TEL AVIV UNIVERSITY
The Raymond & Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

REGRET MINIMIZATION IN REINFORCEMENT LEARNING

A thesis submitted toward the degree of
Doctor of Philosophy

by

Aviv Rosenberg

August 2022

TEL AVIV UNIVERSITY
The Raymond & Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

REGRET MINIMIZATION IN REINFORCEMENT LEARNING

A thesis submitted toward the degree of
Doctor of Philosophy

by

Aviv Rosenberg

This research was carried out at Tel Aviv University
in the Blavatnik School of Computer Science
The Raymond & Beverly Sackler Faculty of Exact Sciences
under the supervision of Prof. Yishay Mansour

August 2022

Acknowledgments

I would like to thank my partner, parents, family, and friends for their love and support. I would like to attribute a special thanks to my advisor, Professor Yishay Mansour. Yishay is the ultimate advisor and he has made my Ph.D. an unforgettable experience. His incredible knowledge and inexhaustible creativity serve as a constant source of personal inspiration. I enjoyed our joint work and I deeply appreciate your support in my research career (and hopefully your future guidance). I also wish to express my deepest gratitude to all the great students and researchers with whom I had the pleasure to collaborate during my Ph.D. – Gal Chechik, Liyu Chen, Alon Cohen, Gal Dalal, Yonathan Efroni, Assaf Hallak, Tiancheng Jin, Haim Kaplan, Tal Lancewicki, Haipeng Luo, Shie Mannor, and Lior Shani – many thanks to all of you for giving me the opportunity to learn through our interaction.

Abstract

Reinforcement Learning (RL) studies the most basic question in Artificial Intelligence (AI) – *how can an agent learn to make good decisions through interaction with the environment?*

While RL has seen impressive empirical success in various settings, the performance of RL algorithms changes dramatically between domains and they may even fail to learn in certain environments. This could happen for many reasons but in this thesis we focus on the following three fundamental reasons:

1. *Exploration.* Many popular algorithms rely on simple heuristics for exploration, such as ε-greedy. Therefore, they might fail in environments where it is hard to reach certain areas of the state space.

2. *Non-stationary.* The most popular model in RL is the Markov Decision Process (MDP), which is entirely stochastic and does not change over time. However, in many real-world applications, the environment is not stationary and changes even throughout the learning process. Many algorithms fail to adapt to these changes.

3. *Inaccurate Model.* The RL literature mostly studies MDPs with the finite-horizon, discounted return or average-reward performance criteria. However, many scenarios (such as navigation and routing) do not fit into these frameworks. Thus, many algorithms are not able to capture them adequately.

This thesis provides new algorithms and theory for tackling all of these issues. Our algorithms face the challenges of exploration in several environments, and therefore their success is measured by the *regret* – the difference between the cumulative cost of the agent through the learning process and the expected cost of the best policy in hindsight.

The thesis consists of two main lines of research: *adversarial MDPs* and *Stochastic Shortest Path (SSP)*. After studying both models, we also study a new model, *adversarial SSP*, which combines them to construct a much more robust and general model.

Adversarial MDPs aim to tackle non-stationary. As opposed to standard MDPs that are stochastic and do not change over time, in adversarial MDPs the cost function can change arbitrarily (while still assuming a fixed stochastic transition function). This model is much more general than standard MDPs since it allows for the costs to be chosen by an adversary, instead of just being drawn from some unknown distribution. In this work we significantly advance our understanding of adversarial MDPs. We present the first high-probability regret bounds for adversarial MDPs with unknown transitions and full-information feedback, where the agent observes the entire cost function after it has changed. Moreover, we present the first regret bounds for the much more realistic model of adversarial MDPs with unknown transitions and bandit feedback, where the agent observes only the costs that she suffers. Our algorithms are built on entropic regularization methodologies, which are known to be highly effective in practice.

Stochastic shortest path (SSP) is one of the most basic models in reinforcement learning. It includes the discounted return model and the finite-horizon model as special cases. In SSP the goal of the agent is to reach a predefined goal state in minimum expected cost. This setting captures a wide variety of realistic scenarios, such as car navigation, game playing and drone flying; i.e., tasks carried out in episodes that eventually terminate. In this work we present the first near-optimal regret bounds for SSP. Then, we develop an improved algorithm based on a reduction to the finite-horizon setting, and prove that it attains optimal regret (up to logarithmic factors).

Table of Contents

1 Introduction

Reinforcement Learning (RL) is a branch of Machine Learning (ML) which studies sequential decision making under uncertainty. It provides a general framework for many practical problems in Artificial Intelligence (AI). In the basic RL setup, an agent interacts with an uncertain environment in order to perform a task by taking a sequence of actions. More precisely, the agent needs to learn the optimal actions in order to maximize its long term payoff, or equivalently, minimize its long term losses. Reinforcement Learning provides algorithmic tools to optimize the strategy of the agent.

There have been impressive empirical successes driven by Deep Learning (DL) that demonstrate how Reinforcement Learning can solve challenging tasks. These include playing a range of Atari video games [MKS$^+$15], achieving human-level performance in Go [SSS$^+$17] and many more. However, the potential applications of RL go beyond games. It is a natural framework for optimizing recommender systems [LCLS10, CBC$^+$19], but also for optimizing adaptive treatments in health-care [LNSL$^+$12], dialog systems [SLKW02] and instruction schedules in intelligent tutoring systems [MLL$^+$14].

The framework adopted in reinforcement learning is that of Markov Decision Processes (MDP) [Put14]. In an MDP, also called an environment, there is an agent that transitions between states by taking actions, and whose aim is to accumulate as little cost as possible. Namely, the agent interacts with the environment with a sequence of actions that allow the agent to visit different states and suffer costs. The objective is to accumulate as little cost as possible – which can stand for financial loss, user dissatisfaction, energy or time – and the plan or strategy to achieve this goal is called a policy. The “goodness” of being in a particular state / position is represented by the value function, which is the expected sum of future costs until the end of the interactions with the environment.

1.1 Regret Minimization in RL

The major focus of this dissertation is on the exploration problem, a fundamental dilemma in RL that is absent in classical prediction-oriented machine learning. The exploration problem arises whenever an RL agent needs to trade-off between exploiting the current best policy and exploring uncertain policies. Playing an uncertain policy may temporarily hurt the agent's performance, yet this is crucial to find better actions to improve future performance. In other words, the exploration problem refers to the process of consciously taking potentially sub-optimal actions to learn more about the environment, and trade-off a temporary decrease in performance for a potentially lower long-term accumulated cost.

A natural way to measure the agent's performance is to look at the difference between the expected costs accumulated by the agent throughout the learning process and those accumulated by the best policy in hindsight, which we generally denote by $\pi^\star$. This is called the *regret* of the agent.

Throughout the dissertation, the regret will be our primary measure for the performance of an agent; that is, we will seek to design agents capable of minimizing the regret. We focus on two models that generalize standard MDPs: *Adversarial MDPs* and *Stochastic Shortest Path (SSP)*. These models tackle fundamental challenges that many RL algorithms need to face in practice.

Adversarial MDPs aim to tackle non-stationarity, i.e., changes in the environment over time. The most popular model in RL is the Markov Decision Process (MDP), which is entirely stochastic and does not change over time. However, in many real-world applications, the environment is not stationary and changes even throughout the learning process. Many algorithms fail to adapt to these changes. As opposed to standard MDPs that are stochastic and do not change over time, in adversarial MDPs the cost function can change arbitrarily (while still assuming a fixed stochastic transition function). This model is much more general than standard MDPs since it allows for the costs to be chosen by an adversary, instead of just being drawn from some unknown distribution. In this work we significantly advance our understanding of adversarial MDPs.

Stochastic shortest path (SSP) is one of the most basic models in reinforcement learning. It includes the discounted return model and the finite-horizon model as special cases. Yet, the RL literature mostly studies MDPs with the finite-horizon, discounted return or average-reward performance criteria, although many scenarios do not fit into these frameworks. In SSP the goal of the agent is to reach a predefined goal state in minimum expected cost. This setting captures a wide variety of realistic scenarios, such as car navigation, game playing and drone flying; i.e., tasks carried out in episodes that eventually terminate. In this work we present the first near-optimal regret bounds for SSP.

1.2 Our Contributions

1.2.1 Adversarial MDP with Full-Information Feedback

We present the first high-probability regret bounds for adversarial MDPs where the agent does not know the transition function in advance and observes full-information feedback, i.e., the entire cost function is revealed after it changes. We bound the regret by $\widetilde{O}(H^2 S\sqrt{AK})$, where S is the number of states, A is the number of actions, H is the horizon (i.e., episode length), and K is the number of episodes.

This improves upon the only previous regret bound for this setting [NGS12] by a factor of $\sqrt{A}$ (this regret bound only holds in expectation and not with high-probability). Moreover, it is larger than the currently best known lower bound of $\Omega(\sqrt{H^3 SAK})$ [OVR16, JAZBJ18] by a factor of only $\sqrt{HS}$ (and logarithmic factors).

Our algorithm UC-O-REPS is built on entropic regularization methodology, which is known to be highly effective in practice. Our main technical contribution here is a novel method to combine the "optimism in face of uncertainty" principle, which helps us estimate the unknown transition function, with the online mirror descent (OMD) method, which is a popular framework for handling adversarial cost functions in online learning. While a naive combination of these two methods yields a non-convex optimization problem (making the computational complexity of the algorithm not polynomial), our method manages to use a convex optimization problem which can be solved efficiently. Moreover, we present a novel analysis to our methods which splits the regret into two separate terms: one for the error in the transition function estimation, and the other for the regret caused by the unknown sequence of cost functions chosen by the adversary.

1.2.2 Adversarial MDP with Bandit Feedback

We propose the first algorithms for the adversarial MDPs with bandit feedback and an unknown transition function. Our algorithms are based on our UC-O-REPS algorithm, that assumes unknown transition function but full-information feedback. Our first algorithm, "Bounded Bandit UC-O-REPS", assumes that any state is reachable under any policy with probability $\beta > 0$ and achieves a regret bound of $\widetilde{O}(H^2 S\sqrt{AK}/\beta)$. Our second algorithm, "Shifted Bandit UC-O-REPS", removes this assumption and achieves a regret bound of $\widetilde{O}(H^2 S A^{1/4} K^{3/4})$.

Here our technical contributions are as follows. First, we show that the UC-O-REPS algorithm can be extended efficiently to pick only policies that reach all states with probability at least β. Second, we highlight the unique challenge that appears only when both the transition function is unknown and the feedback is bandit (and not full-information): it is not possible to construct an unbiased estimator to the cost function. Third, we provide a novel analysis that bounds the error that our biased estimator introduces. Finally, we propose a novel method to generalize the case that there is a positive lower on β to the general case: through a perturbation of the confidence sets and the estimator.

1.2.3 Near-Optimal Regret for SSP

Our first result in the SSP model is the first near-optimal regret bound. We improve upon the work of [TGV$^+$20], which is the only previous regret minimization algorithm specifically designed for SSP. First, we remove the dependency on $c_{\min}$ and allow for zero costs while maintaining regret of $\widetilde{O}(\sqrt{K})$. Second, we give a much simpler algorithm in which the computation of the optimistic policy has a simple solution. Our main regret term is $\widetilde{O}(B_\star S\sqrt{AK})$, where $B_\star$ is an upper bound on the expected cost of the optimal policy (note that $B_\star \le D$). Moreover, we show that this is almost optimal by giving the first lower bound for SSP. It scales as $\Omega(B_\star\sqrt{SAK})$.

We obtain a major improvement in the regret bound through the use of confidence sets that are based on Bernstein inequality [AOM17], that is highly sensitive to variance, instead of Hoeffding inequality. In both our algorithm and the one of [TGV$^+$20], the regret scales with the square root of the total variance. When using Hoeffding-based confidence sets, similarly to UCRL2 [JOA10], this variance is trivially bounded by $B_\star^2$ at each step, which leads to a regret of $\widetilde{O}(\sqrt{B_\star^2 T})$, where T is the number of time-steps taken by the algorithm. However, the use of Bernstein inequalities enables us to bound the total expected variance in a time interval, of roughly $B_\star/c_{\min}$ time-steps, by an identical magnitude of $O(B_\star^2)$. Therefore, the regret bound for our algorithm improves upon the regret of [TGV$^+$20] by a factor of $\sqrt{B_\star/c_{\min}}$, that is, $\widetilde{O}(B_\star S\sqrt{AK})$ compared to $\widetilde{O}(D^{3/2}S\sqrt{AK/c_{\min}})$, where D is the diameter of the SSP.

Our technical contribution is as follows. To better explain our main Bernstein-based algorithm, we start by assuming that the costs are lower bounded by $c_{\min}$ and give an algorithm based on Hoeffding inequalities that is simple to analyze and achieves a regret bound of $\widetilde{O}(B_\star^{3/2} S \sqrt{AK/c_{\min}})$. Note that this bound is comparable to the one of [TGV$^+$20], yet our algorithm and its analysis are significantly simpler and more intuitive. In addition, its analysis contains many of the key ideas of the proof of the Bernstein-based algorithm, and is much easier to follow. We subsequently present the Bernstein-based algorithm. This algorithm is simpler than our first one mainly since picking the parameters of the optimistic model is particularly easy. The analysis, however, is somewhat more delicate. Eventually, we achieve our final regret bound by perturbing the instantaneous costs to be at least $\varepsilon > 0$. The additional cost due to this perturbation has a small effect since the dependency of our regret on $c_{\min}$ is additive and does not multiply any term depending on K.

1.2.4 Minimax Optimal Regret for SSP

Our first algorithm for regret minimization in SSP leaves a gap of $\sqrt{S}$ between the upper and lower bounds. Moreover, for simplicity, there we assumed that the cost function is deterministic and known. Now we consider the case where the costs are i.i.d. and initially unknown. We prove upper and lower bounds for this case, proving that the optimal regret is of order $\widetilde{\Theta}(\sqrt{(B_\star^2 + B_\star)SAK})$.

Both our first algorithm and that of [TGV$^+$20] were based on a direct application of the "Optimism in the Face of Uncertainty" principle to the SSP model, following the ideas behind the UCRL2 algorithm [JOA10] for average-reward MDPs. Here we take a different approach. We propose a novel black-box reduction to finite-horizon MDPs, showing that the SSP problem is not harder than the finite-horizon setting assuming prior knowledge on the expected time it takes for the optimal policy to reach the goal state. While the reduction itself is simple, the analysis is highly nontrivial as one has to show that the goal state is indeed reached in every episode without incurring excessive costs in the process.

The idea of reducing SSP to finite-horizon was previously used by [CLW21, CL21] for SSP with adversarially changing costs. However, they run one finite-horizon episode in every SSP episode and then simply try to reach the goal as fast as possible, while we restart a new finite-horizon episode every H steps. This modification is what enables us to obtain the optimal and improved dependence in the number of states.

In addition, we provide a new algorithm for regret minimization in finite-horizon MDPs called ULCVI. We show that (for large enough number of episodes) its regret depends polynomially on the expected cost of the optimal policy $B_\star$, and only logarithmically on the horizon length H. This implies that the correct measure for the regret is the expected cost of the optimal policy and not the length of the horizon. We note that regret with logarithmic dependence in the horizon H was also obtained by [ZJD21], yet they make a much stronger assumption: that the cumulative cost of every trajectory is bounded by 1. In contrast, we only assume that the expected cost of the optimal policy is bounded by some constant $B_\star$, while other policies may suffer a cost of H.

Our reduction, when combined with our finite-horizon algorithm ULCVI, guarantees SSP regret of $\widetilde{O}(\sqrt{(B_\star^2 + B_\star)SAK})$. This matches our first lower bound $B_\star \geq 1$ up to logarithmic factors. However, this lower bound does not hold for $B_\star < 1$ suggesting that this is not the correct rate in this case. Indeed, we prove a tighter lower bound of $\Omega(\sqrt{B_\star SAK})$ for $B_\star < 1$, showing that our regret guarantees are minimax optimal in all cases.

1.2.5 Adversarial SSP

We present the adversarial SSP model that introduces adversarially changing costs to the classical SSP model. Formally, the agent interacts with an SSP instance for K episodes, and the cost function changes arbitrarily between episodes. The agent's objective is to reach the goal state in all episodes while minimizing its total expected cost.

As pointed out by [TGV$^+$20], in the general SSP problem we face new challenges that do not arise in the loop-free version (i.e., finite-horizon MDPs). Notably, it features two possibly conflicting objectives – reaching the goal vs minimizing cost; and it requires handling unbounded value functions and episode lengths. In the adversarial SSP model, these difficulties are further amplified as the adversary might encourage the learner to use "slow" policies and then punish her with large costs.

We propose the first algorithms for regret minimization in adversarial SSPs without any restrictive assumptions (namely, loop-free assumption). While we leverage algorithmic and technical tools from both SSP and finite-horizon adversarial MDP, tackling the general SSP problem in the presence of an adversary requires novel techniques and careful analysis. Our algorithms are based on the popular online mirror descent (OMD) framework for online convex optimization (OCO). However, naive application of OMD to SSP cannot overcome the challenges mentioned above as we show, and we use carefully designed mechanisms to establish our theoretical guarantees.

Our main contributions are as follows. First, we formalize the adversarial SSP model and define the notion of learning and regret. Second, we establish an efficient implementation of OMD in the SSP model with known transitions and study the conditions under which it guarantees near-optimal $\sqrt{K}$ expected regret, showing that some modifications are necessary. Then, we illustrate the challenge of obtaining regret bounds in high-probability in adversarial SSPs, and present a novel method that allows OMD to obtain its regret with high-probability. Finally, we tackle unknown transitions. We describe the crucial adaptations that allow OMD to be combined with optimistic estimates of the transition function and guarantee $\sqrt{K}$ regret when all costs are strictly positive, and $K^{3/4}$ regret in the general case.

1.3 Prior Work

Regret Minimization in Stochastic MDP. The works of [JOA10, BT09] initiated the study on regret minimization in MDPs. They prove regret bounds of $\widetilde{O}(H^2S\sqrt{AK})$, and their algorithms use the "optimism in face of uncertainty" principle, which proves to be highly useful in adversarial environments as well. Later, the works of [AOM17, ZB19, EMGM19, DLWB19] managed to design improved algorithms, based on similar principles, that attain the optimal regret of $\widetilde{O}(\sqrt{H^3SAK})$ (ignoring logarithmic factors). While all previous algorithms are model-based, [JAZBJ18] presented an optimistic version of the popular model-free algorithm Q-learning with similar regret guarantees. The lower bound of $\Omega(\sqrt{H^3SAK})$ is due to [JOA10, OVR16, JAZBJ18].

Regret Minimization in Adversarial MDP. The work of [EKM09], which presented the adversarial MDP model, assumes full knowledge of the transition function and full-information feedback about the losses. They propose an algorithm, MDP-E, which uses an experts algorithm in each state and achieves $O(\tau\sqrt{T\log A})$ regret, where τ is a bound on the mixing time of the MDP and T is the number of time steps. Another early work in this setting, by [YMS09], achieves an $\widetilde{O}(T^{2/3})$ regret. In the bandit setting, but still assuming full knowledge of the transition function, the work of [NGS10] achieves an $\widetilde{O}(H^2\sqrt{AK}/\alpha)$ regret, where $\alpha > 0$ is a lower bound on the probability to reach some state s under some policy π. Later [NGSA14] eliminate the dependence on α but achieve only $\widetilde{O}(K^{2/3})$ regret. A later work, by [ZN13], proposed the O-REPS algorithm which guarantees an $\widetilde{O}(H\sqrt{SAK})$ regret. The setting where the transition function is unknown is much more challenging and only one algorithm was previously presented for it, and assumed full-information feedback. The FPOP algorithm [NGS12] achieves $\widetilde{O}(H^2SA\sqrt{K})$ regret.

SSP. Early work by [BT91] studied the problem of planning in SSPs, that is, computing the optimal strategy efficiently in a known SSP instance. They established that, under certain assumptions, the optimal strategy is a deterministic stationary policy (a mapping from states to actions) and can be computed efficiently using standard planning algorithms, e.g., Value Iteration or Policy Iteration. The only regret minimization algorithm specifically designed for SSP is that of [TGV$^+$20] that assumes that all costs are bounded away from zero (i.e., there is a $c_{\min} > 0$ such that all costs are in the range $[c_{\min}, 1]$). They show a regret bound that scales as $\widetilde{O}(D^{3/2}S\sqrt{AK/c_{\min}})$, where D is the minimum expected time of reaching the goal state from any state. In addition, they show that the algorithm's regret is $\widetilde{O}(K^{2/3})$ when the costs are arbitrary (namely, may be zero).

1.4 Organization

Chapter 2 covers some background on Markov decision processes and reinforcement learning. It defines formally all the models and notations used throughout this thesis, and presents some fundamental algorithms and results.

Chapters 3 and 4 present our results on adversarial MDPs. They are based on the papers [RM19a] and [RM19b], respectively. Chapter 3 considers the model of full-information feedback, while Chapter 4 focuses on the model of bandit feedback.

Chapters 5 and 6 present our results on SSPs. They are based on the papers [RCMK20] and [CEMR21], respectively. Chapter 5 presents the first near-optimal regret for SSP and the first lower bound, while Chapter 6 presents an improved algorithm that achieves optimal regret guarantees.

Chapter 7 presents our results on adversarial SSPs, and is based on the paper [RM21b].

1.5 Excluded Work

This dissertation contains my main lines of work. I have contributed to other works during my PhD studies which are to varying extent beyond this scope. These works are:

- Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic Policy Optimization with Bandit Feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020.*

- Aviv Rosenberg and Yishay Mansour. Oracle-Efficient Regret Minimization in Factored MDPs with Unknown Structure. In *Advances in Neural Information Processing Systems, NeurIPS 2021.*

- Tal Lancewicki, Aviv Rosenberg and Yishay Mansour. Learning Adversarial Markov Decision Processes with Delayed Feedback. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2022*.

- Liyu Chen, Haipeng Luo and Aviv Rosenberg. Policy Optimization for Stochastic Shortest Path. In *Proceedings of the Conference on Learning Theory, COLT 2022*.

- Tal Lancewicki, Aviv Rosenberg and Yishay Mansour. Cooperative Online Learning in Stochastic and Adversarial MDPs. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022*.

- Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour and Aviv Rosenberg. Near-Optimal Regret for Adversarial MDP with Delayed Bandit Feedback. In *Advances in Neural Information Processing Systems, NeurIPS 2022*.

- Aviv Rosenberg, Assaf Hallak, Shie Mannor, Gal Chechik and Gal Dalal. Planning and Learning with Adaptive Lookahead. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2023*.

2 Model and Preliminaries

2.1 Finite-Horizon Markov Decision Processes

Markov Decision Processes (MDPs) [Put14, Ber95] are one of the most important and well-known frameworks for stochastic decision making. In this thesis, we supply results on both finite-horizon MDPs which are defined as follows and goal-oriented MDPs, i.e., Stochastic Shortest Path (SSP), which are defined in the following section.

A finite-horizon MDP $\mathscr{M}$ is defined by the tuple $(\mathscr{S}, \mathscr{A}, P, c, H)$. $\mathscr{S}$ and $\mathscr{A}$ are finite state and action spaces. Their sizes are denoted by $|\mathscr{S}| = S$ and $|\mathscr{A}| = A$, respectively. The parameter H represents the horizon, i.e., the length of the interaction. $P = \{P_h : \mathscr{S} \times \mathscr{A} \to \Delta_S\}_{h=1}^{H}$ is the transition model. It is a collection of H mappings from a state-action pair $(s,a) \in \mathscr{S} \times \mathscr{A}$ to a probability distribution over $\mathscr{S}$, denoted by Δ_S. We denote by $P_h(s' \mid s,a)$ the probability to transition to state s' when taking action a in state s at the h-th time step. $c = \{c_h : \mathscr{S} \times \mathscr{A} \to \mathbb{R}\}_{h=1}^{H}$ is the reward model, which is bounded in $[0,1]$.

