1}$$

for any $j' \geqslant 2$. Then λ^* is also the desired solution. The proof is completed. □

E.2 Proof of Lemma 2

Lemma 2 (Restatement) *Let $\mathcal{P} = \otimes_{(h,s,a)} \mathcal{P}_{h,s,a}$ be a set of transition models such that $\mathcal{P}_{h,s,a} \subset \Delta^S$ is convex for any (h,s,a). Let $\{(\pi^i, P^i)\}_{i=1}^n$ be a sequence of policy-transition pairs such that $P^i \in \mathcal{C}$. For any $\{\lambda_i\}_{i=1}^n$ such that $\lambda_i \geqslant 0$ for $i \geqslant 1$ and $\sum_i \lambda_i = 1$, there exists a policy π and $P \in \mathcal{P}$, satisfying that*

$$W^{\pi}(\boldsymbol{1}_{h,s,a}, P) = \sum_i \lambda_i W^{\pi^i}(\boldsymbol{1}_{h,s,a}, P^i) \tag{61}$$

for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Furthermore, the time complexity to find $\{\pi, P\}$ could be bounded by $O(nS^3A^2H^2)$.

Proof. By induction on n, it suffices to prove for the case $n = 2$. Our target is to find (π, p) such that

$$W^{\pi}(\mathbf{1}_{h,s,a}, P) = \lambda_1 W^{\pi^1}(\mathbf{1}_{h,s,a}, P^1) + (1-\lambda_1) W^{\pi^2}(\mathbf{1}_{h,s,a}, P^2) \tag{62}$$

holds for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. We will prove this by induction on h. For the case $h = 1$, since the initial distribution is fixed, we finish by letting

$$\pi_1(a|s) = \lambda_1 \pi_1^1(a|s) + (1 - \lambda_1)\pi_1^2(a|s) \tag{63}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

Suppose (62) holds for any $1 \leqslant h' \leqslant h$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$. Then $\lambda_{h,s,a} = \frac{\lambda_1 W^{\pi^1}(\mathbf{1}_{h,s,a}, P^1)}{W^{\pi}(\mathbf{1}_{h,s,a}, P)}$ is well-defined. We set

$$P_{h,s,a} = \lambda_{h,s,a} P^1_{h,s,a} + (1 - \lambda_{h,s,a}) P^2_{h,s,a}$$

for any $(s, a)] \in \mathcal{S} \times \mathcal{A}$. By the inductive assumption

$$W^{\pi}(\mathbf{1}_{h,s,a}, P) = \lambda_1 W^{\pi^1}(\mathbf{1}_{h,s,a}, P^1) + (1 - \lambda_1) W^{\pi^2}(\mathbf{1}_{h,s,a}, P^2), \tag{64}$$

we have that for any (s, a)

$$P_{h,s,a} W^{\pi}(\mathbf{1}_{h,s,a}, P) = \lambda_1 W^{\pi^1}(\mathbf{1}_{h,s,a}, P^1) P^1_{h,s,a} + (1 - \lambda_1) W^{\pi^2}(\mathbf{1}_{h,s,a}, P^2) P^2_{h,s,a}.$$

We then have that

$$\begin{aligned}
&\lambda_1 \sum_{s',a'} W^{\pi^1}(\mathbf{1}_{h,s',a'}, P^1) P^1_{h,s',a',s} + (1 - \lambda_1) \sum_{s',a'} W^{\pi^2}(\mathbf{1}_{h,s',a'}, P^2) P^2_{h,s,',a',s} \\
&= \sum_{s',a'} \lambda_{h,s',a'} W^{\pi}(\mathbf{1}_{h,s',a'}, P) P^1_{h,s',a',s} + \sum_{s',a'} (1 - \lambda_{h,s',a'}) W^{\pi}(\mathbf{1}_{h,s',a'}, P) P^2_{h,s',a',s} \qquad (65) \\
&= \sum_{s',a'} W^{\pi}(\mathbf{1}_{h,s',a'}, P) P_{h,s',a',s}, \qquad (66)
\end{aligned}$$

which implies that

$$W^{\pi}(\mathbf{1}_{h+1,s}, P) = \lambda_1 W^{\pi^1}(\mathbf{1}_{h+1,s} P^1) + (1 - \lambda_1) W^{\pi^2}(\mathbf{1}_{h+1,s} P^2) \tag{67}$$

for any $s \in \mathcal{S}$, where the reward function $\mathbf{1}_{h+1,s} = \sum_a \mathbf{1}_{h+1,s,a}$. Let

$$\pi_{h+1}(a|s) = \frac{\lambda_1 W^{\pi^1}(\mathbf{1}_{h+1,s}, P^1)\pi^1_{h+1}(a|s) + (1 - \lambda_1) W^{\pi^2}(\mathbf{1}_{h+1,s}, P^2)\pi^2_{h+1}(a|s)}{W^{\pi}(\mathbf{1}_{h+1,s}, P)}. \tag{68}$$

Then it is easy to verify that

$$W^{\pi}(\mathbf{1}_{h+1,s,a}, P) = W^{\pi}(\mathbf{1}_{h+1,s}, P)\pi_{h+1}(a|s) = \lambda_1 W^{\pi^1}(\mathbf{1}_{h+1,s,a}, P^1) + (1 - \lambda_1) W^{\pi^2}(\mathbf{1}_{h+1,s,a}, P^2).$$

Also note that the process above costs at most $O(S^3 A^2 H^2)$ time, so the total computational cost is bounded by $O(nS^3 A^2 H^2)$. The proof is completed. □