A policy $\pi = \{\pi_h : \mathscr{S} \to \Delta_A\}_{h=1}^{H}$ is a collection of H mappings from the state space to a probability distribution over actions. We denote by $\pi_h(a \mid s)$ the probability to take action a in state s at the h-th time step, when playing the policy π. When the policy π is deterministic, we often denote by $\pi_h(s)$ the action a for which $\pi_h(a \mid s) = 1$. The expected cost of policy π from an initial state $s \in \mathscr{S}$ in time step h, referred to as the *value function*, is defined as follows:

$$V_h^{\pi}(s;c,P) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi, P\right], \tag{2.1}$$

where the expectations is taken w.r.t. to all existing randomness (the transition model and the policy). When clear from context, we sometimes omit the notations of c and P in the value function, i.e., we use $V_h^{\pi}(s)$ or $V_h^{\pi}(s;c)$. The following linear equations hold for the

Algorithm 1 Value Function Computation

```
init: $V^{\pi}_{H+1}(s) = 0 \; \forall s \in \mathcal{S}$.
for $h = H, H-1, \dots, 1$ do
    $\forall s \in \mathcal{S},\; V^{\pi}_h(s) \leftarrow \sum_{a \in \mathcal{A}} \pi_h(a \mid s)\big(c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^{\pi}_{h+1}(s')\big)$.
end for
return: $V^{\pi}$.
```

value function:

$$\forall s \in \mathcal{S}\; \forall h \in [H],\; V^{\pi}_h(s) = \sum_{a \in \mathcal{A}} \pi_h(a \mid s) \left(c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^{\pi}_{h+1}(s') \right). \tag{2.2}$$

Thus, computing the value function can be done in $O(S^2AH)$ time via dynamic programming (see Algorithm 1).

In many cases we are not only interested in estimating the expected value of a policy, but also want to optimize over it. That is, we want to find the *optimal policy* of an MDP. It is well known that the optimal strategy is a deterministic policy that does not depend on the initial state [Put14, Ber95]. We define it as the policy which minimizes the value $V^{\pi}_h(s)$ for every $h \in [H]$ and $s \in \mathcal{S}$, where $[H] = \{1, 2, \dots, H\}$, and denote it by $\pi^\star = \{\pi^\star_h\}_{h=1}^{H}$. The value of the optimal policy is called the *optimal value function* and denote by $V^\star$. Concretely,

$$\pi^\star \in \arg\min_{\pi} V^{\pi}$$
$$V^\star = \min_{\pi} V^{\pi}.$$

It is well known that the optimal value function satisfies the *Bellman equations* [Put14]:

$$\forall s \in \mathcal{S}\; \forall h \in [H],\; V^\star_h(s) = \min_{a \in \mathcal{A}} c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^\star_{h+1}(s'), \tag{2.3}$$

and that the optimal policy can be extracted from the optimal value by

$$\forall s \in \mathcal{S}\; \forall h \in [H],\; \pi^\star_h(s) \in \arg\min_{a \in \mathcal{A}} c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^\star_{h+1}(s'). \tag{2.4}$$

Similarly to value computation, we can compute the optimal policy using the Value Iteration (VI) algorithm (Algorithm 2) in $O(S^2AH)$ computational complexity.

Algorithm 2 Value Iteration

```
**init:** $V^\star_{H+1}(s) = 0 \; \forall s \in \mathcal{S}$.
**for** $h = H, H-1, \dots, 1$ **do**
    $\forall s \in \mathcal{S}$, $V^\star_h(s) \leftarrow \min_{a \in \mathcal{A}} c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^\star_{h+1}(s')$.
    $\forall s \in \mathcal{S}$, $\pi^\star_h(s) \leftarrow \arg\min_{a \in \mathcal{A}} c_h(s,a) + \sum_{s' \in \mathcal{S}} P_h(s' \mid s,a) V^\star_{h+1}(s')$.
**end for**
**return:** $\pi^\star, V^\star$.
```

2.2 Stochastic Shortest Path

An instance of the stochastic shortest path (SSP) problem is a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, c, s_{\text{init}}, g)$ where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. The agent begins at the initial state s_{init}, and ends her interaction with $\mathcal{M}$ by arriving at the goal state g (where $g \notin \mathcal{S}$). Whenever she plays action a in state s, she pays a cost $c(s,a) \in [0,1]$ and the next state $s' \in \mathcal{S}$ is chosen with probability $P(s' \mid s,a)$. To simplify the presentation we avoid addressing the goal state g explicitly – we assume that the probability of reaching the goal state by playing action a at state s is $1 - \sum_{s' \in \mathcal{S}} P(s' \mid s,a)$.

We now review planning in a known SSP instance. Under certain assumptions that we shall briefly discuss, the optimal behaviour of the agent, i.e., the policy that minimizes the expected total cost of reaching the goal state from *any* state, is a stationary, deterministic and proper policy. A stationary and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping that selects action $\pi(s)$ whenever the agent is at state s. A proper policy is defined as follows.

Definition 2.2.1 (Proper and Improper Policies)**.** A policy π is *proper* if playing π reaches the goal state with probability 1 when starting from any state. A policy is *improper* if it is not proper.

Any policy π induces a *cost-to-go function* $V^\pi : \mathcal{S} \to [0, \infty]$ defined as:

$$V^\pi(s) = \lim_{T \to \infty} \mathbb{E}_\pi \Big[\sum_{t=1}^{T} c(s_t, a_t) \mid s_1 = s \Big],$$

where the expectation is taken w.r.t the random sequence of states generated by playing according to π when the initial state is s. For a proper policy π, since the number of states S is finite, it follows that $V^\pi(s)$ is finite for all $s \in S$. However, note that $V^\pi(s)$ may be finite even if π is improper. We additionally denote by $T^\pi(s)$ the expected time it takes for π to reach g starting at s; in particular, if π is proper then $T^\pi(s)$ is finite for all s, and if π is improper there must exist some s such that $T^{\pi}(s) = \infty$. In this work we assume the following about the SSP model.

Assumption 2.2.1. *There exists at least one proper policy.*

With Assumption 2.2.1, we have the following important properties of proper policies. In particular, the first result shows that a policy is proper if and only if its cost-to-go function satisfies the Bellman equations. The second result proves that a policy is optimal if and only if it satisfies the Bellman optimality criterion. Note that they assume that every improper policy has high cost.

Lemma 2.2.2 ([BT91, Lemma 1])**.** *Suppose that Assumption 2.2.1 holds and that for every improper policy π' there exists at least one state $s \in \mathscr{S}$ such that $V^{\pi'}(s) = \infty$. Let π be any policy, then*

(i) If there exists $V : \mathscr{S} \to \mathbb{R}$ such that $V(s) \geq c\big(s, \pi(s)\big) + \sum_{s' \in \mathscr{S}} P\big(s' \mid s, \pi(s)\big) V(s')$ for all $s \in \mathscr{S}$, then π is proper. Moreover, it holds that $V^{\pi}(s) \leq V(s)$, $\forall s \in \mathscr{S}$.

(ii) If π is proper then V^{π} is the unique solution to the equations $V^{\pi}(s) = c\big(s, \pi(s)\big) + \sum_{s' \in \mathscr{S}} P\big(s' \mid s, \pi(s)\big) V^{\pi}(s')$ for all $s \in \mathscr{S}$.

Lemma 2.2.3 ([BT91, Proposition 2])**.** *Under the conditions of Lemma 2.2.2 the optimal policy $\pi^{\star}$ is stationary, deterministic, and proper. Moreover, a policy π is optimal if and only if it satisfies the Bellman optimality equations for all $s \in \mathscr{S}$:*

$$\begin{aligned} V^{\pi}(s) &= \min_{a \in \mathscr{A}} c\big(s, a\big) + \sum_{s' \in \mathscr{S}} P\big(s' \mid s, a\big) V^{\pi}(s'), \\ \pi(s) &\in \arg\min_{a \in \mathscr{A}} c\big(s, a\big) + \sum_{s' \in \mathscr{S}} P\big(s' \mid s, a\big) V^{\pi}(s'). \end{aligned} \tag{2.5}$$

In this work we are not interested in approximating the optimal policy overall, but rather the best *proper* policy. In this case the second requirement in the lemmas above, that for every improper policy π there exists some state $s \in S$ such that $V^{\pi}(s) = \infty$, can be circumvented in the following way [BY13]. First, note that this requirement is trivially satisfied when all instantaneous costs are strictly positive. Then, one can perturb the instantaneous costs by adding a small positive cost $\varepsilon \in [0, 1]$, i.e., the new cost function is $c_{\varepsilon}(s, a) = \max\{c(s, a), \varepsilon\}$. After this perturbation, all proper policies remain proper, and every improper policy has infinite cost-to-go from some state (as all costs are positive). In the modified MDP, we apply Lemma 2.2.3 and obtain an optimal policy $\pi^{\star}_{\varepsilon}$ that is stationary, deterministic and proper and has a cost-to-go function $V^{\star}_{\varepsilon}$. Taking the limit as $\varepsilon \to 0$, we have that $\pi^{\star}_{\varepsilon} \to \pi^{\star}$ and $V^{\star}_{\varepsilon} \to V^{\star}$, where $\pi^{\star}$ is the optimal *proper* policy in the original model that is also stationary and deterministic, and $V^{\star}$ denotes its cost-to-go function.

Algorithm 3 Reinforcement Learning - Computational Model

```
for $k = 1, 2, \ldots, K$ do
    Observe initial state $s_1^k = s_{\text{init}}$.
    Pick a policy $\pi^k$.
    for $h = 1, 2, \ldots, H$ do
        Observe current state $s_h^k$.
        Pick action $a_h^k \sim \pi_h^k(\cdot \mid s_h^k)$.
        if Stochastic MDP then
            Observe and suffer cost $C_h^k = c_h^k(s_h^k, a_h^k)$.
        else if Adversarial MDP then
            Observe and suffer cost $C_h^k \sim c_h(s_h^k, a_h^k)$.
        end if
        Observe next state $s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)$.
    end for
    if Full-Information Feedback then
        Observe cost function $c^k$.
    end if
end for
```

2.3 Adversarial Markov Decision Processes

In many real-world applications, unlike in MDPs, the environment changes over time and even throughout the learning process. To address this issue, the adversarial MDP model [EKM09] was proposed. In this model, the cost function can change arbitrarily (while still assuming a fixed stochastic transition function). Formally, there are K episodes of interaction between the agent and the environment. The cost function in the k-th episode is c^k, i.e., in finite-horizon adversarial MDPs $c^k = \{c_h^k : \mathscr{S} \times \mathscr{A} \to [0,1]\}$ and in adversarial SSPs $c^k : \mathscr{S} \times \mathscr{A} \to [0,1]$. The sequence of cost functions $\{c^k\}_{k=1}^K$ is chosen by an oblivious adversary before the interaction starts.

Importantly, the adversarial MDP model generalizes the standard MDP model. Concretely, stochastic MDPs are a specific case of adversarial MDPs in which the cost $c_h^k(s,a)$ is sampled i.i.d for each $(k,s,a,h) \in [K] \times \mathscr{S} \times \mathscr{A} \times [H]$ from a distribution with expected value $c_h(s,a)$.

2.4 Reinforcement Learning

It is often the case that the exact model of an MDP is unknown, however, interaction with the unknown model is possible. An optimal policy can be learned through interaction with the unknown MDP based on samples. The field of Reinforcement Learning (RL) [SB98] tackles the question of how to learn an optimal policy using samples. Recently, RL witnessed remarkable empirical success, e.g., [MKS$^+$15, LFDA16, SSS$^+$17]. The empirical success acted as driving force to significant theoretical developments. Next we survey the major advancements in the theory of regret analysis in RL.

In the RL problem, the agent needs to trade-off between *exploration* and *exploitation*. That is, to control whether it needs to have a better estimate of the model or whether it can act optimally with respect to (w.r.t.) the 'empirical' model. By generalizing techniques and algorithms from Multi-Armed Bandit literature [LS20, Sli19] many RL algorithms were suggested and analyzed in the last two decades [AJO09, KS02, BT02, AOM17, JAZBJ18].

Large portion of recent research was devoted to RL for the case the environment is an unknown finite-horizon MDP. For stochastic MDPs, the considered computational model assumes an episodic interaction, in which an RL agent interacts with the finite-horizon MDP for H time steps. Then, the state is initialized to the initial state s_{init}. For adversarial MDPs, the interaction is similar but the cost function changes between episodes. In the end of each episode, the agent observes either the entire cost function for *full-information feedback*, or only the suffered costs for *bandit feedback*. See the full interaction in Algorithm 3.

The common performance measure is the *regret*, which compares the cost suffered by the agent with that suffered by the best fixed policy in hindsight. For stochastic MDPs, it is defined as:

$$R_K = \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}; c, P) - V_1^{\star}(s_{\text{init}}) = \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}) - V_1^{\star}(s_{\text{init}}),$$

and for adversarial MDPs:

$$R_K = \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}; c^k, P) - \min_{\pi} \sum_{k=1}^{K} V_1^{\pi}(s_{\text{init}}; c^k, P) = \sum_{k=1}^{K} V_1^{k,\pi^k}(s_{\text{init}}) - V_1^{k,\pi^\star}(s_{\text{init}}).$$

The definitions for SSP are slightly different and are described in the appropriate chapter.

In [OVR16] a lower bound of $\Omega(\sqrt{H^3 SAK})$ was established for RL in finite-horizon MDPs. Note that this lower bound also applies to adversarial MDPs, as they are a more general model. Furthermore, [ZB19] analyzed the EULER algorithm and established an $\widetilde{O}(\sqrt{H^3 SAK})$ [1] upper bound, which shows the lower bound is tight.

[1] We omit poly-logarithmic factors in the $\widetilde{O}(\cdot)$ notation.

3 Learning Adversarial MDPs with Unknown Transition Function and Full-Information Feedback

This chapter is based on:

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019.*

This chapter presents the first high probability regret bound for adversarial MDP with unknown transitions. We start the chapter by introducing the concept of occupancy measures and reviewing the O-REPS algorithm [ZN13] for regret minimization in adversarial MDP with known transitions. Then, we present an extension of occupancy measures to the case where the transitions are unknown to the agent, and finally present our algorithm and analyze its regret.

3.1 Occupancy Measures

It is beneficial to introduce the concept of occupancy measures on the state-action space $\mathscr{S} \times \mathscr{A} \times [H]$. For a policy π we define the occupancy measure q^{π} as follows:

$$q_h^{\pi}(s,a) = \Pr\left[s_h = s, a_h = a \mid \pi\right].$$

It is easy to see that the occupancy measure of any policy π satisfies

$$\sum_{a \in \mathscr{A}} q_{h+1}^{\pi}(s,a) = \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_h^{\pi}(s',a) P_h(s \mid s',a) \ \forall (s,h) \in \mathscr{S} \times [H-1],$$

with $q_1^\pi(s,a) = \pi_1(a \mid s)\mathbb{1}\{s = s_{\text{init}}\}$. The set of all occupancy measures satisfying the above equality in the MDP M will be denoted as $\Delta(\mathscr{M})$. The policy π is said to generate the occupancy measure $q \in \Delta(\mathscr{M})$ if $\pi_h(a \mid s) = q(s,a)/q(s)$ holds for all $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$, where $q(s) = \sum_{a\in\mathscr{A}} q(s,a)$. It is clear that there exists a unique generating policy for all measures in $\Delta(\mathscr{M})$ and vice versa. The policy generating q will be denoted as π^q. In what follows, we will redefine the task of the learner from having to select policies π^k to having to select occupancy measures $q^{\pi^k} \in \Delta(\mathscr{M})$ in each episode k. To see why this notion simplifies the treatment of the problem, observe that:

$$\begin{aligned} V_1^\pi(s_{\text{init}}; c) &= \mathbb{E}\left[\sum_{h=1}^{H} c_h(s_h, a_h) \mid s_1 = s_{\text{init}}, \pi, P\right] \\ &= \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} q_h^\pi(s,a) c_h(s,a) \\ &\overset{\text{def}}{=} \langle q^\pi, c\rangle. \end{aligned}$$

3.2 Reduction to Online Linear Optimization and the O-REPS Algorithm

Using the notation from the previous section, we can reformulate our original problem as an instance of online linear optimization with decision space $\Delta(\mathscr{M})$. Assuming that the learner selects occupancy measure q^k in episode k, the regret can be rewritten as:

$$R_K = \max_{q\in\Delta(\mathscr{M})} \sum_{k=1}^{K} \langle q^k - q, c^k\rangle = \sum_{k=1}^{K} \langle q^k - q^\star, c^k\rangle.$$

The O-REPS algorithm [ZN13] is an instance of online linear optimization methods usually referred to as Follow-the-Regularized-Leader (FTRL) or Online Mirror Descent (OMD). Before describing the algorithm, some more definitions are in order. First, define $\mathrm{KL}(q \parallel q')$ as the unnormalized Kullback-Leibler divergence between two occupancy measures q and q':

$$\mathrm{KL}(q \parallel q') = \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} q_h(s,a) \log \frac{q_h(s,a)}{q_h(s,a)} - q_h(s,a) + q'_h(s,a).$$

Note that $\sum_{h,s,a} q'_h(s,a) - q_h(s,a) = 0$, but adding these terms will help simplify some of the derivations. Let $R(q)$ define the unnormalized negative entropy of the occupancy measure q:

$$R(q) = \sum_{h=1}^{H} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_h(s,a) \log q_h(s,a) - q_h(s,a).$$

In the first episode, O-REPS chooses the uniform policy with $\pi_h^1(a \mid s) = 1/A$ for all s and a, and we let $q^1 = q^{\pi^1}$. Then, the algorithm proceeds as follows. After observing the feedback from episode k, it selects the occupancy measure that solves the optimization problem:

$$q^{k+1} = \arg \min_{q \in \Delta(\mathscr{M})} \eta \langle q, c^k \rangle + \mathrm{KL}(q \parallel q^k), \tag{3.1}$$

where $\eta > 0$ is a learning rate. This optimization can be reformulated as first solving the unconstrained optimization problem and then projecting the result to $\Delta(\mathscr{M})$, i.e.,

$$\begin{aligned} \tilde{q}^{k+1} &= \arg\min_{q} \eta \langle q, c^k \rangle + \mathrm{KL}(q \parallel q^k) \\ q^{k+1} &= \arg \min_{q \in \Delta(\mathscr{M})} \mathrm{KL}(q \parallel \tilde{q}^{k+1}). \end{aligned}$$

The first step can be simply carried out by setting $\tilde{q}_h^{k+1}(s,a) = q_h^k(s,a) e^{-\eta c_h^k(s,a)}$. The projection step can be performed using the following lemma.

Lemma 3.2.1 ([ZN13], Proposition 1)**.** *It holds that*

$$q_h^{k+1}(s,a) = \frac{q_h^k(s,a) e^{B_h^k(s,a \mid v^k)}}{Z_h^k(v^k)},$$

for:

$$\begin{aligned} B_h^k(s,a \mid v) &= v_h(s) - \eta c_h^k(s,a) - \sum_{s'} P_h(s' \mid s,a) v_{h+1}(s') \\ Z_h^k(v) &= \sum_{s,a} q_h^k(s,a) e^{B_h^k(s,a \mid v)} \\ v^k &= \arg\min_{v} \sum_{h} \log Z_h^k(v). \end{aligned}$$

Minimizing the expression on the right-hand side of the last equation is an unconstrained convex optimization problem and can be solved efficiently. [ZN13] also show that this algorithm achieves regret of $\widetilde{O}(H\sqrt{K})$ which optimal up to logarithmic factors.

3.3 Extending Occupancy Measures to Unknown Transitions

When the transition function is unknown to the learner, we cannot compute the occupancy measure of a policy π or the constraints that define the set of occupancy measures. A naive solution could be to treat $\{P_h(s' \mid s,a)\}_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}$ as additional variables in the optimization problem solved by the O-REPS algorithm in each episode. However, this leads to a non-convex optimization problem which cannot be solved efficiently.

Instead we propose to extend the definition of occupancy measures such that it contains not only the policy, but also the transition function. Namely, we define the occupancy measure $q^{\pi,P}$ of the policy π and the transition function P as follows:

$$q_h^{\pi,P}(s,a,s') = \Pr\left[s_h = s, a_h = a, s_{h+1} = s' \mid \pi, P\right].$$

We start with two basic properties that hold for every occupancy measure q. By standard flow constraints, it holds that in each episode the learner will go through every layer. Therefore, for every $h = 1,\ldots,H$:

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} q_h(s,a,s') = 1. \tag{3.2}$$

Moreover, the probability to enter a state when coming from the previous layer is exactly the probability to visit that state. Thus, for every $h = 2,\ldots,H$ and every $s\in\mathcal{S}$:

$$\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} q_h(s,a,s') = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} q_{h-1}(s',a,s). \tag{3.3}$$

Notice that every occupancy measure q induces a transition function and a policy. We denote them as P^q and π^q respectively, and they can be computed as follows:

$$\begin{aligned} P_h^q(s' \mid s,a) &= \frac{q_h(s,a,s')}{q_h(s,a)} \\ \pi_h^q(a \mid s) &= \frac{q_h(s,a)}{q_h(s)}, \end{aligned}$$

where $q_h(s,a) = \sum_{s'\in\mathcal{S}} q_h(s,a,s')$ and $q_h(s) = \sum_{a\in\mathcal{A}} q_h(s,a)$. The following lemma characterizes $\Delta(\mathcal{M})$ and its proof is straightforward.

Lemma 3.3.1. *For every $q = \{q_h : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]\}_{h\in[H]}$ it holds that $q\in\Delta(\mathcal{M})$ if and only if* (3.2) *and* (3.3) *hold, and $P^q = P$ (where P is the transition function of $\mathcal{M}$).*

The regret is reformulated with occupancy measures similarly to the previous section, but as we will see next, the extended occupancy measures will enable us to extend the O-REPS algorithm to unknown transitions efficiently.

3.4 The UC-O-REPS Algorithm

Our algorithm “Upper Confidence Online Relative Entropy Policy Search” (UC-O-REPS) is presented in Algorithms 4 and 5. It is inspired by the O-REPS algorithm [ZN13] in the sense that it picks occupancy measures instead of policies. However, unlike our algorithm, O-REPS assumes full knowledge of the transition function. To the best of our knowledge, the only algorithm that handles unknown transition probabilities in adversarial MDPs is FPOP [NGS12], which uses a Follow the Pertubed Leader method [KV03] in the space of the policies.

Recall that the adversarial MDP has a stochastic element - the transition function, and an adversarial element - the cost functions.

To handle the stochastic transition function we use the framework of epochs and confidence sets, first introduced by the UCRL-2 algorithm [JOA10]. In this framework, the algorithm maintains confidence sets that contain the actual MDP with high probability, but also shrink as time progresses. We translated this method to the occupancy measures space, and the full details can be found in Section 3.4.1.

The core of the algorithm is the way we choose the occupancy measure for each episode from within the confidence set. This is done by the Online Mirror Descent method [Sha12] for online linear optimization, since we deal with an arbitrary sequence of cost functions. The full details of adapting OMD to our setting can be found in Section 3.4.2.

The combination of these two methods is done using an important principle in reinforcement learning - “optimism in face of uncertainty”. On the one hand, we keep confidence sets to handle the uncertainty, but on the other hand, within these confidence sets, we solve an OMD optimization problem optimistically (without thinking about the transition function estimation).

3.4.1 Confidence Sets

Since the learner does not know the transition function, it has to estimate P from its experience. Using this estimate we define confidence sets, and choose occupancy measures from within them. Notice that these occupancy measures might not be in $\Delta(\mathscr{M})$, i.e., their induced transition function may differ from P. Nevertheless, we can still use them to compute policies and execute those policies.

The algorithm proceeds in epochs of random length, and in the beginning of each epoch the confidence set is updated. The first epoch E_1 starts at episode $k = 1$, and each epoch E_i ends when the number of visits at some state-action pair (s,a) is doubled. Let k_i denote the index of the first episode in epoch E_i, and $i(k)$ denote the index of the epoch that includes episode k. Let $N_h^i(s,a)$ and $M_h^i(s' \mid s,a)$ denote the number of times state-action pair (s,a) was visited (in step h) and the number of times this event was followed by a transition to s' up to episode k_i, respectively. That is

$$N_h^i(s,a) = \sum_{j=1}^{k_i-1} \mathbb{I}\left\{s_h^j = s, a_h^j = a\right\}$$

$$M_h^i(s' \mid s,a) = \sum_{j=1}^{k_i-1} \mathbb{I}\left\{s_h^j = s, a_h^j = a, s_{h+1}^j = s'\right\}.$$

Our estimate $\bar{P}^i$ for the transition function in epoch E^i is

$$\bar{P}_h^i(s' \mid s,a) = \frac{M_h^i(s' \mid s,a)}{\max\left\{1, N_h^i(s,a)\right\}},$$

and we define our confidence set $\Delta(\mathscr{M}, i)$ in epoch E^i to include all the occupancy measures that their induced transition function is "close enough" to $\bar{P}^i$. More formally, given a confidence parameter $\delta > 0$, we define:

$$\varepsilon_h^i(s,a) = \sqrt{\frac{2S \ln \frac{KHSA}{\delta}}{\max\{1, N_h^i(s,a)\}}},$$

and say that $\Delta(\mathscr{M}, i)$ consists of all $q = \{q_h : \mathscr{S} \times \mathscr{A} \times \mathscr{S} \to [0,1]\}_{h \in [H]}$ for which (3.2) and (3.3) hold, and

$$\left\|P_h^q(\cdot \mid s,a) - \bar{P}_h^i(\cdot \mid s,a)\right\|_1 \leq \varepsilon_h^i(s,a) \tag{3.4}$$

for every $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$.

Notice that these confidence sets shrink as time progresses, but the following lemma [JOA10, NGS12] shows that they still contain $\Delta(\mathscr{M})$ with high probability.

Lemma 3.4.1. *For any* $0 < \delta < 1$,

$$\left\|P_h(\cdot \mid s,a) - \bar{P}_h^i(\cdot \mid s,a)\right\|_1 \leq \sqrt{\frac{2S \ln \frac{KHSA}{\delta}}{\max\{1, N_h^i(s,a)\}}}$$

holds with probability at least $1-\delta$ *simultaneously for all* $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$ *and all epochs.*

3.4.2 Optimization Problem

In order to choose the occupancy measure q^k for episode k, the algorithm follows the OMD method. The idea behind this method is to choose an occupancy measure that minimizes the cost in episode k, while not straying too far from the previously chosen occupancy measure. Formally, given a parameter $\eta > 0$,

$$q^{k+1} = \arg \min_{q \in \Delta(\mathscr{M}, i(k))} \eta \left\langle q, c^k \right\rangle + \mathrm{KL}(q \parallel q^k),$$

where $\mathrm{KL}(q \parallel q^k)$ is the unnormalized KL divergence between two occupancy measures defined as

$$\mathrm{KL}(q \parallel q') = \sum_{h=1}^{H} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h(s,a,s') \log \frac{q_h(s,a,s')}{q_h(s,a,s')} - q_h(s,a,s') + q'_h(s,a,s').$$

We now proceed to show that this optimization problem can be solved efficiently. From the theory of OMD it is known that we can split this problem as follows: we start by solving the unconstrained problem, and then project the unconstrained minimizer into the feasible set, namely,

$$\begin{aligned} \tilde{q}_{k+1} &= \arg\min_{q} \eta \left\langle q, c^k \right\rangle + \mathrm{KL}(q \parallel q^k) \\ q^{k+1} &= \arg \min_{q \in \Delta(\mathscr{M}, i(k))} \mathrm{KL}(q \parallel \tilde{q}^{k+1}). \end{aligned} \tag{3.5}$$

The unconstrained problem can be solved by setting $\tilde{q}_h^{k+1}(s,a,s') = q_h^k(s,a,s') e^{-\eta c_h^k(s,a)}$ for every $(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$. Theorem 3.4.2 shows that the second optimization problem can be reduced to a convex optimization problem with only non-negativity constraints (and no constraints about the relations between the variables), which can be solved efficiently using iterative methods [BV04].

Theorem 3.4.2. *It holds that*

$$q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s') e^{B_h^k(s,a,s' | v^{\mu^k}, e^{\mu^k, \beta^k})}}{Z_h^k(v^{\mu^k}, e^{\mu^k, \beta^k})},$$

for:

$$\begin{aligned}
B_h^k(s,a,s' \mid v,e) &= e_h(s,a,s') + v_h(s,a,s') - \eta c_h^k(s,a) - \sum_{s''} \bar{P}_h^k(s'' \mid s,a) v_{h+1}(s,a,s'') \\
v_h^{\mu}(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\
e_h^{\mu,\beta}(s,a,s') &= \beta_{h+1}(s') - \beta_h(s) + \sum_{s''} (\mu_h^-(s,a,s'') + \mu_h^+(s,a,s''))\varepsilon_h^k(s'' \mid s,a) \\
\varepsilon_h^k(s' \mid s,a) &= \sqrt{\frac{16\bar{P}_h^k(s'|s,a)\log\frac{10HSAK}{\delta}}{n_h^k(s,a) \vee 1}} + \frac{10\log\frac{10HSAK}{\delta}}{n_h^k(s,a) \vee 1} \\
Z_h^k(v,e) &= \sum_{s,a,s'} q_h^k(s,a,s') e^{B_h^k(s,a,s'|v,e)} \\
\mu^k, \beta^k &= \arg\min_{\beta,\mu \geq 0} \sum_{h=1}^{H} \log Z_h^k(v^{\mu}, e^{\mu,\beta}),
\end{aligned} \tag{3.6}$$

where $\bar{P}^k = \bar{P}^{i(k)}$ *and* $n^k = N^{i(k)}$.

Proof. First of all we would like to reformulate optimization problem (3.5) as a convex optimization problem. Notice that the target function is convex (since it is the KL-divergence) and so are constraints (3.2), (3.3) of $\Delta(\mathscr{M}, i)$ (where $i = i(k)$). As for constraint (3.4), we will need to write it differently.

Let $(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$, we can replace

$$\left\| \frac{q_h(s,a,\cdot)}{\sum_{y\in\mathscr{S}} q_h(s,a,y)} - \bar{P}_h^i(\cdot \mid s,a) \right\|_1 \leq \varepsilon_h^i(s,a).$$

with $S+1$ constraints as follows. For each $s' \in \mathscr{S}$ we bound the difference in the transition probability with a new variable $\varepsilon_h'(s,a,s')$ and then we bound their sum with the original bound $\varepsilon_h^i(s,a)$. That is,

$$\begin{aligned}
\left| \frac{q_h(s,a,s')}{\sum_{y\in\mathscr{S}} q_h(s,a,y)} - \bar{P}_h^i(s' \mid s,a) \right| &\leq \varepsilon_h'(s,a,s') \\
\sum_{s'\in\mathscr{S}} \varepsilon_h'(s,a,s') &\leq \varepsilon_h^i(s,a).
\end{aligned}$$

Now we can get rid of the denominator by multiplying the equation and then replacing $\varepsilon_h'(s,a,s')$ with a different variable $\varepsilon_h(s,a,s') = \varepsilon_h'(s,a,s')\sum_{y\in\mathscr{S}} q_h(s,a,y)$. Moreover, we will discard the absolute value by replacing it with two linear constraints. The resulting

Algorithm 4 UC-O-REPS Algorithm

```
Input: state space $\mathscr{S}$, action space $\mathscr{A}$, number of episodes $K$, optimization parameter $\eta$ and confidence parameter $\delta$.
Initialization: $i(1) \leftarrow 1, k_1 \leftarrow 1, \pi_h^1(a|s) \leftarrow 1/A, q_h^1(s,a,s') \leftarrow 1/(S^2A), n_h^1(s,a) \leftarrow 0, N_h^1(s,a) \leftarrow 0, m_h^1(s'|s,a) \leftarrow 0, M_h^1(s'|s,a) \leftarrow 0 \; \forall (s,a,s',h)$.
for $k = 1,\dots,K$ do
    Play policy $\pi^k$ and observe trajectory $\{s_h^k, a_h^k\}_{h=1}^H$.
    Observe cost function $c^k$.
    Update epoch counters for $h = 1,\dots,H$:
        $n_h^{i(k)}(s_h^k, a_h^k) \leftarrow n_h^{i(k)}(s_h^k, a_h^k) + 1$
        $m_h^{i(k)}(s_{h+1}^k \mid s_h^k, a_h^k) \leftarrow m_h^{i(k)}(s_{h+1}^k \mid s_h^k, a_h^k) + 1.$
    if $\exists (s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]. \quad n_h^{i(k)}(s,a) \geq N_h^{i(k)}(s,a)$ then
        Start new epoch:
            $i(k+1) \leftarrow i(k) + 1 \quad ; \quad k_{i(k+1)} \leftarrow k + 1.$
        Initialize epoch counters $\forall (s,a,s',h)$:
            $n_h^{i(k+1)}(s,a) \leftarrow 0 \quad ; \quad m_h^{i(k+1)}(s' \mid s,a) \leftarrow 0.$
        Update total counters $\forall (s,a,s',h)$:
            $N_h^{i(k+1)}(s,a) \leftarrow N_h^{i(k)}(s,a) + n_h^{i(k)}(s,a)$
            $M_h^{i(k+1)}(s' \mid s,a) \leftarrow M_h^{i(k)}(s' \mid s,a) + m_h^{i(k)}(s' \mid s,a).$
        compute probability estimate $\forall (s,a,s',h)$:
            $\bar{P}^{i(k+1)}(s' \mid s,a) \leftarrow \frac{M_h^{i(k+1)}(s' \mid s,a)}{\max\left\{1, N_h^{i(k+1)}(s,a)\right\}}.$
    else
        Continue in the same epoch: $i(k+1) \leftarrow i(k)$.
    end if
    Compute policy for next episode:
        $q^{k+1}, \pi^{k+1} \leftarrow \texttt{Comp-Policy}(q^k, \bar{P}^{i(k+1)}, c^k).$
end for
```

Algorithm 5 Comp-Policy Procedure

```
**Input:** previous occupancy measure $q^k$, transition function estimate $\bar{P}^{i(k+1)}$ and current cost function $c^k$.
Solve optimization problem (3.6):
    $\mu^k, \beta^k = \arg\min_{\beta,\mu \geq 0} \sum_{h=1}^{H} \log Z_h^k(v^{\mu}, e^{\mu,\beta}).$
Compute next occupancy measure $\forall (s,a,s',h)$:
    $q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s') e^{B_h^k(s,a,s' | v^{\mu^k}, e^{\mu^k,\beta^k})}}{Z_h^k(v^{\mu^k}, e^{\mu^k,\beta^k})}.$
Compute next policy $\forall (s,a,h)$:
    $\pi_h^{k+1}(a \mid s) = \frac{\sum_{s' \in \mathcal{S}} q_h^{k+1}(s,a,s')}{\sum_{b \in \mathcal{A}} \sum_{s' \in \mathcal{S}} q_h^{k+1}(s,b,s')}.$
```

constraints are:

$$\begin{aligned} q_h(s,a,s') - \bar{P}_h^i(s' \mid s,a) \sum_{y \in \mathcal{S}} q_h(s,a,y) &\leq \varepsilon_h(s,a,s') \\ \bar{P}_h^i(s' \mid s,a) \sum_{y \in \mathcal{S}} q_h(s,a,y) - q_h(s,a,s') &\leq \varepsilon_h(s,a,s') \\ \sum_{s' \in \mathcal{S}} \varepsilon_h(s,a,s') &\leq \varepsilon_h^i(s,a) \sum_{s' \in \mathcal{S}} q_h(s,a,s'). \end{aligned}$$

This gives us a convex optimization problem with linear constraints. This problem obtains strong duality because: (1) The target function is bounded from below because KL-divergence is non-negative, (2) The target function and all constraints are convex, (3) Slater condition holds (easy to check).

Thus we can use the method of Lagrange multipliers, and we are ensured that the solution we get is optimal and finite. The full derivation can be found in the supplementary material and yields the aforementioned result. □

3.5 Analysis

In this section we bound the regret of the UC-O-REPS algorithm, by combining ideas from the regret analyses of OMD and UCRL-2. First we partition the regret into two terms: $\hat{R}^{APP}_{1:K}$ - which includes the error that comes from the estimation of the unknown transition function, and $\hat{R}^{ON}_{1:K}$ - which includes the error that comes from choosing sub-optimal policies. Formally,

$$\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}; c^k, P) - \min_{\pi} \sum_{k=1}^{K} V_1^{\pi}(s_{\text{init}}; c^k, P) \\
&= \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}; c^k, P) - V_1^{\pi^k}(s_{\text{init}}; c^k, P^k) \\
&\quad + \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}}; c^k, P^k) - \min_{\pi} \sum_{k=1}^{K} V_1^{\pi}(s_{\text{init}}; c^k, P) \\
&\overset{\text{def}}{=} \hat{R}^{APP}_{1:K} + \hat{R}^{ON}_{1:K},
\end{aligned}$$

where $P^k = P^{q^k}$ and $\pi^k = \pi^{q^k}$.

Theorems 3.5.2 and 3.5.3 bound each of these terms, which yields our main result.

Theorem 3.5.1. *Running UC-O-REPS in an adversarial MDP $\mathscr{M} = \left(\mathscr{S}, \mathscr{A}, H, P, \{c^k\}_{k=1}^K\right)$ yields the following regret,*

$$R_K \le O\left(H^2 S \sqrt{AK \log \frac{KHSA}{\delta}}\right).$$

3.5.1 Bounding $\hat{R}^{APP}_{1:K}$

The term $\hat{R}^{APP}_{1:K}$ is a result of the learner's lack of knowledge about the environment's dynamics. Since the dynamics are stochastic the learner estimates the transition probabilities to build confidence sets. It then selects occupancy measures from within these confidence sets, but they are not exactly occupancy measures of $\mathscr{M}$.

In this section we bound the difference between the loss of the learner's chosen policies in $\mathscr{M}$ and the loss of these policies in the "optimistic" MDP (the one induced by the occupancy measure q^k).

The way the algorithm minimizes this difference is through shrinking of the confidence sets. The following bound on $\hat{R}^{APP}_{1:K}$ is adapted from arguments in the regret analysis of UCRL-2, and the proof can be found in the supplementary material.

Theorem 3.5.2. *Let $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP. With probability at least $1-2\delta$, UC-O-REPS obtains,*

$$\hat{R}_{1:K}^{APP} \le O\left(H^2 S \sqrt{AK \log \frac{KHSA}{\delta}}\right).$$

3.5.2 Bounding $\hat{R}_{1:K}^{ON}$

The term $\hat{R}_{1:K}^{ON}$ is a result of the learner's lack of knowledge about the cost functions. Since the sequence of cost functions can be arbitrary, the learner handles it with tools from online convex optimization.

In this section we ignore the fact that the occupancy measures chosen by the learner are not exactly occupancy measures of $\mathcal{M}$, since this issue was already addressed in the previous section bounding $\hat{R}_{1:K}^{APP}$. Recall that

$$\hat{R}_{1:K}^{ON} = \sum_{k=1}^{K} \left\langle q^k - q, c^k \right\rangle,$$

for some occupancy measure $q \in \Delta(\mathcal{M})$ which is best in hindsight.

Now we can use arguments from online linear optimization. Specifically, the following theorem is an adaptation of OMD regret analysis to our setting.

Theorem 3.5.3. *Let $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP. With probability at least $1-\delta$, UC-O-REPS obtains the following for every $q \in \Delta(\mathcal{M})$.*

$$\hat{R}_{1:K}^{ON} \le O\left(\eta H K + \frac{H \log(HS^2A)}{\eta}\right),$$

and setting $\eta = \sqrt{\frac{\log(HS^2A)}{K}}$ yields

$$\hat{R}_{1:K}^{ON} \le O\left(H\sqrt{K \log(HS^2A)}\right).$$

Proof. By standard arguments of OMD regret analysis we have that

$$\sum_{k=1}^{K} \left\langle q^k - q, c^k \right\rangle \le \sum_{k=1}^{K} \left\langle q^k - \tilde{q}^{k+1}, c^k \right\rangle + \frac{\mathrm{KL}(q \parallel q^1)}{\eta}.$$

However, these arguments assume that q^k are chosen from within $\Delta(\mathcal{M})$ so we need to show that they are still valid. From Lemma 3.4.1 we know that $\Delta(\mathscr{M}) \subseteq \Delta(\mathscr{M}, i)$ for every i with probability at least $1-\delta$. Therefore, by choosing approximate occupancy measures we can only improve the regret so the arguments are indeed valid.

Using the exact form of $\tilde{q}^{k+1}$ and the fact that $e^x \geq 1+x$, we get that

$$\tilde{q}_h^{k+1}(s,a,s') \geq q_h^k(s,a,s') - \eta q_h^k(s,a,s')c_h^k(s,a),$$

and therefore

$$\begin{aligned}\sum_{k=1}^{K} \left\langle q^k - \tilde{q}^{k+1}, c^k \right\rangle &\leq \eta \sum_{k=1}^{K} \sum_{h,s,a,s'} q_h^k(s,a,s')c_h^k(s,a)^2 \\ &\leq \eta \sum_{k=1}^{K} \sum_{h,s,a,s'} q_h^k(s,a,s') = \eta HK.\end{aligned}$$

For the second term, $\mathrm{KL}(q \parallel q^1)/\eta$, we use the fact that the unnormalized KL divergence is the Bregman divergence associated with the unnormalized negative entropy, defined as follows:

$$R(q) = \sum_{h,s,a,s'} q_h(s,a,s') \log q_h(s,a,s') - q_h(s,a,s').$$

Now from standard arguments we obtain

$$\begin{aligned}\mathrm{KL}(q \parallel q^1) &\leq R(q) - R(q^1) \\ &\leq \sum_{h=1}^{H} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h^1(s,a,s') \log \frac{1}{q_h^1(s,a,s')} \\ &\leq H \log(HS^2A).\end{aligned}$$

Putting these two bounds together completes the proof. □

4 Learning Adversarial MDPs with Unknown Transition Function and Bandit Feedback

This chapter is based on:

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems, NeurIPS 2019*.

This chapter presents the first regret bounds for adversarial MDPs with unknown transition function and bandit feedback.

4.1 The O-REPS Algorithm for Bandit Feedback

Recall that under bandit feedback the agent observes only the costs that she suffers, as opposed to full-information feedback in which the entire cost function is revealed in the end of the episode. Formally, in the end of episode k, instead of observing $\{c_h^k(s,a)\}_{(s,a,h)\in\mathscr{S}\times\mathscr{A}\times[H]}$ we only observe $\{c_h^k(s_h^k,a_h^k)\}_{h\in[H]}$.

We start this chapter by reviewing the bandit version of the O-REPS algorithm when the transition function is known to the learner. The algorithm is almost the same as the one presented in the previous chapter, but with a single modification that we shall now describe.

Notice that O-REPS uses the cost function c^k when solving the optimization problem in Equation (3.1). Now, that not the entire cost function is known, [ZN13] propose to estimate it with a standard importance sampling estimator, defined by:

$$\hat{c}^k_h(s,a) = \begin{cases} \frac{c^k_h(s,a)}{q^k_h(s,a)}, & \text{if } s^k_h = s \text{ and } a^k_h = a \\ 0, & \text{otherwise} \end{cases}.$$

They show that this estimator is unbiased. Moreover, they prove that combining O-REPS with this importance sampling estimator yields optimal regret (up to logarithmic factors) of $\widetilde{O}(H\sqrt{SAK})$.

4.2 Our Algorithms for Bandit Feedback with Unknown Transitions

We define $\beta(\mathcal{M})$ as the minimum probability to visit some state under the worst exploratory policy, i.e., $\beta(\mathcal{M}) = \min_\pi \min_{h\in[H]} \min_{s\in\mathcal{S}} q^{P,\pi}_h(s)$. Moreover, we define $p_{min}(\mathcal{M})$ as the minimal transition probability, that is, $p_{min}(\mathcal{M}) = \min_{h,s,a,s'} P_h(s' \mid s,a)$.

Our first algorithm, "Bounded Bandit UC-O-REPS", is aimed for MDPs where there is a known positive lower bound on $\beta(\mathcal{M})$. Our second algorithm, "Shifted Bandit UC-O-REPS", works in general episodic adversarial MDPs and makes use of the first algorithm.

4.2.1 Bounded Bandit UC-O-REPS

The "Bounded Bandit UC-O-REPS" algorithm runs UC-O-REPS but with two crucial changes.

Firstly, instead of using c^k (which we do not have) we use $\hat{c}^k$ which is our estimate of c^k defined as follows:

$$\hat{c}^k_h(x,a) = \begin{cases} \frac{c^k_h(x,a)}{q^k_h(s,a)}, & \text{if } s^k_h = s \text{ and } a^k_h = a \\ 0, & \text{otherwise} \end{cases}.$$

Notice that this is a biased estimator since P^k may be different from P,

$$\begin{aligned}\mathbb{E}^k\left[\hat{c}_h^k(s,a)\right] &= q_h^{P,\pi^k}(s,a)\frac{c_h^k(s,a)}{q_h^{P^k,\pi^k}(s,a)} \\ &= q_h^{P,\pi^k}(s)\pi_h^k(a \mid s)\frac{c_h^k(s,a)}{q^{P^k,\pi^k}(s)\pi_h^k(a \mid s)} \\ &= q_h^{P,\pi^k}(s)\frac{c_h^k(s,a)}{q_h^{P^k,\pi^k}(s)},\end{aligned} \tag{4.1}$$

where the notation $\mathbb{E}^k[\cdot]$ means that we take the expectation conditioning on every thing that happened before the beginning of episode k.

Secondly, because of the bandit feedback we want to ensure that our algorithm performs enough exploration. For this purpose we constrain the confidence sets to contain only occupancy measures that visit every state with probability of at least α, where $0 < \alpha < 1$ is a parameter. That is, we define our confidence set for epoch i as $\Delta_\alpha(\mathscr{M}, i) = \Delta(\mathscr{M}, i) \cap \{q : q_h(s) \geq \alpha \quad \forall h, s\}$.

Thus our algorithm performs the following steps in each episode:

$$\begin{aligned}\tilde{q}^{k+1} &= \arg\min_q \eta \langle q, \hat{c}^k\rangle + \mathrm{KL}(q \parallel q^k) \\ q^{t+1} &= q^{P^{k+1},\pi^{k+1}} = \arg\min_{q\in\Delta_\alpha(\mathscr{M},i(k))} \mathrm{KL}(q \parallel \tilde{q}^{k+1}).\end{aligned}$$

If $\Delta_\alpha(\mathscr{M}, i(t)) = \emptyset$, then q^{k+1} is chosen to be an arbitrary occupancy measure. The efficient implementation of this algorithm is similar to the one of the original UC-O-REPS algorithm, and is described in details in the supplementary material (together with full pseudo-code).

4.2.2 Shifted Bandit UC-O-REPS

The "Shifted Bandit UC-O-REPS" algorithm runs "Bounded Bandit UC-O-REPS" with $\alpha = \frac{\rho}{S}$ (where $0 < \rho < 1$ is a parameter) but it makes the following change in order to handle the unknown $\beta(\mathscr{M})$ (which may be zero). It shifts the confidence sets by changing the empirical transition function. That is, instead of using $\bar{P}^i$ as the empirical transition function for epoch i it uses $\widetilde{P}^i$ which is defined as follows for every $h = 1, \ldots, H$ and for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$:

$$\widetilde{P}_h^i(s' \mid s,a) = (1-\rho)\bar{P}_h^i(s' \mid s,a) + \frac{\rho}{S}.$$

To sum up, the new confidence sets are denoted as $\widetilde{\Delta}_\alpha(\mathscr{M},i)$ and they contain all occupancy measures $q^{P',\pi}$ such that $q_h^{P',\pi}(s) \geq \alpha$ for every (h,s), and for every (h,s,a),

$$\|P_h'(\cdot \mid s,a) - \widetilde{P}_h^i(\cdot \mid s,a)\|_1 \leq \varepsilon_h^i(s,a).$$

Clearly this algorithm can be implemented efficiently, given the efficient implementation of "Bounded Bandit UC-O-REPS" (full pseudo-code can be found in the supplementary material for completeness).

4.3 Regret Analysis - Bounded Bandit UC-O-REPS

In this case we assume that $\beta(\mathscr{M}) > 0$ and it is known to the learner (or some positive lower bound on it). This assumption is quite strong but it holds if, for example, the minimum transition probability is not zero, i.e., $p_{min}(\mathscr{M}) > 0$. In this case $\beta(\mathscr{M}) \geq p_{min}(\mathscr{M})$.

Notice that if we run "Bounded Bandit UC-O-REPS" with $\alpha = \beta(\mathscr{M})$, then $\Delta(\mathscr{M}) = \Delta_\alpha(\mathscr{M}) \overset{def}{=} \Delta(\mathscr{M}) \cap \{q : q_h(s) \geq \alpha \quad \forall h,s\}$. Therefore, using the proof of UC-O-REPS, we have that all the confidence sets contain $\Delta(\mathscr{M})$ with probability at least $1-\delta$.

Let $q \in \Delta(\mathscr{M}) = \Delta_\alpha(\mathscr{M})$, and partition the regret into two terms as follows,

$$R_K = \sum_{k=1}^{K} \langle q^{P,\pi^k} - q, c^k \rangle = \left(\sum_{k=1}^{K} \langle q^{P,\pi^k} - q^{P^k,\pi^k}, c^k \rangle \right) + \left(\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, c^k \rangle \right).$$

The first term includes the error that comes from the estimation of the unknown transition function and will be denoted as $\hat{R}_{1:K}^{APP}$. The second term includes the error that comes from choosing sub-optimal policies and will be denoted as $\hat{R}_{1:K}^{ON}$.

Sections 4.3.1 and 4.3.2 bound these two terms and give us the following regret bound.

Theorem 4.3.1. *Let $\mathscr{M} = \left(\mathscr{S}, \mathscr{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP, and assume that $\beta(\mathscr{M}) > 0$. Then, "Bounded Bandit UC-O-REPS" with $\alpha = \beta(\mathscr{M})$ obtains the following regret bound:*

$$\mathbb{E}[R_K] \leq O\left(\frac{H^2 S \sqrt{AK} \log(KHSA)}{\beta(\mathscr{M})} \right).$$

4.3.1 Bounding $\hat{R}^{APP}_{1:K}$

Recall that $\hat{R}^{APP}_{1:K}$ is the difference between the loss of the learner's chosen policies in $\mathscr{M}$ and the loss of these policies in the "optimistic" MDPs (the ones induced by the occupancy measures q^k). The algorithm minimizes this difference through shrinking of the confidence sets. Notice that:

$$\hat{R}^{APP}_{1:K} = \sum_{k=1}^{K} \langle q^{P,\pi^k} - q^{P^k,\pi^k}, c^k \rangle \leq \sum_{k=1}^{K} \|q^{P,\pi^k} - q^{P^k,\pi^k}\|_1 \|c^k\|_\infty \leq \sum_{k=1}^{K} \|q^{P,\pi^k} - q^{P^k,\pi^k}\|_1.$$

Since the algorithm uses the same framework of confidence sets as the original UC-O-REPS (and all the confidence sets contain $\Delta(\mathscr{M})$ with high probability), we can use the following theorem from [RM19a] to bound this difference.

Theorem 4.3.2. *Let $\{\pi^k\}_{k=1}^K$ be policies and let $\{P^k\}_{k=1}^K$ be transition functions such that $q^{P^k,\pi^k} \in \Delta(\mathscr{M}, i(k))$ for every k. Then, when setting $\delta = \frac{1}{K}$:*

$$\mathbb{E}\left[\hat{R}^{APP}_{1:K}\right] \leq \mathbb{E}\left[\sum_{k=1}^{K} \|q^{P,\pi^k} - q^{P^k,\pi^k}\|_1\right] \leq O\left(H^2 S\sqrt{AK\log(KHSA)}\right).$$

4.3.2 Bounding $\hat{R}^{ON}_{1:K}$

Recall that $\hat{R}^{ON}_{1:K}$ is the regret for the performance of the online algorithm's chosen occupancy measures. Notice that the learner performs the original UC-O-REPS algorithm with respect to the sequence of loss functions $\{\hat{c}^k\}_{k=1}^K$ and the set of occupancy measures $\Delta_\alpha(\mathscr{M})$. Therefore, we can use the regret analysis of the original algorithm to obtain the following result (full proof in the supplementary material).

Lemma 4.3.3. *Let $\mathscr{M} = \left(\mathscr{S}, \mathscr{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP. Then, for every $q \in \Delta_\alpha(\mathscr{M})$, "Bounded Bandit UC-O-REPS" obtains:*

$$\mathbb{E}\left[\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, \hat{c}^k \rangle\right] \leq O\left(\frac{\eta HAK}{\alpha} + \frac{H\log(HSA)}{\eta}\right).$$

Now we show that the sequence of occupancy measures chosen by the algorithm performs similarly on $\{\hat{c}^k\}_{k=1}^K$ and $\{c^k\}_{k=1}^K$ in expectation, and therefore we can derive a bound on $\hat{R}^{ON}_{1:K}$.

Lemma 4.3.4. *Let $\mathscr{M} = \left(\mathscr{S}, \mathscr{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP. Then, for every $q \in \Delta_\alpha(\mathscr{M})$, "Bounded Bandit UC-O-REPS" obtains:*

$$\left|\mathbb{E}\left[\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, \hat{c}^k\rangle\right] - \mathbb{E}\left[\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, c^k\rangle\right]\right| \le O\left(\frac{H^2 S\sqrt{AK}\log(KHSA)}{\alpha}\right).$$

Proof. First we use the linearity of expectation and the fact that $q^k = q^{P^k,\pi^k}$ to obtain:

$$\left|\mathbb{E}\left[\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, \hat{c}^k\rangle\right] - \mathbb{E}\left[\sum_{k=1}^{K} \langle q^{P^k,\pi^k} - q, c^k\rangle\right]\right| = \left|\mathbb{E}\left[\sum_{k=1}^{K} \langle q^k - q, \hat{c}^k - c^k\rangle\right]\right|.$$

From the law of total expectation we have,

$$\left|\mathbb{E}\left[\sum_{k=1}^{K} \langle q^k - q, \hat{c}^k - c^k\rangle\right]\right| = \left|\mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}^k\left[\langle q^k - q, \hat{c}^k - c^k\rangle\right]\right]\right|. \tag{4.2}$$

Now for every k we can use the definition of $\hat{c}^k$ and (4.1) to obtain,

$$\mathbb{E}^k\left[\langle q^k - q, \hat{c}^k - c^k\rangle\right] = \sum_{h,s,a}\left(q_h^k(s,a) - q_h(s,a)\right)\left(q_h^{P,\pi^k}(s)\frac{c_h^k(s,a)}{q_h^{P^k,\pi^k}(s)} - c_h^k(s,a)\right).$$

Substituting this back into (4.2) we get,

$$\begin{aligned}
\left|\mathbb{E}\Big[\sum_{k=1}^{K} \langle q^k - q, \hat{c}^k - c^k\rangle\Big]\right| &= \left|\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h,s,a}\left(q_h^k(s,a) - q_h(s,a)\right)\left(q_h^{P,\pi^k}(s)\frac{c_h^k(s,a)}{q_h^{P^k,\pi^k}(s)} - c_h^k(s,a)\right)\right]\right| \\
&\le \mathbb{E}\left[\left|\sum_{k=1}^{K}\sum_{h,s,a} c_h^k(s,a)\left(q_h^k(s,a) - q_h(s,a)\right)\frac{q_h^{P,\pi^k}(s) - q_h^{P^k,\pi^k}(s)}{q_h^{P^k,\pi^k}(s)}\right|\right] \\
&\le \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h,s}\frac{|q_h^{P,\pi^k}(s) - q_h^{P^k,\pi^k}(s)|}{q_h^{P^k,\pi^k}(s)}\left|\sum_{a} c_h^k(s,a)\left(q_h^k(s,a) - q_h(s,a)\right)\right|\right] \\
&\le \frac{1}{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h,s}|q^{P,\pi^k}(x) - q^{P^k,\pi^k}(x)|\right],
\end{aligned}$$

where the last inequality follows because $q_h^{P^k,\pi^k}(s) \ge \alpha$, $0 \le \sum_a q_h^k(s,a)c_h^k(s,a) \le 1$ and

$0 \leq \sum_a q_h(s,a)c^k(x,a) \leq 1$. Finally, we use Theorem 4.3.2 to conclude that

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{k=1}^{K}\langle q^k - q, \hat{c}^k - c^k\rangle\right]\right| &\leq \frac{1}{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h,s}\left|\sum_{a,s'} q_h^{P,\pi^k}(s,a,s') - q_h^{P^k,\pi^k}(s,a,s')\right|\right] \\
&\leq \frac{1}{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h,s,a,s'}|q_h^{P,\pi^k}(s,a,s') - q_h^{P^k,\pi^k}(s,a,s')|\right] \\
&= \frac{1}{\alpha}\mathbb{E}\left[\sum_{k=1}^{K}\|q^{P,\pi^k} - q^{P^k,\pi^k}\|_1\right] \leq O\left(\frac{H^2 S\sqrt{AK}\log(KHSA)}{\alpha}\right).
\end{aligned}$$

□

Corollary 4.3.5. *Let $\mathscr{M} = \left(\mathscr{S},\mathscr{A},H,P,\{c^k\}_{k=1}^K\right)$ be an adversarial MDP. Then, when setting $\eta = \sqrt{\frac{\log(KHSA)}{AK}}$ and $\delta = \frac{1}{K}$, "Bounded Bandit UC-O-REPS" obtains:*

$$\mathbb{E}\left[\hat{R}_{1:K}^{ON}\right] \leq O\left(\frac{H^2 S\sqrt{AK}\log(KHSA)}{\alpha}\right).$$

4.4 Regret Analysis - Shifted Bandit UC-O-REPS

We remove the assumption that $\beta(\mathscr{M}) > 0$, and for this case will use the "Shifted Bandit UC-O-REPS" algorithm. Notice that the key insight for the regret analysis of "Bounded Bandit UC-O-REPS" is that by setting $\alpha = \beta(\mathscr{M})$, we get that all the confidence sets contain $\Delta(\mathscr{M})$ with high probability. The idea behind "Shifted Bandit UC-O-REPS" is to work on an imaginary MDP $\widetilde{\mathscr{M}}$ that is close to $\mathscr{M}$ but has the property $\beta(\widetilde{\mathscr{M}}) > 0$.

The transition function for the MDP $\widetilde{\mathscr{M}} = \left(\mathscr{S},\mathscr{A},H,p,\{c^k\}_{k=1}^K\right)$ is defined as follows for every $h = 1,\dots,H$ and for every $(s,a,s') \in \mathscr{S}\times\mathscr{A}\times\mathscr{S}$:

$$p_h(s' \mid s,a) = (1-\rho)P_h(s' \mid s,a) + \frac{\rho}{S}.$$

This means that the minimal transition probability is positive, i.e., $p_{min}(\widetilde{\mathscr{M}}) \geq \frac{\rho}{S} > 0$. Therefore, $\beta(\widetilde{\mathscr{M}}) \geq \frac{\rho}{S} > 0$ and we can run "Bounded Bandit UC-O-REPS" on $\widetilde{\mathscr{M}}$. The problem is that our data is sampled from $\mathscr{M}$, but we need to build confidence sets that contain $\Delta(\widetilde{\mathscr{M}})$ and not $\Delta(\mathscr{M})$. The following lemma shows that shifting the confidence sets obtains this desired property: all the confidence sets contain $\Delta(\widetilde{\mathscr{M}})$ with probability at least $1-\delta$.

Lemma 4.4.1. *If $\Delta(\mathscr{M}) \subseteq \Delta(\mathscr{M},i)$, then $\Delta(\widetilde{\mathscr{M}}) \subseteq \widetilde{\Delta}_\alpha(\mathscr{M},i)$.*

Proof. Let $q^{p,\pi} \in \Delta(\widetilde{\mathscr{M}})$. First of all, since $\beta(\widetilde{\mathscr{M}}) \geq \frac{\rho}{S} = \alpha$ we have that $q_h^{p,\pi}(s) \geq \alpha$ for every h, s. Now, Since $\Delta(\mathscr{M}) \subseteq \Delta(\mathscr{M}, i)$ we have that for every (h, s, a),

$$\|\bar{P}_h^i(\cdot \mid s,a) - P_h(\cdot \mid s,a)\|_1 \leq \varepsilon_h^i(s,a).$$

By the definition of $\bar{P}_i^\star$ and $P^\star$ we have that,

$$\begin{aligned}
\|\widetilde{P}_h^i(\cdot \mid s,a) - p_h(\cdot \mid s,a)\|_1 &= \sum_{s'} |\widetilde{P}_h^i(s' \mid s,a) - p_h(s' \mid s,a)| \\
&= \sum_{s'} |(1-\rho)\bar{P}_h^i(s' \mid s,a) + \frac{\rho}{S} - (1-\rho)P_h(s' \mid s,a) - \frac{\rho}{S}| \\
&= (1-\rho)\sum_{s'} |\bar{P}_h^i(s' \mid s,a) - P_h(s' \mid s,a)| \\
&= (1-\rho)\|\bar{P}_h^i(\cdot \mid s,a) - P_h(\cdot \mid s,a)\|_1 \leq \varepsilon_h^i(s,a),
\end{aligned}$$

and therefore $q^{p,\pi} \in \widetilde{\Delta}_\alpha(\mathscr{M}, i)$ and $\Delta(\widetilde{\mathscr{M}}) \subseteq \widetilde{\Delta}_\alpha(\mathscr{M}, i)$. □

Now we divide the regret into two parts: the regret of "Bounded Bandit UC-O-REPS" in $\widetilde{\mathscr{M}}$ and the difference in the performance of policies in $\mathscr{M}$ and $\widetilde{\mathscr{M}}$. Formally, the regret of any $q = q^{P,\pi} \in \Delta(\mathscr{M})$ is partitioned as follows:

$$\begin{aligned}
R_K &= \sum_{k=1}^K \langle q^{P,\pi^k} - q^{P,\pi}, c^k \rangle \\
&= \left(\sum_{k=1}^K \langle q^{P,\pi^k} - q^{p,\pi^k}, c^k \rangle\right) + \left(\sum_{k=1}^K \langle q^{p,\pi^k} - q^{p,\pi}, c^k \rangle\right) + \left(\sum_{k=1}^K \langle q^{p,\pi} - q^{P,\pi}, c^k \rangle\right).
\end{aligned}$$

Since $\|P_h(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq 2\rho$ for every (h, s, a), we can use [RM19b, Corollary E.2] to bound the first and third terms as $O(\rho H^2 K)$. The second term includes the regret of "Bounded Bandit UC-O-REPS" in $\widetilde{\mathscr{M}}$ so according to Theorem 4.3.1 we can bound it as $O\left(\frac{H^2 S \sqrt{AK \log(KHSA)}}{\rho/S}\right) = O\left(\frac{H^2 S^2 \sqrt{AK \log(KHSA)}}{\rho}\right)$. Thus we get the following regret bound.

Theorem 4.4.2. *Let $\mathscr{M} = \left(\mathscr{S}, \mathscr{A}, H, P, \{c^k\}_{k=1}^K\right)$ be an adversarial MDP. Then, "Shifted Bandit UC-O-REPS" with $\rho = S\sqrt[4]{\frac{A \log(KHSA)}{K}}$ obtains the following regret bound,*

$$\mathbb{E}[R_K] \leq O\left(H^2 S A^{1/4} K^{3/4} \log^{1/4}(KHSA)\right).$$

5 Near-Optimal Regret for Stochastic Shortest Path

This chapter is based on:

Aviv Rosenberg, Alon Cohen, Yishay Mansour and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020.*

This chapter presents the first lower bound and the first near-optimal regret bound for stochastic shortest path (SSP). We start the chapter by describing the regret minimization formulation in SSP, then we present the main results, and finally dive deeper into our algorithms and lower bound.

5.1 Regret Minimization in SSP

For simplicity, in this chapter we assume that the costs are deterministic and known to the learner. However, the transition probabilities P are fixed but unknown to the learner. The learner interacts with the model in episodes: each episode starts at the initial state s_{init}, and ends when the learner reaches the goal state g (note that she might *never* reach the goal state). Success is measured by the learner's regret over K such episodes, that is the difference between her total cost over the K episodes and the total expected cost of the optimal proper policy:

$$R_K = \sum_{k=1}^{K} \sum_{i=1}^{I^k} c(s_i^k, a_i^k) - K \cdot \min_{\pi \in \Pi_{\text{proper}}} V^{\pi}(s_{\text{init}}),$$

where I^k is the time it takes the learner to complete episode k (which may be infinite), Π_{proper} is the set of all stationary, deterministic and proper policies (that is not empty by

Assumption 2.2.1), and (s_i^k, a_i^k) is the i-th state-action pair at episode k. In the case that I^k is infinite for some k, we define $R_K = \infty$.

We denote the optimal proper policy by $\pi^\star$, i.e., $V^\star(s) = \arg\min_{\pi \in \Pi_{\text{proper}}} V^\pi(s)$ for all $s \in \mathscr{S}$. Moreover, let $B_\star > 0$ be an upper bound on the values of $V^\star$ and let $T_\star > 0$ be an upper bound on the times $T^{\pi^\star}$, i.e., $B_\star \geq \max_{s \in \mathscr{S}} V^\star(s)$ and $T_\star \geq \max_{s \in \mathscr{S}} T^{\pi^\star}(s)$.

5.2 Summary of Results

In Section 5.3 we present our Hoeffding-based algorithms (Algorithms 6 and 14) and their analysis. While they achieve similar regret bounds to [TGV$^+$20], their presentation is important in order to lay the foundations for our Bernstein-based algorithm (Algorithm 7) and its improved regret bound shown in Section 5.4. Finally, in Section 5.5 we give a lower bound on the learner's regret showing that Algorithm 7 is near-optimal.

The learner must reach the goal state otherwise she has infinite regret. Therefore, she has to trade-off two objectives, one is to reach the goal state and the other is to minimize the cost. Under the following assumption, the two objectives essentially coincide.

Assumption 5.2.1. *All costs are positive, i.e., there exists $c_{min} > 0$ such that $c(s,a) \geq c_{min}$ for every $(s,a) \in \mathscr{S} \times \mathscr{A}$.*

This assumption allows us to upper bound the running time of the algorithm by its total cost up to a factor of c_{min}^{-1}. In particular, it guarantees that any policy that does not reach the goal state has infinite cost, so any bounded regret algorithm has to reach the goal state. We eventually relax Assumption 5.2.1 by a technique similar to that of [BY13]. We add a small positive perturbation to the instantaneous costs and run our algorithms on the model with the perturbed costs. This provides a regret bound that scales with the expected running time of the optimal policy.

We now summarize our results. For ease of comparison, we first present our regret bounds for both the Hoeffding and Bernstein-based algorithms when Assumption 5.2.1 holds, and subsequently show the regret bounds of both algorithms for the general case. In order to simplify the presentation of our results, we assume that $S \geq 2, A \geq 2$ and $K \geq S^2 A$ throughout. In addition, we denote $L = \log(K B_\star SA/\delta c_{\text{min}})$. The complete proofs of all statements are found in Appendix C.

Positive costs. The following results hold when Assumption 5.2.1 holds (recall that we always assume Assumption 2.2.1). In particular, when this assumption holds the optimal policy overall is proper (Theorem 2.2.3) hence the regret bounds below are with respect to the best overall policy.

Theorem 5.2.2. *Suppose that Assumption 5.2.1 holds. With probability at least* $1-\delta$ *the regret of Algorithm 14 is bounded as follows:*

$$R_K = O\left(\sqrt{\frac{B_\star^3 S^2 AK}{c_{min}}}L + \frac{B_\star^3 S^2 A}{c_{min}^2} L^2\right).$$

The main issue with the regret bound in Theorem 5.2.2 is that it scales with $\sqrt{K/c_{\min}}$ which cannot be avoided regardless of how large K is with respect to $c_{\min}^{-1}$. This problem is alleviated in Algorithm 7 that uses the tighter Bernstein-based confidence bounds.

Theorem 5.2.3. *Assume that Assumption 5.2.1 holds. With probability at least* $1-\delta$ *the regret of Algorithm 7 is bounded as follows:*

$$R_K = O\left(B_\star S\sqrt{AK}L + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{min}}}L^2\right).$$

Note that when $K \gg B_\star S^2 A/c_{\min}$, the regret bound above scales as $\widetilde{O}(B_\star S\sqrt{AK})$ thus obtaining a near-optimal rate.

Arbitrary costs (i.e., $c(s,a) \in [0,1]$). Recall that in this case we can no longer assume that the optimal policy is proper. Therefore, the regret bounds below are with comparison to the best *proper* policy. Assumption 5.2.1 can be easily alleviated by adding a small fixed cost to the cost of all state-action pairs. Following the perturbation of the costs, we obtain regret bounds from Theorems 5.2.2 and 5.2.3 with $c_{\min} \leftarrow \varepsilon$ and $B_\star \leftarrow B_\star + \varepsilon T_\star$, and the learner also suffers an additional cost of $\varepsilon T_\star K$ due to the misspecification of the model caused by the perturbation. By picking ε to balance these terms we get the following corollaries (letting $\widetilde{L} = \log(KB_\star T_\star SA/\delta)$).

Corollary 5.2.4. *Running Algorithm 14 using costs* $c_\varepsilon(s,a) = \max\{c(s,a),\varepsilon\}$ *defined for* $\varepsilon = (S^2A/K)^{1/3}$ *gives the following regret bound with probability at least* $1-\delta$*:*

$$R_K = O\left(T_\star^3 S^{2/3} A^{1/3} K^{2/3} \widetilde{L} + T_\star^3 S^2 A \widetilde{L}^2\right).$$

Corollary 5.2.5. *Running Algorithm 7 using costs $c_\varepsilon(s,a) = \max\{c(s,a),\varepsilon\}$ defined for $\varepsilon = S^2A/K$ gives the following regret bound with probability at least $1-\delta$:*

$$R_K = O\left(B_\star^{3/2} S\sqrt{AK}\widetilde{L} + T_\star^{3/2} S^2 A \widetilde{L}^2 \right).$$

Moreover, when the algorithm knows $B_\star$ and $K \gg S^2AT_\star^2$, then choosing $\varepsilon = B_\star S^2A/K$ gives a near-optimal regret bound of $\widetilde{O}(B_\star S\sqrt{AK})$.

Lower bound. In Section 5.5 we show that Corollary 5.2.5 is nearly-tight using the following theorem.

Theorem 5.2.6. *There exists an SSP problem instance $\mathscr{M} = (\mathscr{S},\mathscr{A},P,c,s_{init},g)$ in which $V^\star(s) \le B_\star$ for all $s \in \mathscr{S}$, $S \ge 2$, $A \ge 16$, $B_\star \ge 2$, $K \ge SA$, and $c(s,a) = 1$ for all $s \in \mathscr{S}, a \in \mathscr{A}$, such the expected regret of any learner after K episodes satisfies*

$$\mathbb{E}[R_K] \ge \frac{1}{1024} B_\star \sqrt{SAK}.$$

5.3 Hoeffding-type Confidence Bounds

We start with a simpler case in which $B_\star$ is known to the learner. In Section 5.3.2 we alleviate this assumption with a penalty of an additional log-factor in the regret bound. For now, we prove the following bound on the learner's regret.

Theorem 5.3.1. *Suppose that Assumption 5.2.1 holds. With probability at least $1-\delta$ the regret of Algorithm 6 is bounded as follows:*

$$R_K = O\left(\sqrt{\frac{B_\star^3 S^2 AK}{c_{min}}} L + \frac{B_\star^3 S^2 A}{c_{min}^2} L^{3/2} \right).$$

Our algorithm follows the known concept of optimism in face of uncertainty. That is, it maintains confidence sets that contain the true transition function with high probability and picks an optimistic optimal policy—a policy that minimizes the expected cost over all policies and all transition functions in the current confidence set. The computation of the optimistic optimal policy can be done efficiently using Extended Value Iteration as shown by [TGV$^+$20]. Construct an augmented MDP whose states are S and its action set consists of tuples $(a,\widetilde{P})$ where $a \in \mathscr{A}$ and $\widetilde{P}$ is any transition function such that

Algorithm 6 HOEFFDING-TYPE CONFIDENCE BOUNDS AND KNOWN $B_\star$

```
input: state space $\mathscr{S}$, action space $\mathscr{A}$, bound on cost-to-go of optimal policy $B_\star$, confidence parameter $\delta$.
initialization: $\forall(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} : N(s,a,s') \leftarrow 0, N(s,a) \leftarrow 0$, an arbitrary policy $\tilde{\pi}$, $t \leftarrow 1$.
for $k = 1, 2, \dots$ do
    set $s_t \leftarrow s_{\text{init}}$.
    while $s_t \neq g$ do
        follow optimistic optimal policy: $a_t \leftarrow \tilde{\pi}(s_t)$.
        observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
        update: $N(s_t,a_t,s_{t+1}) \leftarrow N(s_t,a_t,s_{t+1}) + 1$, $N(s_t,a_t) \leftarrow N(s_t,a_t) + 1$.
        if $N(s_{t+1}, \tilde{\pi}(s_{t+1})) \leq \frac{5000 B_\star^2 S}{c_{\min}^2} \log \frac{B_\star SA}{\delta c_{\min}}$ or $s_{t+1} = g$ then
            # start new interval
            compute empirical transition function $\bar{P}$ as $\bar{P}(s'|s,a) = N(s,a,s')/N_+(s,a)$ where $N_+(s,a) = \max\{N(s,a), 1\}$.
            compute optimistic policy $\tilde{\pi}$ by minimizing expected cost over transition functions $\widetilde{P}$ that satisfy Equation (5.1).
        end if
        set $t \leftarrow t + 1$.
    end while
end for
```

$$\left\|\widetilde{P}(\cdot|s,a) - \bar{P}(\cdot|s,a)\right\|_1 \leq 5\sqrt{\frac{S \log(SAN_+(s,a)/\delta)}{N_+(s,a)}} \tag{5.1}$$

where $\bar{P}$ is the empirical estimate of P. It can be shown that the optimistic policy and the optimistic model, i.e., those that minimize the expected total cost over all policies and feasible transition functions, correspond to the optimal policy of the augmented MDP.

To ensure that the algorithm reaches the goal state in every episode, we define a state-action pair (s,a) as *known* if the number of visits to this pair is at least $\frac{5000 B_\star^2 S}{c_{\min}^2} \log \frac{B_\star SA}{\delta c_{\min}}$ and as *unknown* otherwise. We show with high probability the optimistic policy chosen by the algorithm will be proper once all state-action pairs are known. However, when some pairs are still unknown, our chosen policies may be improper. This implies that the strategy of keeping the policy fixed throughout an episode, as done usually in episodic RL, will fail. Consequently, our algorithm changes policies at the start of every episode and also every time we reach an unknown state-action pair.

Formally, we split the time into *intervals*. The first interval begins at the first time step, and every interval ends by reaching the goal state or a state s such that $(s, \tilde{\pi}(s))$ is unknown (where $\tilde{\pi}$ is the current policy followed by the learner). Recall that once all state-action pairs are known, the optimistic policy will eventually reach the goal state. Therefore, recomputing the optimistic policy at the end of every interval ensures that the algorithm will eventually reach the goal state with high probability. Note that the total number of intervals is at most the number of visits to an unknown state-action pair plus the number of episodes.

Observation 5.3.2. *The total number of intervals, M, is*

$$O\left(K + \frac{B_\star^2 S^2 A}{c_{min}^2} \log \frac{B_\star SA}{\delta c_{min}}\right).$$

5.3.1 Analysis

The proof of Theorem 5.3.1 begins by defining the "good event" in which our confidence sets contain the true transition function and the total cost in every interval is bounded. This in turn implies that all episodes end in finite time. We prove that the good event holds with high probability.

Then, independently, we give a high-probability bound on the regret of the algorithm when the good event holds. To do so, recall that at the beginning of every interval m, the learner computes an optimistic policy by minimizing over all policies and over all transition functions within the current confidence set. We denote the chosen policy by $\tilde{\pi}^m$ and let $\widetilde{P}_m$ be the minimizing transition function (i.e., the optimistic model). A key observation is that by the definition of our confidence sets, $\widetilde{P}_m$ is such that there is always some positive probability to transition to the goal state directly from any state-action. This implies that all policies are proper in the optimistic model and that the cost-to-go function of $\tilde{\pi}^m$ defined with respect to $\widetilde{P}_m$, and denoted by $\widetilde{V}^m$, is finite. By Theorem 2.2.2, the following Bellman optimality equations hold for all $s \in \mathscr{S}$,

$$\widetilde{V}^m(s) = \min_{a \in \mathscr{A}} c(s,a) + \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s,a) \widetilde{V}^m(s'). \tag{5.2}$$

High probability events. For every interval m, we let Ω^m denote the event that the confidence set for interval m contains the true transition function P. Formally, let $\bar{P}_m$ denote the empirical estimate of the transition function at the beginning of interval m, let $N_m(s,a)$ denote the number of visits to state-action pair (s,a) up to interval m (not including), and let $n_m(s,a)$ be the number of visits to (s,a) during interval m. Then we say that Ω^m holds if for all $(s,a) \in \mathscr{S} \times \mathscr{A}$, we have $(N_+^m(s,a) = \max\{1, N_m(s,a)\})$

$$\|P(\cdot|s,a) - \bar{P}_m(\cdot|s,a)\|_1 \leq 5\sqrt{\frac{S\log\big(SAN_+^m(s,a)/\delta\big)}{N_+^m(s,a)}}. \tag{5.3}$$

In the following lemma we show that, with high probability, the events Ω^m hold and that the total cost in each interval is bounded. Combining this with Observation 5.3.2 we get that all episodes terminate within a finite number of steps, with high probability.

Lemma 5.3.3. *With probability at least* $1-\delta/2$*, for all intervals m simultaneously, we have that* Ω^m *holds and that* $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \leq 24B_\star \log\frac{4m}{\delta}$*, where* H^m *denotes the length of interval m,* s_h^m *is the observed state at time h of interval m and* $a_h^m = \tilde{\pi}^m(s_h^m)$ *is the chosen action. This implies that the total number of steps of the algorithm is*

$$T = O\left(\frac{KB_\star}{c_{min}}L + \frac{B_\star^3 S^2 A}{c_{min}^3}L^2\right).$$

Proof sketch. The events Ω^m hold with high probability due to standard concentration inequalities, and thus it remains to address the high probability bound on the total cost within each interval.

This proof consists of three parts. In the first, we show that when Ω^m occurs we have that $\widetilde{V}^m(s) \leq V^\star(s) \leq B_\star$ for all $s \in \mathscr{S}$ due to the optimistic nature of the computation of $\tilde{\pi}^m$. In the second part, we postulate that had all state-action pairs been known, then having Ω^m hold implies that $V^m(s) \leq 2B_\star$ for all $s \in \mathscr{S}$. That is, when all state-action pairs are known, not only $\tilde{\pi}^m$ is proper in the true model, but its expected cumulative cost is at most $2B_\star$.

The third part of the proof deals with the general case when not all state-action pairs are known. Fix some interval m. Since the interval ends when we reach an unknown state-action, it must be that all but the first state-action pair visited during the interval are known. For this unknown first state-action pair, it follows from the Bellman equations (Equation (5.2)) and from $\widetilde{V}^m(s) \leq B_\star$ for all $s \in \mathscr{S}$ that $\tilde{\pi}^m$ never picks an action whose instantaneous cost is larger than $B_\star$. Therefore, the cost of this first unknown state-action pair is at most $B_\star$, and we focus on bounding the total cost in the remaining time steps with high probability.

To that end, we define the following modified MDP $M^{\text{know}} = (S^{\text{know}}, A, P^{\text{know}}, c, s_{\text{init}})$ in which every state $s \in \mathscr{S}$ such that $(s, \tilde{\pi}^m(s))$ is unknown is contracted to the goal state.

Let P^{know} be the transition function induced in M^{know} by P, and let V^m_{know} be the cost-to-go of $\tilde{\pi}^m$ in M^{know} w.r.t P^{know}. Similarly, define $\widetilde{P}^{\text{know}}_m$ as the transition function induced in M^{know} by $\widetilde{P}_m$, and $\widetilde{V}^m_{\text{know}}$ as the cost-to-go of $\tilde{\pi}^m$ in M^{know} w.r.t $\widetilde{P}^{\text{know}}_m$. It is clear that $\widetilde{V}^m_{\text{know}}(s) \le \widetilde{V}^m(s)$ for every $s \in \mathcal{S}$ from whence $\widetilde{V}^m_{\text{know}}(s) \le B_\star$. Moreover, since all states $s \in \mathcal{S}$ for which $(s, \tilde{\pi}^m(s))$ is unknown were contracted to the goal state, in M^{know} all remaining states-action pairs are known. Therefore, by the second part of the proof, $V^m_{\text{know}}(s) \le 2B_\star$ for all $s \in \mathcal{S}$. Note that reaching the goal state in M^{know} is equivalent to reaching either the goal state or an unknown state-action pair in the true model hence the latter argument shows that the total expected cost in doing so is at most $2B_\star$. We further obtain the high probability bound by a probabilistic amplification argument using the Markov property of the MDP. □

Regret analysis. In what follows, instead of bounding R_K, we bound

$$\widetilde{R}_K = \sum_{m=1}^{M} \sum_{h=1}^{H^m} c(s^m_h, a^m_h) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}}),$$

where $\mathbb{I}$ is the indicator function. Note that according to Theorem 5.3.3, we have that $\widetilde{R}_K = R_K$ with high probability.

The definition of $\widetilde{R}_K$ allows the analysis to disentangle two dependent probabilistic events. The first is the intersection of the events Ω^m which is dealt with in Theorem 5.3.3. The second holds when, for a fixed policy, the costs suffered by the learner do not deviate significantly from their expectation. In the following lemma we bound $\widetilde{R}_K$.

Lemma 5.3.4. *With probability at least* $1 - \delta/2$, *we have*

$$\widetilde{R}_K \le O\Bigg(\underbrace{\frac{B_\star^3 S^2 A}{c_{min}^2} \log \frac{B_\star SA}{c_{min}\delta}}_{(1)} + B_\star \sqrt{T \log \frac{T}{\delta}} + B_\star \sqrt{S \log \frac{SAT}{\delta}} \underbrace{\sum_{s,a} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N^m_+(s,a)}}}_{(2)} \Bigg).$$

Here we only explain how to interpret the resulting bound. The term (1) bounds the total cost spent in intervals that ended in unknown state-action pairs (it does not depend on K). The term (2) is at most $O(\sqrt{SAT})$ when Theorem 5.3.3 holds, and then the dominant term in Lemma 5.3.4 becomes $\widetilde{O}(B_\star S \sqrt{AT})$. Theorem 5.3.1 is finally obtained by applying a union bound on Lemma 5.3.3 and Lemma 5.3.4, and using Theorem 5.3.3 to bound T.

5.3.2 Unknown Cost Bound

In this section we relax the assumption that $B_\star$ is known to the learner. Instead, we keep an estimate $\widetilde{B}$ that is initialized to $c_{\min}$ and doubles every time the cost in interval m (denoted as C_m) reaches $24\widetilde{B}\log\frac{4m}{\delta}$. By Theorem 5.3.3, with high probability, $\widetilde{B} \leq 2B_\star$. We end an interval as before (once the goal state is reached or an unknown state-action pair is reached), but also when $\widetilde{B}$ is doubled. The algorithm for this case is presented in Appendix C (Algorithm 14). Since $\widetilde{B}$ changes, every state-action pair can become known once for every different value of $\widetilde{B}$.

Observation 5.3.5. *When $B_\star$ is unknown to the learner, the number of times a state-action pair can become known is at most $\log_2(B_\star/c_{min})$. The number of intervals M is*

$$O\left(K + \frac{B_\star^2 S^2 A}{c_{min}^2}\log^2\frac{B_\star SA}{\delta c_{min}}\right).$$

Lemma 5.3.6. *When $B_\star$ is unknown, with probability at least $1-\delta/2$, for all intervals m simultaneously, we have that Ω^m holds and that $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \leq 24B_\star\log\frac{4m}{\delta}$. This implies that the total number of steps of the algorithm is*

$$T = O\left(\frac{KB_\star}{c_{min}}L + \frac{B_\star^3 S^2 A}{c_{min}^3}L^3\right).$$

The analysis follows that of Algorithm 6. In particular, Lemma 5.3.4 still holds (with $2B_\star$ instead of $B_\star$), and jointly with Lemma 5.3.6 imply Theorem 5.2.2.

5.4 Bernstein-type Confidence Bounds

Algorithm 6 has two drawbacks. The first one is the use of Hoeffding-style confidence bounds which we improve with Bernstein-style confidence bounds. The second is the number of times the optimistic optimal policy is computed. In this section we propose to compute it in a way similar to UCRL2, i.e., once the number of visits to some state-action pair is doubled. Note that this change also eliminates the need to know or to estimate $B_\star$.

The algorithm is presented in Algorithm 7. It consists of *epochs*. The first epoch starts at the first time step, and each epoch ends once the number of visits to some state-action pair is doubled. An optimistic policy is computed at the end of every epoch using (empirical) Bernstein confidence bounds. In contrast to Algorithm 6, Algorithm 7 defines a

Algorithm 7 BERNSTEIN-TYPE CONFIDENCE BOUNDS

```
input: state space $\mathscr{S}$, action space $\mathscr{A}$ and confidence parameter $\delta$.
initialization: $i \leftarrow 1, t \leftarrow 1$, arbitrary policy $\tilde{\pi}_1$ , $\forall(s,a,s') : N_1(s,a,s') \leftarrow 0, N_1(s,a) \leftarrow 0$, $n_1(s,a,s') \leftarrow 0$, $n_1(s,a) \leftarrow 0$.
for $k = 1,2,\ldots$ do
  set $s_t \leftarrow s_{\text{init}}$.
  while $s_t \neq g$ do
    follow optimistic optimal policy: $a_t \leftarrow \tilde{\pi}_i(s_t)$.
    observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
    set: $n_i(s_t,a_t) \leftarrow n_i(s_t,a_t)+1$, $n_i(s_t,a_t,s_{t+1}) \leftarrow n_i(s_t,a_t,s_{t+1})+1$.
    if $n_i(s_{t+1},\tilde{\pi}_i(s_{t+1})) < N_i(s_{t+1},\tilde{\pi}_i(s_{t+1}))$ then
      set $t \leftarrow t+1$ and continue.
    end if
    # start new epoch
    set: $N_{i+1}(s,a,s') \leftarrow N_i(s,a,s') + n_i(s,a,s')$, $N_{i+1}(s,a) \leftarrow N_i(s,a) + n_i(s,a)$, $n_{i+1}(s,a) \leftarrow 0$, $n_{i+1}(s,a,s') \leftarrow 0$ for all $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$.
    compute empirical transition function $\bar{P}$ as $\bar{P}(s' \mid s,a) = N(s,a,s')/N_+(s,a)$ for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$ where $N_+(s,a) = \max\{N(s,a),1\}$.
    compute optimistic transition function $\widetilde{P}$ using Equation (5.4).
    compute optimal policy $\tilde{\pi}$ w.r.t $\widetilde{P}$.
    $i \leftarrow i+1, t \leftarrow t+1$.
  end while
end for
```

confidence range for each state, action, and next state, separately, around its empirical estimate (i.e., we use an L_∞ "ball" rather than an L_1 "ball" around the empirical estimates). This allows us to disentangle the computation of the optimistic policy from the computation of the optimistic model. Indeed, the computation of the optimistic model becomes very easy: one simply has to maximize the probability of transition directly to the goal state at every state-action pair which means minimizing the probability of transition to all other states and setting them at the lowest possible value of their confidence range. This results in the following formula for $\widetilde{P}(s' \mid s,a)$ for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$:

$$\max\{\bar{P}(s'|s,a) - 28A(s,a) - 4\sqrt{\bar{P}(s'|s,a)A(s,a)}, 0\}, \tag{5.4}$$

where $A(s,a) = \log(SAN_+(s,a)/\delta)/N_+(s,a)$ and the remaining probability mass goes to $\widetilde{P}(g \mid s,a)$. The optimistic policy is then the optimal policy in the SSP model defined by the transition function $\widetilde{P}$.

5.4.1 Analysis

In this section we prove Theorem 5.2.3. We start by showing that our new confidence sets contain P with high probability which implies that each episode ends in finite time with high probability. Consequently, we are able to bound the regret through summation of our confidence bounds.

We once again distinguish between *known* and *unknown* state-action pairs similarly to Algorithm 6. A state-action pair (s,a) becomes *known* at the end of an epoch if the total number of visits to (s,a) has passed $\alpha \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$ at some time step during the epoch (for some constant $\alpha > 0$). Note that at the end of the epoch, the visit count of (s,a) may be strictly larger than $\alpha \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$ but at most twice as much by the definition of our algorithm. Furthermore, we split each epoch into *intervals* similar to what did in Section 5.3. The first interval starts at the first time step and each interval ends once (1) the total cost in the interval accumulates to at least $B_\star$; (2) an unknown state-action pair is reached; (3) the current episode ends; or (4) the current epoch ends. We have the following observation.

Observation 5.4.1. *Let C_M denote the cost of the learner after M intervals. Observe that the total cost in each interval is at least $B_\star$ unless the interval ends in the goal state, in an unknown state-action pair or the epoch ends. Thus the total number of intervals satisfies*

$$M \le \frac{C_M}{B_\star} + 2SA \log T + K + O\left(\frac{B_\star S^2 A}{c_{min}} \log \frac{B_\star SA}{\delta c_{min}}\right),$$

and the total time satisfies $T \le C_M / c_{min}$.

Recall that in the analysis of Algorithm 6 we show that once all state-action pairs are known, the optimistic policies generated by the algorithm are proper in the true MDP. The same holds true for Algorithm 7, yet we never prove this directly. Instead, our proof goes as follows.[1] We prove that C_M, the cost accumulated by the learner during the first M intervals, is at most $K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{M}$ with high probability as long as no more than K episodes have been completed during these M intervals. We notice that once all state-action pairs are known, the total cost in each interval is at least $B_\star$ (ignoring intervals that end with the end of an epoch or an episode), which implies that the total number of intervals M is bounded by $C_M / B_\star$. This allows us to get a bound on C_M that is independent of the number of intervals by solving the inequality $C_M \lesssim K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{M} \lesssim K \cdot V^\star(s_{\text{init}}) + \sqrt{B_\star \cdot C_M}$. From this, and since the instantaneous costs are strictly positive (by

[1] We neglect low order terms here.

Assumption 5.2.1), it must be that the learner eventually completes all K episodes; i.e., there must be a time from which Algorithm 7 generates only proper policies.

Notation. The epoch that interval m belongs to is denoted by $i(m)$, other notations are as in Section 5.3.1. Note that since the optimistic policy is computed at the end of an epoch and not at the end of an interval, it follows that $\tilde{\pi}^m = \tilde{\pi}^{i(m)}$ and $\widetilde{V}^m = \widetilde{V}^{i(m)}$. The trajectory visited in interval m is denoted by $U^m = (s_1^m, a_1^m, \dots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$, where a_h^m is the action taken in s_h^m, and H^m is the length of the interval. In addition, the concatenation of the trajectories of the intervals up to and including interval m is denoted by $\bar{U}^m$, that is $\bar{U}^m = \cup_{m'=1}^m U^{m'}$.

High probability events. Throughout the analysis we denote $S^+ = S \cup \{g\}$. For every interval m we let Ω^m denote the event that the confidence set for epoch $i = i(m)$ contains the actual transition function P. Formally, if Ω^m holds then for all $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$, we have (denote $N_+^m(s,a) = \max\{1, N^m(s,a)\}$, $A_h^m = A(s_h^m, a_h^m)$)

$$|P(s'|s,a) - \bar{P}_m(s'|s,a)| \le 28A_h^m + 4\sqrt{\bar{P}_m(s'|s,a)A_h^m}. \tag{5.5}$$

In the following lemma we show that the events Ω^m hold with high probability.

Lemma 5.4.2. *With probability at least $1 - \delta/2$, Ω^m holds for all intervals m simultaneously.*

Regret analysis. In the following section, instead of bounding R_K, we bound $\widetilde{R}_M = \sum_{m=1}^M \sum_{h=1}^{H^m} c(s_h^m, a_h^m)\mathbb{I}\{\Omega^m\} - KV^\star(s_{\text{init}})$ for any number of intervals M. This implies Theorem 5.2.3 by the following argument. Theorem 5.4.2 implies that $\widetilde{R}_M = R_M$ with high probability for any number of intervals M (R_M is the true regret within the first M intervals). In particular, when M is the number of intervals in which the first K episodes elapse, this implies Theorem 5.2.3 (we show that the learner indeed completes these K episodes).

To bound $\widetilde{R}_M$, we use the next lemma to decompose $\widetilde{R}_M$ into two terms which we bound independently.

Lemma 5.4.3. *It holds that $\widetilde{R}_M = \sum_{m=1}^M \widetilde{R}_m^1 + \sum_{m=1}^M \widetilde{R}_m^2 - K \cdot V^\star(s_{init})$, where*

$$\widetilde{R}_m^1 = \big(\widetilde{V}^m(s_1^m) - \widetilde{V}^m(s_{H^m+1}^m)\big)\mathbb{I}\{\Omega^m\}, \quad \text{and}$$
$$\widetilde{R}_m^2 = \left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\}.$$

The lemma breaks down $\widetilde{R}_M$ into two terms. The first term accounts for the number of times in which the learner changes her policy in the middle of an episode which is at most the number of epochs. The second term sums the errors between the cost-to-go of the observed next state and its estimated expectation.

Indeed, $\sum_{m=1}^{M} \widetilde{R}_m^1$ is related to the total number of epochs which is at most $SA \log_2 T$ due to the following lemma.

Lemma 5.4.4. *It holds that* $\sum_{m=1}^{M} \widetilde{R}_m^1 \leq 2B_\star SA \log T + KV^\star(s_{init})$.

The next lemma shows that $\sum_{m=1}^{M} \widetilde{R}_m^2$ does not deviate from $\sum_{m=1}^{M} \mathbb{E}[\widetilde{R}_m^2 \mid \bar{U}^{m-1}]$ significantly.

Lemma 5.4.5. *With probability at least* $1 - \delta/4$,

$$\sum_{m=1}^{M} \widetilde{R}_m^2 \leq \sum_{m=1}^{M} \mathbb{E}\big[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\big] + 3B_\star \sqrt{M \log \frac{8M}{\delta}}.$$

The key property of the lemma is that the deviations between $\sum_{m=1}^{M} \widetilde{R}_m^2$ and its corresponding expectation is of order $\sqrt{M}$ and do not scale with T.

To prove the lemma, we recall that an interval ends at most at the first time step in which the accumulated cost in the interval surpasses $B_\star$. We show in our analysis that $\widetilde{V}^m(s) \leq V^\star(s) \leq B_\star$ for all $s \in \mathcal{S}$ due to the optimistic computation of $\tilde{\pi}^m$. Therefore, $\tilde{\pi}^m$ never picks an action whose instantaneous cost is more than $B_\star$. This implies that the total cost within each interval is at most $2B_\star$. Then, we use the Bellman equations to bound $\widetilde{R}_m^2$ by order of the total cost in the interval, and the lemma follows by an application of Azuma's concentration inequality.

Theorem 5.4.6 below bounds $\mathbb{E}\big[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\big]$ for every interval m by a sum of the confidence bounds used in Algorithm 7.

Lemma 5.4.6. *For every interval m,*

$$\begin{aligned}\mathbb{E}\big[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\big] \leq 16\mathbb{E}\left[\sum_{h=1}^{H^m} \sqrt{S\mathbb{V}_h^m A_h^m} \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] \\ + 272\mathbb{E}\left[\sum_{h=1}^{H^m} B_\star S A_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right],\end{aligned} \tag{5.6}$$

where $\mathbb{V}_h^m$ *is the empirical variance defined as* $\mathbb{V}_h^m = \sum_{s' \in \mathcal{S}^+} P(s' \mid s_h^m, a_h^m)\big(\widetilde{V}^m(s') - \mu_h^m\big)^2$, *and* $\mu_h^m = \sum_{s' \in \mathcal{S}^+} P(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')$.

The next step is the part of our proof in which our analysis departs from that of Algorithm 6. Note that when Ω^m holds, $\mathbb{V}_h^m \le B_\star^2$. Using this bound for each time step separately will result in a bound similar to that of Theorem 5.2.2. However, this bound is loose due to the following intuitive argument. Suppose that we replace $\widetilde{V}^m$ with the true cost-to-go function of $\tilde{\pi}^m$, V^m, in the definition of $\mathbb{V}_h^m$. Note that from the Bellman equations (Equation (2.5)) we have $V^m(s_h^m) > V^m(s_{h+1}^m)$ in expectation on consecutive time steps h and $h+1$ hence we surmise that in expectation $\mathbb{V}_h^m$ would also decrease on consecutive time steps. A similar argument holds when in reality we use $\widetilde{V}^m$ because all-but-one of the state-action pairs in the interval are known, and $\widetilde{V}^m$ is a "close enough" approximation of V^m on known state-action pairs since they have been sampled sufficiently many times. Indeed, in Theorem 5.4.7 we use the technique of [AOM17] to show that (up to a constant) $B_\star^2$ bounds the expected sum of the variances over the time steps of an interval.

Lemma 5.4.7. $\mathbb{E}\big[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \le 44 B_\star^2$.

Armed with Theorem 5.4.7, we upper bound $\sum_{m=1}^{M} \mathbb{E}\big[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\big]$ by applying some algebraic manipulation on Equation (5.6), and summing over all intervals which gives the next lemma.

Lemma 5.4.8. *With probability at least* $1-\delta/4$,

$$\sum_{m=1}^{M} \mathbb{E}\big[\widetilde{R}_m^2 \,|\bar{U}^{m-1}\big] \le 614 B_\star \sqrt{M S^2 A \log^2 \frac{TSA}{\delta}} \\ + 8160 B_\star S^2 A \log^2 \frac{TSA}{\delta}.$$

Theorem 5.2.3 is obtained by first applying a union bound on Theorems 5.4.2, 5.4.5 and 5.4.8, plugging in the bounds of Theorems 5.4.4, 5.4.5 and 5.4.8 into Theorem 5.4.3, and bounding T and M using Observation 5.4.1. This results in a quadratic inequality in $\sqrt{C_M}$ and solving it yields the theorem.

5.5 Lower Bound

In this section we give an overview of the proof for Theorem 5.2.6. For clarity we restate the theorem.

Theorem (restatement of Theorem 5.2.6). There exists an SSP problem instance $M = (S, A, P, c, s_{\text{init}})$ in which $V^\star(s) \le B_\star$ for all $s \in \mathscr{S}$, $S \ge 2$, $A \ge 16$, $B_\star \ge 2$, $K \ge SA$, and

$c(s,a) = 1$ for all $s \in \mathscr{S}, a \in A$, such the expected regret of any learner after K episodes satisfies

$$\mathbb{E}[R_K] \geq \frac{1}{1024} B_\star \sqrt{SAK}.$$

The proof of our lower bound takes similar steps to the one in [JOA10]. Note that one cannot simply use a reduction to the average-cost setting in our case because the number of steps taken by the algorithm is potentially unbounded, and not the same as the number of steps taken by the optimal policy.

Still, our lower bound matches the one for finite-horizon MDPs of $\Omega(\sqrt{HSAT})$, where H is the horizon and T is the total number of time steps. Since the length of each episode is H, we have that $T = HK$ and the lower bound takes the form of $\Omega(H\sqrt{SAK})$. In our case, $B_\star$ replaces the horizon H as an upper bound on the expected cost of the optimal policy, and we get the same linear dependence in this parameter.

Before constructing an instance for which we can prove the general lower bound, we consider a simpler instance that consists of only the initial state s_{init} and the goal state g. The actions are all the same, except for one optimal action $a^\star$ which is chosen uniformly at random. While all actions (including $a^\star$) suffer a cost of 1, $a^\star$ has a better probability of transitioning to the goal state, that is, $P(g \mid s_{\text{init}}, a^\star) = 1/B_\star$ compared to $P(g \mid s_{\text{init}}, a) = (1-\varepsilon)/B_\star$ for all other actions $a \neq a^\star$.

Notice that the optimal policy $\pi^\star$ chooses $a^\star$ and has an expected cost of exactly $B_\star$. Therefore, the job of the learner is simply to identify $a^\star$. In the supplementary material we show that the regret of the learner in this case must be $\Omega(B_\star\sqrt{AK})$.

Subsequently, we build our general hard instance by taking S copies of the aforementioned simple MDP and picking the initial state in every episode uniformly at random. Since the copies are not connected in any way, the lower bound applies to each of them separately. Notice that every state will be visited K/S times in expectation, so the expected regret will be

$$\Omega\left(\sum_{s\in\mathscr{S}} B_\star \sqrt{A\frac{K}{S}}\right) = \Omega(B_\star\sqrt{SAK}).$$

Interestingly enough, although playing proper policies was a major concern in the construction of our algorithms, the hard instance we have built does not have any improper polices at all.

6 Minimax Regret for Stochastic Shortest Path

This chapter is based on:

Alon Cohen, Yonathan Efroni, Yishay Mansour and Aviv Rosenberg. Minimax regret for stochastic shortest path. In *Advances in Neural Information Processing Systems, NeurIPS 2021.*

This chapter presents the first minimax optimal regret bound for stochastic shortest path (SSP). While the algorithms in Chapter 4 were based on a direct application of the optimism in face of uncertainty principle, in this chapter we present algorithms based on a reduction to RL in finite-horizon MDPs.

6.1 Summary of Results

In Section 6.2 we present a novel black-box reduction from SSP to finite-horizon MDPs (Algorithm 8), that yields $\sqrt{K}$ regret bounds when combined with a certain class of optimistic algorithms for regret minimization in finite-horizon MDPs that we call *admissible* (Definition 6.2.1). The regret analysis for the reduction is described in Section 6.3, and in Section 6.4 we present an admissible algorithm for regret minimization in finite-horizon MDPs called `ULCVI`. We show that it guarantees the following optimal regret in the finite-horizon setting (stated formally in Theorem 6.4.1). Note that (for large enough number of episodes) this bound depends only on the expected cost of the optimal policy and not on the horizon H.

Theorem 6.1.1. *Running* `ULCVI` *(Algorithm 9 in Section 6.4) in a finite-horizon MDP guarantees, with probability at least* $1-\delta$*, a regret bound of*

$$O\left(\sqrt{(B_\star^2+B_\star)SAM}\log\frac{MHSA}{\delta}+H^4B_\star^{-1}S^2A\log^{3/2}\frac{MHSA}{\delta}\right),$$

for any number of episodes $M \geq 1$ simultaneously.

Combining `ULCVI` with our reduction yields the following minimax optimal regret bound for SSP.

Theorem 6.1.2. *Running the reduction in Algorithm 8 with the finite-horizon regret minimization algorithm* `ULCVI` *ensures, with probability at least $1-\delta$,*

$$R_K = O\left(\sqrt{(B_\star^2+B_\star)SAK}\log\frac{KT_\star SA}{\delta} + T_\star^5 B_\star^{-2} S^2 A \log^6\frac{KT_\star SA}{\delta}\right).$$

Remark 1. An important observation is that this regret bound is meaningful even for small K. Unlike finite-horizon MDPs, where linear regret is trivial, in SSP ensuring finite regret is not easy. Our regret bound also implies that if we play for only one episode, i.e., we are only interested in the time it takes to reach the goal state, then it will take us at most $\widetilde{O}(T_\star^5 B_\star^{-2} S^2 A)$ time steps to do so.

Remark 2. Note that our algorithm needs to know an upper bound on $T_\star$ in advance. However, if all costs are strictly positive (i.e., at least $c_{\min} > 0$), then there is a trivial upper bound of $B_\star / c_{\min}$. In this case, our algorithm keeps an optimal regret bound for large enough K, since the bound on $T_\star$ only appears in the additive factor. Some previous work used a perturbation argument to generalize their results from the $c_{\min}$ case to general costs [TGV$^+$20, RCMK20, RM21b]. In our case, it will not work since the dependence on $1/c_{\min}$ in the additive term is too large. This may be an inherent shortcoming of using finite-horizon reduction to solve SSPs, as it also appears in the works of [CLW21, CL21] for the adversarial setting.

Remark 3. In practice, one can think of $T_\star$ as a parameter of the algorithm that controls computational complexity and the number of steps to complete K episodes. By choosing the parameter $T_\star = x$ for example, we can guarantee that the regret bound of Theorem 6.1.2 holds against the best proper policy with expected time to the goal of at most x (assuming there exists one), and we can also guarantee that the total computational complexity of the algorithm is $\widetilde{O}(x \log K)$ (see Remark 5). Furthermore, the algorithm will take at most $\widetilde{O}(xK + poly(x,S,A))$ steps to complete K episodes.

Remark 4. While the additive term in our regret bound is standard for most cases, it becomes large when $B_\star$ is extremely small because of the dependence in $B_\star^{-1}$. This was not an issue in previous work [TGV$^+$20, RCMK20] since they assumed that the costs are deterministic and known. We believe that this dependence is an artifact of our analysis that may be avoided with a more careful definition of $\omega_{\texttt{ALG}}$ (see Definition 6.2.1) that depends on the actual cost in each state-action pair and not just $B_\star$. Nevertheless, the main focus of this paper is on establishing that the minimax optimal regret for SSP is $\widetilde{\Theta}(\sqrt{(B_\star^2+B_\star)SAK})$, and not on optimizing lower order terms. By that we also show that this is the minimax optimal regret for finite-horizon which is independent of the horizon H (up to logarithmic factors). Tightening the additive term and eliminating its dependence in $B_\star^{-1}$ is left as an interesting future direction.

In Section D.4 we prove that our regret bound is indeed minimax optimal. To complement the $\Omega(B_\star\sqrt{SAK})$ lower bound of [RCMK20] that assumes $B_\star \geq 1$, we provide the following tighter lower bound for the case that $B_\star < 1$.

Theorem 6.1.3. *Let $B_\star \leq \frac{1}{2}$. There exists an SSP problem instance $\mathscr{M} = (\mathscr{S}, \mathscr{A}, P, c, s_{init}, g)$ in which $V^\star(s) \leq B_\star$ for all $s \in S$, $S \geq 2$, $A \geq 2$, $K \geq B_\star SA$, such the expected regret of any learner after K episodes satisfies*

$$\mathbb{E}[R_K] \geq \frac{1}{32}\sqrt{B_\star SAK}.$$

6.2 A Black-Box Reduction from SSP to Finite-Horizon

Our algorithm takes as input an algorithm `ALG` for regret minimization in finite-horizon MDPs, and uses it to perform a black-box reduction. The algorithm is depicted below as Algorithm 8.

The algorithm breaks the individual time steps that comprise each of the K episodes into *intervals* of H time steps. If the agent reaches the goal state before H time steps, we simply assume that she stays in g until H time steps are elapsed. We see each interval as one episode of a finite-horizon model $\widehat{\mathscr{M}} = (\widehat{\mathscr{S}}, \mathscr{A}, \widehat{P}, H, \hat{c}, \hat{c}_f)$, where $\widehat{\mathscr{S}} = \mathscr{S} \cup \{g\}$ and $\hat{c}_f : \widehat{\mathscr{S}} \to \mathbb{R}$ is a set of terminal costs defined by $\hat{c}_f(s) = 8B_\star \mathbb{I}\{s \neq g\}$, where $\mathbb{I}\{s \neq g\}$ is the indicator function that equals 1 if $s \neq g$ and 0 otherwise. Moreover, $\widehat{P}, \hat{c}$ are the natural extensions of P, c to the goal state. That is, $\hat{c}(s,a) = c(s,a)\mathbb{I}\{s \neq g\}$ and

$$\widehat{P}(s' \mid s,a) = \begin{cases} P(s' \mid s,a), & s \neq g; \\ 1, & s = g, s' = g; \\ 0, & s = g, s' \neq g. \end{cases}$$

The horizon H (which we will set to be roughly $T_\star$) is chosen such that the optimal SSP policy will reach the goal state in H time steps with high probability (recall that the expected hitting time of the optimal policy is bounded by $T_\star$). The additional terminal cost is there to encourage the agent to reach the goal state within H steps, which otherwise is not necessarily optimal with respect to the planning horizon.

Algorithm 8 REDUCTION FROM SSP TO FINITE-HORIZON MDP

```
input: state sapce $\mathcal{S}$, action space $\mathcal{A}$, initial state $s_{\text{init}}$, goal state $g$, confidence parameter $\delta$, number of episodes $K$, bound on the expected cost of the optimal policy $B_\star$, bound on the expected time of the optimal policy $T_\star$ and algorithm ALG for regret minimization in finite-horizon MDPs.
initialize ALG with state space $\widehat{\mathcal{S}} = \mathcal{S} \cup \{g\}$, action space $\mathcal{A}$, horizon $H = 8T_\star \log(8K)$, confidence parameter $\delta/4$, terminal costs $\hat{c}_f(s) = 8B_\star \mathbb{I}\{s \neq g\}$ and bound on the expected cost of the optimal policy $9B_\star$.
initialize intervals counter $m \leftarrow 0$ and time steps counter $t \leftarrow 1$.
for $k = 1, \dots, K$ do
  set $s_t \leftarrow s_{\text{init}}$.
  while $s_t \neq g$ do
    set $m \leftarrow m + 1$, feed initial state $s_t$ to ALG and obtain policy $\pi^m = \{\pi^m_h : \widehat{\mathcal{S}} \to \mathcal{A}\}_{h=1}^H$.
    for $h = 1, \dots, H$ do
      play action $a_t = \pi^m_h(s_t)$, suffer cost $C_t \sim c(s_t, a_t)$, and set $s^m_h = s_t, a^m_h = a_t, C^m_h = C_t$.
      observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ and set $t \leftarrow t + 1$.
      if $s_t = g$ then
        pad trajectory to be of length $H$ and BREAK.
      end if
    end for
    set $s^m_{H+1} = s_t$.
    feed trajectory $U^m = (s^m_1, a^m_1, \dots, s^m_H, a^m_H, s^m_{H+1})$ and costs $\{C^m_h\}_{h=1}^H$ to ALG.
  end while
end for
```

The algorithm ALG is initialized with the state and action spaces as in the original SSP instance, the horizon length H, a confidence parameter $\delta/4$, a set of terminal costs $\hat{c}_f$ and a bound on the expected cost of the optimal policy in the finite-horizon model $9B_\star$. At the beginning of each interval, it takes as input an initial state and outputs a policy to be used throughout the interval. In the end of the interval it receives the trajectory and costs observed through the interval.

Note that while Algorithm 8 may run any finite-horizon regret minimization algorithm, in the analysis we require that ALG possesses some properties (that most optimistic algorithms already have) in order to establish our regret bound. We specifically require ALG to be an *admissible* algorithm—a model-based optimistic algorithm for regret minimization in finite-horizon MDPs, e.g., UCBVI [AOM17] and EULER [ZB19]. Admissible algorithms are defined formally as follows.

Definition 6.2.1. A model-based algorithm ALG for regret minimization in finite-horizon MDPs is called *admissible* if, when running ALG with confidence parameter δ, there is a good event that holds with probability at least $1-\delta$, under which the following hold:

(i) ALG provides anytime regret guarantees without prior knowledge of the number of episodes, and when the initial state of each episode is arbitrary. The regret bound that ALG guarantees for M episodes is denoted by $\widehat{\mathscr{R}}_{\texttt{ALG}}(M)$, for some non-decreasing function $\widehat{\mathscr{R}}_{\texttt{ALG}}$.

(ii) The policy π^m that ALG picks in episode m is greedy with respect to an estimate of the optimal policy's Q-function.

(iii) The algorithm's estimate $\underline{V}^m$ of $\widehat{V}^\star$ (the cost-to-go function associated with the optimal finite-horizon policy) is optimistic, i.e., $\underline{V}^m_h(s) \le \widehat{V}^\star_h(s)$ for every $s \in S$ and $h = 1,\dots,H+1$.

(iv) ALG computes $\underline{V}^m$ using estimates $\tilde{c}^m, \widetilde{P}^m$ of the cost function $\hat{c}$ and the transition function $\widehat{P}$, respectively. There exists $\omega_{\texttt{ALG}}$ which is a function of H,S,A such that: if state-action pair (s,a) was visited at least $\omega_{\texttt{ALG}} \log \frac{MHSA}{\delta}$ times, then $|\tilde{c}^m_h(s,a) - \hat{c}(s,a)| \le B_\star/H$ and $\|\widetilde{P}^m(\cdot \mid s,a) - \widehat{P}(\cdot \mid s,a)\|_1 \le 1/(9H)$.

Using an admissible algorithm in Algorithm 8 enables us to bound the total number of intervals, thus ensuring that the agent reaches the goal state in almost every interval. This is because, as ALG is optimistic, it will try to avoid the terminal cost (which is suffered in all states except for g) by reaching the goal state. In addition, ALG will succeed in doing so once it has a good enough estimation of the transition function. Armed with the notion of admissibility, in the sequel we prove the following regret bound for any admissible algorithm ALG. The proof of Theorem 6.1.2 is now given by combining Theorem 6.2.1 with the regret bound of ULCVI in Theorem 6.1.1.

Theorem 6.2.1. *Let ALG be an admissible algorithm for regret minimization in finite-horizon MDPs and denote its regret in M episodes by $\widehat{\mathscr{R}}_{\texttt{ALG}}(M)$. Then, running Algorithm 8* *with* `ALG` *ensures that, with probability at least* $1-\delta$,

$$R_K \le \widehat{\mathscr{R}}_{\texttt{ALG}}\left(4K + 4\cdot 10^4 SA\omega_{\texttt{ALG}}\log\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta}\right) \\ + O\left(\sqrt{(B_\star^2+B_\star)K\log\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta}} + T_\star\omega_{\texttt{ALG}}SA\log^2\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta}\right),$$

where $\omega_{\texttt{ALG}}$ *is a quantity that depends on the algorithm* `ALG` *and on* S,A,H.

Remark 5 (Computational complexity). Our reduction directly inherits the computational complexity of the finite-horizon algorithm `ALG` in M episodes, where $M \approx K + poly(S,A,T_\star)$ by Theorem 6.3.3. The computational complexity of `ULCVI` is $O(HS^3A^2\log(MH))$, and therefore our optimal regret for SSP is achieved in total computational complexity of $O\big(T_\star S^3A^2\log^2\frac{KT_\star SA}{\delta}\big)$ which is only logarithmic in the number of episodes.

6.2.1 Unknown expected optimal cost

Inspired by techniques for estimation of the SSP-diameter in the adversarial SSP literature [RM21b, CL21], in Section D.3 we show that our reduction does not need to know $B_\star$ in advance, but can instead estimate it on the fly.

We can obtain a reasonable estimate (up to a constant multiplicative factor) of the cost-to-go from state s by running the `Bernstein-SSP` algorithm of [RCMK20] for regret minimization in SSPs (that does not need to know $B_\star$) with initial state s for roughly $T_\star^2S^2A$ episodes. Thus, we can apply our reduction while utilizing our first visits to each state in order to estimate its cost-to-go.

We operate in *phases* where each phase ends when some state is visited at least $T_\star^2S^2A$ times, and all states that were not visited enough are treated as the goal state. Once we reach a poorly visited state, we simply run an episode of the corresponding `Bernstein-SSP` algorithm. Notice that this comes at a computational cost that is independent of the number of episodes K (since we use `Bernstein-SSP` for a small number of episodes), and in Section D.3 we show that it achieves similar regret bounds with only an additional additive factor of $\widetilde{O}(T_\star^3S^3A)$.

6.3 Regret Analysis

In this section we prove Theorem 6.2.1. Below we give a high-level overview of the proofs and defer the details to Section D.1. We start the analysis with a regret decomposition that states that the SSP regret can be bounded by the sum of two terms: the expected regret of the finite-horizon algorithm, and the deviation of the actual cost in each interval from its expected value. To that end, we use the notations: M for the total number of intervals, $U^m = (s_1^m, a_1^m, \dots, s_h^m, a_h^m, s_{H+1}^m)$ for the trajectory visited in interval m, C_h^m for the cost suffered in step h of interval m, π^m for the policy chosen by `ALG` for interval m, and $\widehat{V}_h^\pi(s)$ for the expected finite-horizon cost when playing policy π starting from state s in time step h.

Lemma 6.3.1. *For $H = 8T_\star \log(8K)$, we have the following bound on the regret of Algorithm 8:*

$$R_K \le \widehat{\mathscr{R}}_{ALG}(M) + \sum_{m=1}^{M} \left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}_1^{\pi^m}(s_1^m) \right) + B_\star. \tag{6.1}$$

The bound in Equation (6.1) is comprised of two summands and an additional constant. The first summand is an upper bound on the expected finite-horizon regret which we acquire by the admissibility of `ALG` (Definition 6.2.1). Note that this bound is in terms of the number of intervals M (i.e., the number of finite-horizon episodes) which is a random variable and not necessarily bounded. In what follows we show that, using the admissibility of `ALG`, we can actually bound M by the number of SSP episodes K plus a constant that depends on $\omega_{\texttt{ALG}}, S, A, T_\star$ (but not on K). The second summand in Equation (6.1) relates to the deviation of the total finite-horizon cost from its expected value.

The proof of Theorem 6.3.1 builds on two key ideas. The first is that, by setting H to be $O(T_\star \log K)$, we ensure that the expected cost of the optimal policy in the SSP model $\mathscr{M}$ is close to that in the finite-horizon model $\widehat{\mathscr{M}}$. The second idea is that if the agent does not reach the goal state in a certain interval, then she must suffer the terminal cost in the finite-horizon model. Therefore, although in a single episode there may be many intervals in which the agent does not reach the goal state, we can upper bound the cost in these extra intervals in $\mathscr{M}$ by the corresponding terminal costs in $\widehat{\mathscr{M}}$.

Next, we bound the deviation of the actual cost in each interval from its expected value which appears as the second summand in Equation (6.1). The bound is due to the following lemma.

Lemma 6.3.2. *Assume that the reduction is performed using an admissible algorithm `ALG`. Then, the following holds with probability at least $1 - 3\delta/8$,*

$$\sum_{m=1}^{M} \left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}_1^{\pi^m}(s_1^m) \right) = O\left(\sqrt{(B_\star^2 + B_\star) M \log \frac{M}{\delta}} + H\omega_{ALG} SA \log \frac{MKT_\star SA}{\delta} \right).$$

The key observation here relies on the notion of *unknown* state-action pairs – pairs that were not visited at least $\omega_{\texttt{ALG}}$ times. After $\omega_{\texttt{ALG}}$ visits to some state-action pair s,a, we have a reasonable estimate of the next-state distribution $P(\cdot \mid s,a)$ therefore we can show that the expected accumulated cost in an interval until reaching an unknown state-action pair or the goal state is of order $B_\star$. Moreover, the second moment of this cost is of order $B_\star^2 + B_\star$. Thus, using Freedman inequality, we bound the deviation by $\widetilde{O}(\sqrt{(B_\star^2+B_\star)M})$, plus a cost of $O(H)$ for each "bad" interval in which we do not reach an unknown state-action pair or the goal state (there are roughly $\omega_{\texttt{ALG}}SA$ such intervals).

Lastly, we need to bound the number of intervals M to obtain a regret bound in terms of K and not M (notice that M is a random variable that is not bounded a-priori).

Lemma 6.3.3. *Assume that the reduction is performed using an admissible algorithm* ALG*. Then, with probability at least* $1 - {}^{3\delta}/_8$, $M \le 4K + 4\cdot 10^4 SA\omega_{\mathtt{ALG}} \log(KT_\star SA\omega_{\mathtt{ALG}}/\delta)$.

The proof shows that in every interval there is a constant probability to reach either the goal state or an unknown state-action pair. Leveraging this observation with a concentration inequality, we can bound the number of intervals by $\widetilde{O}(K + \omega_{\texttt{ALG}}SAH)$.

We can now prove a bound on the regret of Algorithm 8 using any admissible algorithm ALG.

Proof of Theorem 6.2.1. The regret bound of ALG, Theorems 6.3.2 and 6.3.3 all hold with probability at least $1-\delta$, via a union bound. Using Theorems 6.3.1 and 6.3.2 we can write

$$R_K \le \widehat{\mathscr{R}}_{\texttt{ALG}}(M) + O\left(\sqrt{(B_\star^2+B_\star)M\log\frac{M}{\delta}} + H\omega_{\texttt{ALG}}SA\log\frac{MKT_\star SA}{\delta}\right) + B_\star.$$

Finally, we use Theorem 6.3.3 to bound M by $4K + 4\cdot 10^4 SA\omega_{\texttt{ALG}}\log(KT_\star SA\omega_{\texttt{ALG}}/\delta)$. □

6.4 ULCVI: an admissible algorithm for finite-horizon MDPs

In this section we present the Upper Lower Confidence Value Iteration algorithm (ULCVI; Algorithm 9) for regret minimization in finite-horizon MDPs. This result holds independently of our SSP algorithm. Since the algorithm is similar to previous optimistic algorithms for the finite-horizon setting, e.g., UCBVI [AOM17] and ORLC [DLWB19], we defer the analysis to Section D.2 and focus on our technical novelty – bounding the regret in terms of the optimal value function and not the horizon.

In each episode m, the ULCVI algorithm maintains an optimistic lower bound $\underline{V}_h^m(s)$ and a pessimistic upper bound $\bar{V}_h^m(s)$ on the cost-to-go function of the optimal policy $V_h^\star(s)$, and acts greedily with respect to the optimistic estimates. These optimistic and pessimistic estimates are computed based on the empirical transition function $\bar{P}^{m-1}(s' \mid s,a)$ and the empirical cost function $\bar{c}^{m-1}(s,a)$ to which we add an exploration bonus $b_c^m(s,a) + b_p^m(s,a)$, where b_p^m handles the approximation error in the transitions and b_c^m handles the approximation error in the costs. The bonuses are defined as follows,

$$b_c^m(s,a) = \sqrt{\frac{2\overline{\mathrm{Var}}_{s,a}^{m-1}(C) L_m}{\max\{1, n^{m-1}(s,a)\}}} + \frac{5L_m}{\max\{1, n^{m-1}(s,a)\}} \tag{6.2}$$
$$b_p^m(s,a) = \sqrt{\frac{2\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m) L_m}{\max\{1, n^{m-1}(s,a)\}}} + \frac{62H^3 B_\star^{-1} S L_m}{\max\{1, n^{m-1}(s,a)\}} + \frac{B_\star}{16H^2}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')],$$

where $L_m = 3\log(3SAHm/\delta)$ is a logarithmic factor and $n^{m-1}(s,a)$ is the number of visits to (s,a) in the first $m-1$ episodes. Furthermore, $\overline{\mathrm{Var}}_{s,a}^{m-1}(C)$ is the empirical variance of the observed costs in (s,a) in the first $m-1$ episodes.[1] Lastly, the term $\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)$ is the variance of the next state value $\underline{V}_{h+1}^m$ from state-action pair (s,a), calculated via the empirical transition model, i.e., $\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m) = \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\underline{V}_{h+1}^m(s')^2] - \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\underline{V}_{h+1}^m(s')]^2$.

For improved computational complexity, we compute the optimistic policy only in episodes in which the number of visits to some state-action pair was doubled. This ensures that the number of optimistic policy computations grows only logarithmically with the number of episodes, i.e., it is bounded by $3SA\log(MH)$. Since each optimal policy computation costs $O(HS^2A)$ in the finite-horizon MDP model, our algorithm enjoys a total computational complexity of $O(HS^3A^2\log(MH))$.

For clarity, we keep the notation of the finite-horizon MDP as $\widehat{\mathscr{M}} = (\mathscr{S}, \mathscr{A}, \widehat{P}, H, \hat{c}, \hat{c}_f)$, and let $B_\star = \max_{s,h}\widehat{V}_h^\star(s)$ where $\widehat{V}^\pi$ is the value function of policy π (in the case of our SSP reduction this parameter is simply $9B_\star$ by Theorem D.1.1). This implies that $\hat{c}_f(s) \le B_\star$ for every s, and for simplicity, we assume that $B_\star \le H$. Thus, the maximal total cost in an episode is bounded by $H + B_\star \le 2H$. In Section D.2 we prove the following high probability regret bound.

Theorem 6.4.1. ULCVI *(Algorithm 9) is admissible with the following guarantees:*

(i) With probability at least $1-\delta$*, the regret bound of* ULCVI *is*

$$\widehat{\mathscr{R}}_{\texttt{ULCVI}}(M) = O\Big(\sqrt{(B_\star^2 + B_\star)SAM}\log\frac{MHSA}{\delta} + H^4 B_\star^{-1} S^2 A \log^{3/2}\frac{MHSA}{\delta}\Big)$$

[1] The empirical variance of n numbers $a_1, \ldots, a_n$ is defined by $\frac{1}{n}\sum_{i=1}^n\big(a_i - \frac{1}{n}\sum_{j=1}^n a_j\big)^2$.

for any number of episodes $M \geq 1$.

(ii) $\omega_{\texttt{ULCVI}} = O(H^4 B_\star^{-2} S)$.

Our analysis resembles the one in [EMSM21], and is adapted to the stationary MDP setting (i.e., the transition function does not depend on the time step h), and to the setting where we have costs instead of rewards, and terminal costs (which do not appear in previous work). By the definition of the algorithm and the regret bound in Theorem 6.4.1, it is clear that properties (i)-(iii) in Definition 6.2.1 of admissible algorithms hold. For property (iv), we use standard concentration inequalities and the definition of the bonuses in Equation (6.2) in order to show it holds for $\omega_{\texttt{ULCVI}} = O(H^4 B_\star^{-2} S)$.

To obtain a regret bound whose leading term depends on $B_\star$ and not H, we start with a standard regret analysis for optimistic algorithms that establishes the regret scales with the square-root of the variance of the value functions of the agent's policies, i.e.,

$$\widehat{\mathscr{R}}_{\texttt{ULCVI}}(M) \lesssim \sqrt{SA} \sqrt{\sum_{m=1}^{M} \sum_{h=1}^{H} \mathrm{Var}_{P(\cdot|s_h^m, a_h^m)}(V_{h+1}^{\pi^m})} + H^4 B_\star^{-1} S^2 A,$$

up to logarithmic factors and lower order terms. This can be further bounded by the second moment of the cumulative cost in each episode as follows,

$$\widehat{\mathscr{R}}_{\texttt{ULCVI}}(M) \lesssim \sqrt{SA} \sqrt{\sum_{m=1}^{M} \mathbb{E}\left[\left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) \right)^2 \middle| \bar{U}^m \right]} + H^4 B_\star^{-1} S^2 A,$$

where $\bar{U}^m$ is the sequence of state-action pairs observed up to episode m. Leveraging our techniques for the SSP reduction (but independently), we show that the second moment of the cumulative cost until an unknown state-action pair is reached can be bounded by $O(B_\star^2 + B_\star)$. Therefore, we have at most $\widetilde{O}(H^4 B_\star^{-2} S^2 A)$ episodes in which we bound the second moment trivially by $O(H^2)$, and in the rest of the episodes we can bound it by $O(B_\star^2 + B_\star)$. Together this yields the theorem as follows,

$$\widehat{\mathscr{R}}_{\texttt{ULCVI}}(M) \lesssim \sqrt{SA} \sqrt{(B_\star^2 + B_\star) M + H^2 \cdot H^4 B_\star^{-2} S^2 A} \lesssim \sqrt{(B_\star^2 + B_\star) SAM} + H^4 B_\star^{-1} S^2 A.$$

Algorithm 9 Upper Lower Confidence Value Iteration (ULCVI)

```
input: state space $\mathscr{S}$, action space $\mathscr{A}$, horizon $H$, confidence parameter $\delta$, terminal costs $\hat{c}_f$ and upper bound on the expected cost of the optimal policy $B_\star$.
initialize: $n^0(s,a) = 0, n^0(s,a,s') = 0, N^0(s,a) = 0, N^0(s,a,s') = 0\ \forall(s,a,s')$.
initialize: $C^0(s,a) = 0, \bar{c}^0(s,a) = 0, \bar{P}^0(s'|s,a) = \mathbb{I}\{s' = s\}\ \forall(s,a,s')$.
initialize: PlanningTrigger = true.
for $m = 1,2,\dots$ do
    observe initial state $s_1^m$.
    if PlanningTrigger = true then
        set $n^{m-1}(s,a) \leftarrow N^{m-1}(s,a), n^{m-1}(s,a,s') \leftarrow N^{m-1}(s,a,s')\ \forall(s,a,s')$.
        set $\bar{P}^{m-1}(s'|s,a) \leftarrow \frac{n^{m-1}(s,a,s')}{\max\{1,n^{m-1}(s,a)\}}, \bar{c}^{m-1}(s,a) \leftarrow \frac{C^{m-1}(s,a)}{\max\{1,n^{m-1}(s,a)\}}\ \forall(s,a,s')$.
        compute $\{\pi_h^m(s)\}_{s,h}$ via Optimistic-Pessimistic Value Iteration (Algorithm 10).
        set PlanningTrigger ← false.
    else
        set $n^{m-1}(s,a) \leftarrow n^{m-2}(s,a), n^{m-1}(s,a,s') \leftarrow n^{m-2}(s,a,s')\ \forall(s,a,s')$
        set $\bar{P}^{m-1}(s'|s,a) \leftarrow \bar{P}^{m-2}(s'|s,a), \bar{c}^{m-1}(s,a) \leftarrow \bar{c}^{m-2}(s,a)\ \forall(s,a,s')$.
        set $\pi_h^m(s) \leftarrow \pi_h^{m-1}(s)$ for all $s \in S$ and $h = 1,\dots,H$.
    end if
    set $N^m(s,a) \leftarrow N^{m-1}(s,a), N^m(s,a,s') \leftarrow N^{m-1}(s,a,s'), C^m(s,a) \leftarrow C^{m-1}(s,a)\ \forall(s,a,s')$.
    for $h = 1,\dots,H$ do
        pick action $a_h^m = \pi_h^m(s_h^m)$.
        suffer cost $C_h^m \sim \hat{c}(s_h^m,a_h^m)$ and observe next state $s_{h+1}^m \sim \widehat{P}(\cdot \mid s_h^m,a_h^m)$.
        update visits counters $n^m(s_h^m,a_h^m) \leftarrow n^m(s_h^m,a_h^m) + 1, n^m(s_h^m,a_h^m,s_{h+1}^m) \leftarrow n^m(s_h^m,a_h^m,s_{h+1}^m) + 1$.
        update accumulated cost $C^m(s_h^m,a_h^m) \leftarrow C^m(s_h^m,a_h^m) + C_h^m$.
        if $N^m(s_h^m,a_h^m) \geq 2n^{m-1}(s_h^m,a_h^m)$ then
            set PlanningTrigger ← true.
        end if
    end for
    Suffer terminal cost $\hat{c}_f(s_{H+1}^m)$.
end for
```

Algorithm 10 Optimistic-Pessimistic Value Iteration

```
input: $n^{m-1}, \bar{P}^{m-1}, \bar{c}^{m-1}, \hat{c}_f, B_\star$.
initialize $\underline{V}^m_{H+1}(s) = \bar{V}^m_{H+1}(s) = \hat{c}_f(s)$ for all $s \in \mathscr{S}$.
for $h = H, H-1, \ldots, 1$ do
    for $s \in S$ do
        for $a \in A$ do
            set the bonus $b^m_h(s,a) = b^m_c(s,a) + b^m_p(s,a)$ defined in Equation (6.2).
            compute optimistic and pessimistic Q-functions:
                $\underline{Q}^m_h(s,a) = \bar{c}^{m-1}(s,a) - b^m_h(s,a) + \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\underline{V}^m_{h+1}(s')]$
                $\bar{Q}^m_h(s,a) = \bar{c}^{m-1}(s,a) + b^m_h(s,a) + \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\bar{V}^m_{h+1}(s')]$.
        end for
        $\pi^m_h(s) \in \arg\min_{a \in A} \underline{Q}^m_h(s,a)$.
        $\underline{V}^m_h(s) = \max\Big\{\underline{Q}^m_h(s, \pi^m_h(s)), 0\Big\}$, $\bar{V}^m_h(s) = \min\big\{\bar{Q}^m_h(s, \pi^m_h(s)), H\big\}$.
    end for
end for
```

7 Learning Adversarial Stochastic Shortest Path

This chapter is based on:

Aviv Rosenberg and Yishay Mansour. Stochastic Shortest Path with Adversarially Changing Costs. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021.*

This chapter presents the first algorithms for regret minimization in adversarial SSPs. We start by introducing the adversarial SSP model. Then, we present and analyze our algorithms that handle both known and unknown transitions.

7.1 Problem Formulation

An adversarial SSP problem is defined by an MDP $M = (\mathscr{S}, \mathscr{A}, P, s_{\text{init}}, g)$ and a sequence $\{c^k : \mathscr{S} \times \mathscr{A} \to [0,1]\}_{k=1}^K$ of cost functions. The learner interacts with $\mathscr{M}$ in episodes, where c^k is the cost function for episode k. However, it is revealed to the learner only in the end of the episode. Formally, the learner starts each episode k at the initial state $s_1^k = s_{\text{init}}$. In each step i of the episode, the learner observes its current state s_i^k, picks an action a_i^k and moves to the next state s_{i+1}^k sampled from $P(\cdot \mid s_i^k, a_i^k)$. The episode ends when the goal state g is reached, and then the learner observes c^k and suffers cost $\sum_{i=1}^{I^k} c^k(s_i^k, a_i^k)$ where I^k is the length of the episode. Importantly, I^k is a random variable that might be infinite. This is the unique challenge of SSP compared to finite-horizon.

Under the additional assumption that every improper policy suffers infinite expected cost from some state, [BT91] show that the optimal policy is stationary, deterministic and proper; and that every proper policy π satisfies the following Bellman equations for every

$s \in S$:

$$V^{\pi}(s) = \sum_{a \in \mathscr{A}} \pi(a \mid s)\Big(c(s,a) + \sum_{s' \in \mathscr{S}} P(s' \mid s,a) V^{\pi}(s')\Big)$$
$$T^{\pi}(s) = 1 + \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} \pi(a \mid s) P(s' \mid s,a) T^{\pi}(s'). \tag{7.1}$$

Learning Formulation. The learner's goal is to minimize its total cost. Its performance is measured by the *regret* – the difference between the learner's total cost in K episodes and the total expected cost of the best *proper* policy in hindsight:

$$R_K = \sum_{k=1}^{K} \sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) - \min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^{K} V_k^{\pi}(s_{\text{init}}),$$

where V_k^{π} is the cost-to-go of policy π with respect to (w.r.t) cost function c^k, and Π_{proper} is the set of proper policies. If I^k is infinite for some k, we define $R_K = \infty$ forcing the learner to reach the goal in every episode. We also denote by $\pi^\star = \arg\min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^{K} V_k^{\pi}(s_{\text{init}})$ the best policy in hindsight.

Our analysis makes use of the Bellman equations, that hold under the conditions described before Eq. (7.1). To make sure these are met, we assume that the costs are strictly positive.

Assumption 7.1.1. *All costs are positive, i.e., there exists $c_{min} > 0$ such that $c^k(s,a) \geq c_{min}$ for every k and $(s,a) \in \mathscr{S} \times \mathscr{A}$.*

We can easily eliminate Assumption 7.1.1 by applying a perturbation to the instantaneous costs. That is, instead of c^k we use the cost function $\tilde{c}^k(s,a) = \max\{c^k(s,a), \varepsilon\}$ for some $\varepsilon > 0$. This ensures that the effective minimal cost is $c_{\min} = \varepsilon$, at the price of introducing additional bias. Choosing $\varepsilon = \Theta(K^{-1/4})$ ensures that all our algorithms obtain regret bounds of $\widetilde{O}(K^{3/4})$ in the general case.

Occupancy Measures. Every policy π induces an occupancy measure $q^{\pi} : \mathscr{S} \times \mathscr{A} \to [0,\infty]$ such that $q^{\pi}(s,a)$ is the expected number of times to visit state s and take action a when playing according to π, i.e.,

$$q^{\pi}(s,a) = \lim_{T \to \infty} \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a\} \mid P, \pi, s_1 = s_{\text{init}}\Big],$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Note that for a proper policy π, $q^\pi(s,a)$ is finite for every (s,a). In fact, the correspondence between proper policies and finite occupancy measures is 1-to-1, and its inverse[1] for q is given by $\pi^q(a \mid s) = \frac{q(s,a)}{q(s)}$ where $q(s) = \sum_{a \in \mathscr{A}} q(s,a)$ is the expected number of visits to s. The equivalence between policies and occupancy measures is well-known for MDPs (see, e.g., [ZN13]), but also holds for SSPs by linear programming formulation [Man60]. Notice that the expected cost of policy π is linear w.r.t q^π, i.e.,

$$\begin{aligned} V_k^{\pi_k}(s_{\text{init}}) &= \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) \mid P, \pi_k, s_1 = s_{\text{init}}\Big] \\ &= \sum_{s \in S} \sum_{a \in A} q^{\pi_k}(s,a) c^k(s,a) \overset{\text{def}}{=} \langle q^{\pi_k}, c^k \rangle. \end{aligned}$$

Thus, minimizing the expected regret can be written as an instance of online linear optimization in the following manner,

$$\begin{aligned} \mathbb{E}[R_K] &= \mathbb{E}\Big[\sum_{k=1}^{K} V_k^{\pi_k}(s_{\text{init}}) - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}})\Big] \\ &= \mathbb{E}\Big[\sum_{k=1}^{K} \langle q^{\pi_k} - q^{\pi^\star}, c^k \rangle\Big]. \end{aligned}$$

7.2 Known Transition Function

We start with the simpler (yet surprisingly challenging) case where P is known to the learner. Recall that while the transition function is known, the costs change arbitrarily between episodes. In Section 7.2.1 we establish the implementation of the OMD method in SSP, and in Section 7.2.2 we use it to obtain a high probability regret bound.

7.2.1 Online Mirror Descent for SSP

Online mirror descent is a popular framework for OCO and its application to occupancy measures yields the O-REPS algorithms [ZN13, RM19a, RM19b, JJL$^+$20]. Usually these algorithms operate w.r.t to the set of all occupancy measures (which corresponds to the set of all policies), but a naive application of this kind fails in SSP because it does not

[1] If $q(s) = 0$ for some state s then the inverse mapping is not well-defined. However, since s will not be reached, we can pick the action there arbitrarily. More precisely, the correspondence holds when restricting to reachable states.

guarantee that the learner plays proper policies. For example, in the first episode these algorithms play the uniform policy which may suffer exponential cost.

Thus, we propose to apply OMD to the set $\Delta(\mathscr{M})(\tau)$ – occupancy measures of policies π that reach the goal in expected time $T^{\pi}(s_{\text{init}}) \leq \tau$. This set is convex and has a compact representation as we show shortly. Our algorithm SSP-O-REPS operates as follows. In the beginning of episode k, it picks an occupancy measure q_k from $\Delta(\mathscr{M})(\tau)$ which minimizes a trade-off between the current cost function and the distance to the previously chosen occupancy measure. Then, it extracts the policy $\pi_k = \pi^{q_k}$ and plays it through the episode. Formally,

$$q_k = q^{\pi_k} = \arg\min_{q \in \Delta(\mathscr{M})(\tau)} \eta \langle q, c_{k-1} \rangle + \mathrm{KL}(q \parallel q_{k-1}), \tag{7.2}$$

where $\mathrm{KL}(\cdot||\cdot)$ is the KL-divergence, and $\eta > 0$ is a learning rate. Computing q_k is implemented in two steps: first find the unconstrained minimizer and then project it into $\Delta(\mathscr{M})(\tau)$, i.e.,

$$q'_k = \arg\min_q \eta \langle q, c_{k-1} \rangle + \mathrm{KL}(q \parallel q_{k-1}) \tag{7.3}$$

$$q_k = \arg\min_{q \in \Delta(\mathscr{M})(\tau)} \mathrm{KL}(q \parallel q'_k). \tag{7.4}$$

Eq. (7.3) has a closed form $q'_k(s,a) = q_{k-1}(s,a)e^{-\eta c_{k-1}(s,a)}$, and Eq. (7.4) can be formalized as a constrained convex optimization problem with the following linear constraints:

$$\begin{aligned} \forall s.\ \sum_{a \in \mathscr{A}} q(s,a) - \sum_{s' \in \mathscr{S}} \sum_{a' \in \mathscr{A}} q(s',a')P(s|s',a') &= \mathbb{I}\{s = s_{\text{init}}\} \\ \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} q(s,a) &\leq \tau, \end{aligned} \tag{7.5}$$

where we omitted non-negativity constraints. The first set of constraints are standard flow constraints, while the novel constraint (7.5) ensures that $T^{\pi^q}(s_{\text{init}}) \leq \tau$.

Finally, we need to pick the parameter τ. While it needs to upper bound $T^{\pi^\star}(s_{\text{init}})$ in order to have $q^{\pi^\star} \in \Delta(\mathscr{M})(\tau)$, we want it to be as small as possible to get tighter regret guarantees. To that end, define the SSP-diameter [TGV+20] $D = \max_{s \in \mathscr{S}} \min_{\pi \in \Pi_{\text{proper}}} T^{\pi}(s)$ and pick $\tau = D/c_{\min}$. The diameter can be computed efficiently by finding the optimal policy w.r.t the constant cost function $c(s,a) = 1$. We refer to this policy as the fast policy π^f, and it holds that $D = \max_{s \in S} T^{\pi^f}(s)$.

Indeed $q^{\pi^\star} \in \Delta(\mathscr{M})(D/c_{\min})$ because the total cost of the best policy in hindsight in

K episodes is upper bounded by the total cost of any other policy, e.g., the fast policy (which is at most DK), and is lower bounded by the expected time of $\pi^\star$ times the minimal cost, i.e., $V_k^{\pi^\star}(s_{\text{init}}) \geq c_{\min} T^{\pi^\star}(s_{\text{init}})$. In the supplementary material we also show that this choice of τ cannot be smaller in general.

In the supplementary material we provide the full analysis of the algorithm yielding the following regret bound in expectation. Moreover, we show that all the chosen policies must be proper and therefore the goal is reached with probability 1 in all episodes.

Theorem 7.2.1. *Under Assumption 5.2.1, the expected regret of SSP-O-REPS with known transition function and* $\eta = \widetilde{\Theta}(\frac{1}{\sqrt{K}})$ *is*

$$\mathbb{E}[R_K] \leq O\Big(\frac{D}{c_{min}}\sqrt{K \log \frac{DSA}{c_{min}}}\Big) = \widetilde{O}\Big(\frac{D}{c_{min}}\sqrt{K}\Big).$$

7.2.2 High Probability Regret Bound

To obtain high probability regret bounds, we must control the deviation between the learner's suffered cost and its expected value. While this is easily achievable in the finite-horizon setting through an application of Azuma inequality, it appears a major challenge in SSP since there is no finite upper bound on the learner's cost. In fact, the supplementary material illustrates a simple example with 0 expected regret, but constant probability to suffer large regret (linear in K). The idea here is that even though a policy has small cost in expectation, there might be a tiny probability that it suffers huge cost (this cannot happen in finite-horizon since the cost is always bounded by H). Finally, even an event with tiny probability will happen at least once if there is a large number of episodes K.

Our strategy to control the deviation between the learner's actual suffered cost and its expected value is based on the observation that this quantity is closely related to the expected time to reach the goal from any state. This is illustrated by the following lemma whose proof is based on an adaptation of Azuma inequality to unbounded martingales (Theorem E.12.5) which may be of independent interest.

Lemma 7.2.2. *Assume that in each episode k the learner plays a strategy σ_k such that the expected time to reach the goal from any state is at most τ. Then, with probability at least*

$1-\delta$,

$$\sum_{k=1}^{K}\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) \le \sum_{k=1}^{K} \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) \mid P,\sigma_k,s_1^k=s_{init}\Big] + O\Big(\tau\sqrt{K\log^3\frac{K}{\delta}}\Big).$$

Thus, bounding the regret in high probability boils down to guaranteeing that $T^{\pi_k}(s) \le D/c_{\min}$ for all $s \in S$ and not just s_{init}. Unfortunately, these constraints admit a non-convex set of occupancy measures. To bypass this issue we propose the SSP-O-REPS2 algorithm that operates as follows: start every episode k by playing the policy π_k chosen by SSP-O-REPS (i.e., Eq. (7.2)), but once we reach a state s whose expected time to the goal is too long (i.e., $T^{\pi_k}(s) \ge D/c_{\min}$), switch to the fast policy π^f.

Now the conditions of Lemma 7.2.2 are clearly met, so it remains to relate the expected cost of our new strategy σ_k to this of π_k. The key novelty of our mid-episode policy switch is the timing. The naive approach would be to perform the switch when the policy takes too long, but then there is no way to bound the excess cost when compared to that of π_k. Performing the switch only once a “bad” state is reached ensures that the expected cost of σ_k can only be better than π_k.

Theorem 7.2.3. *Under Assumption 5.2.1, with probability* $1-\delta$*, the regret of SSP-O-REPS2 with known transition function is*

$$R_K \le O\Big(\frac{D}{c_{min}}\sqrt{K\log^3\frac{KDSA}{\delta c_{min}}}\Big) = \widetilde{O}\Big(\frac{D}{c_{min}}\sqrt{K}\Big).$$

7.3 Unknown Transition Function

A standard technique to deal with unknown transition function in adversarial MDPs is to use optimistic estimates of P. We follow this approach but, as in the known transitions case, crucial modifications are necessary to apply optimism and obtain regret guarantees. In this section we describe our SSP-O-REPS3 algorithm for unknown transitions.

We start by describing the confidence sets and transition estimates used by the algorithm. SSP-O-REPS3 proceeds in *epochs* and updates the confidence set at the beginning of every epoch. The first epoch begins at the first time step, and an epoch ends once an episode ends or the number of visits to some state-action pair is doubled. Denote by $N^e(s,a)$ the number of visits to (s,a) up to (and not including) epoch e, and by

$N^e(s,a,s')$ the number of times this was followed by a transition to s'. Let $N^e_+(s,a) = \max\{N^e(s,a),1\}$ and define the empirical transition function for epoch e by $\bar{P}_e(s'|s,a) = N^e(s,a,s')/N^e_+(s,a)$. Finally, define the confidence set for epoch e as the set of all transition functions P' such that for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times (\mathscr{S} \cup \{g\})$,

$$|P'(s' \mid s,a) - \bar{P}_e(s' \mid s,a)| \leq \varepsilon_e(s' \mid s,a),$$

where $\varepsilon_e(s'|s,a) = 4\sqrt{\bar{P}_e(s'|s,a)A^e(s,a)} + 28A^e(s,a)$ is the confidence set radius for $A^e(s,a) = \frac{\log\left(SAN^e_+(s,a)/\delta\right)}{N^e_+(s,a)}$. By Bernstein inequality (see, e.g., [AOM17]), these confidence sets contain P with probability $1-\delta$ for all epochs.

Next, we extend our OMD implementation to the unknown transitions case. We follow the elegant approach of [RM19a] that use occupancy measures that are extended to include a transition function as well, that is,

$$q^{P,\pi}(s,a,s') = \lim_{T\to\infty} \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a, s_{t+1} = s'\}\Big],$$

where $\mathbb{E}[\cdot]$ is shorthand for $\mathbb{E}[\cdot \mid P, \pi, s_1 = s_{\text{init}}]$ here. Now an occupancy measure q corresponds to a transition function-policy pair with the inverse mapping given by

$$\pi^q(a \mid s) = \frac{q(s,a)}{q(s)} \quad ; \quad P^q(s' \mid s,a) = \frac{q(s,a,s')}{q(s,a)},$$

where $q(s,a) = \sum_{s' \in \mathscr{S} \cup \{g\}} q(s,a,s')$ is the expected number of visits to (s,a) w.r.t P^q when playing π^q. We extend the set $\Delta(\mathscr{M})(\tau)$ (which we cannot compute without knowing P), and perform OMD on the set $\widetilde{\Delta(\mathscr{M})}_e(\tau)$ that changes through epochs. $\widetilde{\Delta(\mathscr{M})}_e(\tau)$ is defined as the set of occupancy measures q whose induced transition function P^q is in the confidence set of epoch e and the expected time of π^q (w.r.t P^q) from s_{init} to the goal is at most τ. This set is again convex with a compact representation, and it admits the following OMD update step,

$$q_k = q^{P_k,\pi_k} = \arg \min_{q \in \widetilde{\Delta(\mathscr{M})}_{e(k)}(\tau)} \eta \langle q, c_{k-1} \rangle + \mathrm{KL}(q \parallel q_{k-1}), \tag{7.6}$$

where $e(k)$ denotes the first epoch in episode k. Similarly to the known transitions case, this update can be performed efficiently.

In contrast to the known transitions case, this version of OMD cannot even guarantee bounded regret in expectation, because without knowledge of the transition function there is no guarantee that the chosen policies are even proper. Note that in the easier loop-free SSP setting, this OMD version is enough to guarantee a high probability regret bound even with unknown transitions. We now describe the mechanisms that need to be combined with OMD to obtain our regret bound.

Similarly to Section 7.2.2, we must make sure that the learner does not take too much time to reach the goal. The problem now is that we cannot compute its expected time T^{π_k} since P is unknown. Instead, we use the expected time of π_k w.r.t P_k (denoted by $\widetilde{T}_k^{\pi_k}$) which is an estimate of T^{π_k}, but not necessarily an optimistic one. Once a state s is reached such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$ we want to switch to the fast policy π^f which again cannot be computed without knowing P. This policy is replaced with its optimistic estimate $\widetilde{\pi}_e^f$, which we refer to as the optimistic fast policy. Together with the optimistic fast transition function $\widetilde{P}_e^f$, this policy minimizes the expected time to the goal out of all pairs of policies and transition functions from the confidence set of epoch e.

If we were in the known transitions case, this would have been enough. So it seems that it should also suffice with unknown transitions, if we recompute the optimistic fast policy in the end of every epoch similarly to [RCMK20]. However, in the adversarial setting this approach fails for two main reasons. First, we cannot guarantee that $\widetilde{T}_k^{\pi_k}$ is a good enough estimate of T^{π_k} in all states. Second, the learner's policy is stochastic which means that we cannot guarantee all actions are being explored enough (as opposed to [RCMK20] that only play deterministic policies since they do not tackle adversarial costs). To overcome these challenges, we propose to force exploration in the following manner. Define a state to be *unknown* until every action was played at least $\Phi = \alpha \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$ times in this state (for some constant $\alpha > 0$), and *known* afterwards. When reaching an unknown state, we play the least played action so far (forcing exploration), and only then switch to the optimistic fast policy. The idea behind this forced exploration is inspired by [RCMK20] that show that once all states are known, the optimistic fast policy is proper with high probability.

To summarize, SSP-O-REPS3 operates as follows. We start each episode k by playing the policy π_k computed in Eq. (7.6), and maintain confidence sets that are updated at the beginning of every epoch. When we reach a state s such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$, we switch to the optimistic fast policy. In addition, when an unknown state is reached we play the least played action up to this point and then switch to the optimistic fast policy. Finally, we also make the switch to the optimistic fast policy once the number of visits to some state-action pair is doubled, at which point we also recompute it.

Theorem 7.3.1. *Under Assumption 5.2.1, with probability* $1-\delta$*, the regret of SSP-O-*

REPS3 with known SSP-diameter D is

$$R_K \leq \widetilde{O}\Big(\frac{DS}{c_{min}}\sqrt{AK} + \frac{D^2 S^2 A}{c_{min}^2}\Big) = \widetilde{O}\Big(\frac{DS}{c_{min}}\sqrt{AK}\Big),$$

where the last equality holds for $K \geq D^2 S^2 A / c_{min}^2$.

Our analysis builds on ideas from [RCMK20] that analyze optimistic algorithms in SSP with stochastic costs. However, for the many reasons described in this paper and because our algorithm is not optimistic, many novel technical adaptions are needed in order to tackle the new challenges that arise when both the costs are adversarial and the transition function is unknown.

Recall that the learner has two objectives in SSP: minimizing cost and reaching the goal. When transitions were known, we used Lemma 7.2.2 to say that (with high probability) the goal is reached in every episode, and then we could simply focus on bounding the regret. With unknown transitions, the argument for bounding the total time becomes more involved. The idea is that (with high probability) the number of steps between policy switches cannot be too long, as a consequence of our added mechanisms. To that end, we split the time steps into *intervals*. The first interval begins at the first time step, and an interval ends once (1) an episode ends, (2) an epoch ends, (3) an unknown state is reached, or (4) a policy switch is made due to reaching a "bad" state. Intuitively, we bound the length of every interval by $\widetilde{O}(D/c_{\min})$ with high probability, and then use fact that the number of intervals is bounded by $\widetilde{O}(K + DS^2A/c_{\min}^2)$ to bound the total time. Then, we show that the regret of the learner can be bounded by the regret of OMD (analyzed in Section 7.2) plus the square root of the total variance (times S^2A). Finally, we obtain our regret bound by noticing that the total variance is equal to the variance in each interval times the number of intervals, and bounding the variance in an interval by $O(D^2/c_{\min}^2)$.

Estimating the SSP-diameter. When the transition function is unknown, we cannot compute the diameter D. However, a careful look at our algorithms shows that we use it only twice. First, we pick $\tau = D/c_{\min}$ as an upper bound on the expected time of the best policy in hindsight. For this purpose it is enough to use $T^{\pi^f}(s_{\text{init}})/c_{\min}$, and therefore we shall dedicate the first L episodes to computing an estimate $\widetilde{D}(s_{\text{init}})$ of $T^{\pi^f}(s_{\text{init}})$ before running SSP-O-REPS3. Second, D is used to make a switch when a "bad" or unknown state s is reached, but again it is enough to use $T^{\pi^f}(s)$ instead. Similarly, we use the first L visits to s to estimate $T^{\pi^f}(s)$ and then continue executing the algorithm with $\widetilde{D}(s)$ instead of D.

To compute $\widetilde{D}(s)$ we run the algorithm of [RCMK20] for regret minimization in SSP with constant cost of 1 (since it measures time). By their regret bound, we can set $L \approx \sqrt{K}$ and suffer negligible additional regret. This is also enough to yield the two properties we need in order to keep the same regret bound (with high probability): $\widetilde{D}(s)$ is an upper bound on $T^{\pi^f}(s)$ for any $s \in S$, and $\widetilde{D}(s) \leq O(D)$ (i.e., it is not too large).

8 Conclusions and Future Work

In this thesis we provided new algorithms and theory for tackling fundamental issues that hurt the performance of reinforcement learning algorithms in many real-world applications. We focused on three main issues: exploration, non-stationarity and inaccurate models. Our algorithms face these challenges in several environments, and their performance is analyzed in terms of the regret – the difference between the cumulative cost of the agent through the learning process and the expected cost of the best policy in hindsight.

We started by studying adversarial MDPs that model non-stationary environments through adversarially changing costs. We presented the first high-probability regret bounds for adversarial MDPs with unknown transitions and full-information feedback, and the first regret bounds for the case of bandit feedback (and unknown transitions).

Then, we studied the stochastic shortest path model that captures a wide variety of realistic scenarios and includes the discounted return model and the finite-horizon model as special cases. We presented the first near-optimal regret bounds for SSP, and then developed an improved algorithm based on a reduction to the finite-horizon setting that is able to achieve optimal regret (up to logarithmic factors).

Finally, we also combined the two models and presented the adversarial MDP model. This general model is able to better capture the challenges of applying RL to real-world applications, and we provide the first regret minimization algorithms in this setting.

There are many interesting open questions left for future work. The most important direction is extending our regret minimization algorithms to settings where the state space is huge (even infinite) and function approximation must be used. While research has already begun for factored MDPs and linear function approximation [JYWJ20, ZBB$^+$20, ZLKB20, RM21a, VPSS22, MHWG22, CJL22], it is just the tip of the iceberg. For adversarial MDPs, the prominent open question that remains is whether the minimax optimal regret scales with the number of states linearly or via square root. Another important question is whether policy optimization methods (that are highly successful in practice) are able to obtain optimal regret, or is it possible only with occupancy measure based methods. This question was partly answered by [LWL21, CLR22], but it is still unclear if these methods can be parameter-free in the SSP setting, for example. More interesting future directions tackle best-of-both-worlds, delay, cooperation and privacy [JL20, JHL21, LRM22b, JLL$^+$22, LRM22a].

References

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[AJO09] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

[AOM17] 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.

[Ber95] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[BT91] Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

[BT02] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[BT09] Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[BY13] Dimitri P Bertsekas and Huizhen Yu. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909, MIT*, 2013.

[CBC$^+$19] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CEMR21] Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34:28350–28361, 2021.

[CJL22] Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear mdp. In *International Conference on Machine Learning*, pages 3204–3245. PMLR, 2022.

[CL21] Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: the adversarial cost and unknown transition case. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1651–1660. PMLR, 2021.

[CLR22] Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 982–1046. PMLR, 2022.

[CLW21] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1180–1215. PMLR, 2021.

[DLWB19] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

[EKM09] Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov Decision Processes. *Math. Oper. Res.*, 34(3):726–736, 2009. (preliminary version NIPS 2004).

[EMGM19] Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 12203–12213, 2019.

[EMM21] Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7288–7295. AAAI Press, 2021.

[EMSM21] Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2937–2947. PMLR, 2021.

[JAZBJ18] 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.

[JHL21] Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

[JJL$^+$20] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 2020.

[JL20] Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33:16557–16566, 2020.

[JLL$^+$22] Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13172*, 2022.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[KS02] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

[KV03] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. In *16th Annual Conference on Computational Learning Theory (COLT)*, pages 26–40, 2003.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[LFDA16] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[LNSL^{+}12] Huitan Lei, Inbal Nahum-Shani, Kevin Lynch, David Oslin, and Susan A Murphy. A âsmartâ design for building individualized treatment sequences. *Annual review of clinical psychology*, 8, 2012.

[LRM22a] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Cooperative online learning in stochastic and adversarial mdps. *arXiv preprint arXiv:2201.13170*, 2022.

[LRM22b] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7281–7289, 2022.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LWL21] Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

[Man60] Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[MHWG22] Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. In *International Conference on Machine Learning*, pages 15584–15629. PMLR, 2022.

[MKS^{+}15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[MLL^{+}14] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

[NGS10] Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 231–243, 2010.

[NGS12] Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813, 2012.

[NGSA14] Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov Decision Processes under bandit feedback. *IEEE Trans. Automat. Contr.*, 59(3):676–691, 2014.

[OVR16] Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

[Put14] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[RCMK20] Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

[RM19a] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486, 2019.

[RM19b] Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2209–2218, 2019.

[RM21a] Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored mdps with unknown structure. *Advances in Neural Information Processing Systems*, 34:11148–11159, 2021.

[RM21b] Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2936–2942. ijcai.org, 2021.

[SB98] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[SERM20] Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

[Sha12] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[Sli19] Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

[SLKW02] Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.

[SSS$^+$17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[TGV$^+$20] Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, 2020.

[VPSS22] Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pages 22203–22233. PMLR, 2022.

[WOS$^+$03] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[YMS09] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov Decision Processes with arbitrary reward processes. *Math. Oper. Res.*, 34(3):737–757, 2009.

[ZB19] 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.

[ZBB$^+$20] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[ZJD21] 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.

[ZLKB20] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[ZN13] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1583–1591, 2013.

A Supplementary Material for Chapter 3

A.1 Proof of Theorem 3.4.2 Cont.

In the proof of Theorem 3.4.2 we showed that the following optimization problem

$$q^{k+1} = \arg \min_{q \in \Delta(\mathscr{M}, i(k))} \mathrm{KL}(q \parallel \tilde{q}^{k+1})$$

can be reformulated as the following convex optimization problem (from now on we use k instead of i):

$$\begin{aligned}
&\min_q \mathrm{KL}(q \parallel \tilde{q}^{k+1}) \\
&s.t. \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h(s,a,s') = 1 \quad \forall h = 1, \ldots, H \\
&\quad \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_h(s,a,s') = \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_{h-1}(s',a,s) \quad \forall h = 2, \ldots, H \quad \forall s \in \mathscr{S} \\
&\quad q_h(s,a,s') - \bar{P}_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_{(}s,a,y) \leq \varepsilon_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_{(}s,a,y) \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H] \\
&\quad \bar{P}_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_h(s,a,y) - q_h(s,a,s') \leq \varepsilon_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_h(s,a,y) \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H] \\
&\quad q_h(s,a,s') \geq 0 \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H].
\end{aligned}$$

Now we will derive the solution to this problem using Lagrange multipliers. First we write the Lagrangian with $\lambda, \beta, \mu, \mu^+, \mu^-$ as Lagrange multipliers. Notice that we omit the non-negativity constraints, which we can justify since the solution will be non-negative anyway.

$$
\begin{aligned}
\mathscr{L}(q) = \mathrm{KL}(q \parallel \tilde{q}^{k+1}) &+ \sum_{h=1}^{H} \lambda_h \left(\sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h(s,a,s') - 1 \right) \\
&+ \sum_{h=2}^{H} \sum_{s\in\mathscr{S}} \beta_h(s) \left(\sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h(s,a,s') - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_{h-1}(s',a,s) \right) \\
&+ \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \mu_h^+(s,a,s') \left(q_h(s,a,s') - \left(\varepsilon_h^k(s'|s,a) + \bar{P}_h^k(s'|s,a) \right) \sum_{y\in\mathscr{S}} q_h(s,a,y) \right) \\
&+ \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \mu_h^-(s,a,s') \left(-q_h(s,a,s') - \left(\varepsilon_h^k(s'|s,a) - \bar{P}_h^k(s'|s,a) \right) \sum_{y\in\mathscr{S}} q_h(s,a,y) \right).
\end{aligned}
$$

Let $(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$ and consider the derivative with respect to $q_h(s,a,s')$. We denote $\beta_1(s) = 0$ to avoid addressing the edge cases explicitly.

$$
\begin{aligned}
\frac{\partial \mathscr{L}}{\partial q_h(s,a,s')} &= \log q_h(s,a,s') - \log \tilde{q}_h^{k+1}(s,a,s') + \lambda_h + \beta_h(s) - \beta_{h+1}(s') \\
&+ \mu_h^+(s,a,s') - \mu_h^-(s,a,s') \\
&- \sum_{y\in\mathscr{S}} \left(\varepsilon_h^k(s'|s,a) + \bar{P}_h^k(s'|s,a) \right) \mu_h^+(s,a,y)) \\
&- \sum_{y\in\mathscr{S}} \left(\varepsilon_h^k(s'|s,a) - \bar{P}_h^k(s'|s,a) \right) \mu_h^-(s,a,y)).
\end{aligned}
$$

We define the following value function v and error function e parameterized by μ and β, and an estimated Bellman error.

$$
\begin{aligned}
v_h^{\mu}(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\
e_h^{\mu,\beta}(s,a,s') &= \sum_{y\in\mathscr{S}} (\mu_h^-(s,a,y) + \mu_h^+(x,a,y)) \varepsilon_h^k(s' \mid s,a) + \beta_{h+1}(s') - \beta_h(s) \\
B_h^k(s,a,s' \mid v,e) &= e_h(s,a,s') + v_h(s,a,s') - \eta c_h^k(s,a) - \sum_{y\in\mathscr{S}} \bar{P}_h^k(y \mid s,a) v_{h+1}(s,a,y).
\end{aligned}
$$

So the derivative becomes:

$$
\begin{aligned}
\frac{\partial \mathscr{L}}{\partial q_h(s,a,s')} &= \log \frac{q_h(s,a,s')}{\tilde{q}_h^{k+1}(s,a,s')} + \lambda_h - v_h^{\mu}(s,a,s') + \sum_{y\in\mathscr{S}} \bar{P}_h^k(y \mid s,a) v_{h+1}^{\mu}(s,a,y) \\
&= \log q_h(s,a,s') - \log \tilde{q}_h^{k+1}(s,a,s') + \lambda_h - \eta c_h^k(s,a) - B_h^k(s,a,s' \mid v^{\mu}, e^{\mu,\beta}).
\end{aligned}
$$

Setting the gradient to zero and using the explicit form of $\tilde{q}_h^{k+1}(s,a,s')$ we obtain

$$\begin{aligned} q_h^{k+1}(s,a,s') &= \tilde{q}_h^{k+1}(s,a,s')e^{-\lambda_h+\eta c_h^k(s,a)+B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})} \\ &= q_h^k(s,a,s')e^{-\eta c_h^k(s,a)}e^{-\lambda_h+\eta c_h^k(s,a)+B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})} \\ &= q_h^k(s,a,s')e^{-\lambda_h+B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})}. \end{aligned}$$

We can use the first constraint to discover that λ_h is a normalizer for every $h = 1,\dots,H$, i.e.

$$\begin{aligned} 1 &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} q_h^{k+1}(s,a,s') \\ 1 &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} q_h^k(s,a,s')e^{-\lambda_h+B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})} \\ e^{\lambda_h} &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} q_h^k(s,a,s')e^{B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})}, \end{aligned}$$

so defining $Z_h^k(v,e) = \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} q_h^k(s,a,s')e^{B_h^k(s,a,s'|v,e)}$, we obtain

$$q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s')e^{B_h^k(s,a,s'|v^\mu,e^{\mu,\beta})}}{Z_h^k(v^\mu,e^{\mu,\beta})}.$$

Now to find β and μ we consider the dual problem. Substituting q^{k+1} back into $\mathscr{L}$ we obtain the following dual problem.

$$\max_{\beta,\mu\geq 0}\min_q \mathscr{L}(q) = \max_{\beta,\mu\geq 0}\mathscr{L}(q^{k+1}) = \max_{\beta,\mu\geq 0} -\sum_{h=1}^H \log Z_h^k(v^\mu,e^{\mu,\beta}) - 1 + \sum_{h,s,a,s'}\tilde{q}_h^{k+1}(s,a,s').$$

So after ignoring constants we observe that

$$\beta^k,\mu^k = \arg\min_{\beta,\mu\geq 0}\sum_{h=1}^H \log Z_h^k(v^\mu,e^{\mu,\beta}).$$

A.2 Proof of Theorem 3.5.2

We start with a value difference lemma [SERM20]:

$$\begin{aligned}
\hat{R}^{APP}_{1:K} &= \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}};c^k,P^k) - \min_{\pi} \sum_{k=1}^{K} V_1^{\pi}(s_{\text{init}};c^k,P) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\left[\langle P_h(\cdot \mid s_h^k,a_h^k) - P_h^k(\cdot \mid s_h^k,a_h^k), V_{h+1}^{\pi^k}(\cdot;c^k,P^k)\rangle \mid s_1^k = s_{\text{init}}, P, \pi^k\right] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathscr{S}} \mathbb{E}\left[|P_h(s' \mid s_h^k,a_h^k) - P_h^k(s' \mid s_h^k,a_h^k)|V_{h+1}^{\pi^k}(s';c^k,P^k) \mid s_1^k = s_{\text{init}}, P, \pi^k\right] \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} q_h^{\pi^k}(s,a)|P_h(s' \mid s_h^k,a_h^k) - P_h^k(s' \mid s_h^k,a_h^k)|V_{h+1}^{\pi^k}(s';c^k,P^k) \\
&\leq H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_h^{\pi^k}(s,a)\|P_h(\cdot \mid s_h^k,a_h^k) - P_h^k(\cdot \mid s_h^k,a_h^k)\|_1.
\end{aligned}$$

Now, we use the fact that the confidence sets contain P with probability $1-\delta$ to obtain:

$$\begin{aligned}
\hat{R}^{APP}_{1:K} &\leq H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_h^{\pi^k}(s,a)\min\{2,\varepsilon_h^k(s,a)\} \\
&\leq H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} \mathbb{I}\{s_h^k=s,a_h^k=a\}\min\{2,\varepsilon_h^k(s,a)\} + 10H\sqrt{K\log\frac{KHSA}{\delta}} \\
&\leq 10H\sqrt{S\log\frac{KHSA}{\delta}}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} \frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\max\{1,N_h^k(s,a)\}} + 10H\sqrt{K\log\frac{KHSA}{\delta}},
\end{aligned}$$

where the second inequality follows by Azuma inequality.

Finally, by [JOA10] we have that:

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} \frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\max\{1,N_h^k(s,a)\}} &\leq 3\sum_{h=1}^{H}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sqrt{N_h^K(s,a)} \\
&\leq 3H\sqrt{SAK},
\end{aligned}$$

where the last inequality follows by Jensen inequality.

B Supplementary Material for Chapter 4

B.1 Efficient Implementation of "Bounded Bandit UC-O-REPS"

In each episode we need to perform (while maintaining confidence sets):

$$\tilde{q}^{k+1} = \arg\min_{q} \eta \langle q, \hat{c}^k \rangle + \mathrm{KL}(q \parallel q^k) \tag{B.1}$$

$$q^{k+1} = \arg \min_{q \in \Delta_\alpha(\mathscr{M}, i(k))} \mathrm{KL}(q \parallel \tilde{q}^{k+1}). \tag{B.2}$$

The confidence sets are maintained like in the original UC-O-REPS algorithm, and step (B.1) can be easily solved by setting $\tilde{q}_h^{k+1}(s,a,s') = q_h^k(s,a,s')e^{-\eta \hat{c}_h^k(s,a)}$ for every $(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$. Thus, we need to describe an efficient implementation to step (B.2). This step can be reformulated as the following constrained convex optimization problem (from now on we use k instead of i):

$$\begin{aligned}
&\min_{q} \mathrm{KL}(q \parallel \tilde{q}^{k+1}) \\
&s.t. \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h(s,a,s') = 1 \quad \forall h = 1, \dots, H \\
&\quad \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_h(s,a,s') = \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_{h-1}(s',a,s) \quad \forall h = 2, \dots, H \quad \forall s \in \mathscr{S} \\
&\quad q_h(s,a,s') - \bar{P}_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_(s,a,y) \le \varepsilon_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_(s,a,y) \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H] \\
&\quad \bar{P}_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_h(s,a,y) - q_h(s,a,s') \le \varepsilon_h^k(s'|s,a) \sum_{y \in \mathscr{S}} q_h(s,a,y) \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H] \\
&\quad \sum_{s' \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_h(s,a,s') \ge \alpha \quad \forall h = 1, \dots, H \quad \forall s \in \mathscr{S} \\
&\quad q_h(s,a,s') \ge 0 \quad \forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H].
\end{aligned}$$

Now we will derive the solution to this problem using Lagrange multipliers. First we write the Lagrangian with $\lambda, \beta, \mu, \mu^+, \mu^-, g$ as Lagrange multipliers. Notice that we omit the non-negativity constraints, which we can justify since the solution will be non-negative anyway.

$$
\begin{aligned}
\mathscr{L}(q) = \mathrm{KL}(q \parallel \tilde{q}^{k+1}) &+ \sum_{h=1}^{H} \lambda_h \left(\sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h(s,a,s') - 1 \right) \\
&+ \sum_{h=2}^{H} \sum_{s\in\mathscr{S}} \beta_h(s) \left(\sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h(s,a,s') - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_{h-1}(s',a,s) \right) \\
&+ \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \mu_h^+(s,a,s') \left(q_h(s,a,s') - \left(\varepsilon_h^k(s'|s,a) + \bar{P}_h^k(s'|s,a) \right) \sum_{y\in\mathscr{S}} q_h(s,a,y) \right) \\
&+ \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \mu_h^-(s,a,s') \left(-q_h(s,a,s') - \left(\varepsilon_h^k(s'|s,a) - \bar{P}_h^k(s'|s,a) \right) \sum_{y\in\mathscr{S}} q_h(s,a,y) \right) \\
&+ \sum_{h=1}^{H} \sum_{s\in\mathscr{S}} g_h(s) \left(\alpha - \sum_{s'\in\mathscr{S}} \sum_{a\in\mathscr{A}} q_h(s,a,s') \right).
\end{aligned}
$$

Let $(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$ and consider the derivative with respect to $q_h(s,a,s')$. We denote $\beta_1(s) = 0$ to avoid addressing the edge cases explicitly.

$$
\begin{aligned}
\frac{\partial \mathscr{L}}{\partial q_h(s,a,s')} &= \log q_h(s,a,s') - \log \tilde{q}_h^{k+1}(s,a,s') + \lambda_h + \beta_h(s) - \beta_{h+1}(s') \\
&+ \mu_h^+(s,a,s') - \mu_h^-(s,a,s') - g_h(s) \\
&- \sum_{y\in\mathscr{S}} \left(\varepsilon_h^k(s'|s,a) + \bar{P}_h^k(s'|s,a) \right) \mu_h^+(s,a,y)) \\
&- \sum_{y\in\mathscr{S}} \left(\varepsilon_h^k(s'|s,a) - \bar{P}_h^k(s'|s,a) \right) \mu_h^-(s,a,y)).
\end{aligned}
$$

We define the following value function v and error function e parameterized by μ and β, and an estimated Bellman error.

$$
\begin{aligned}
v_h^{\mu}(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\
e_h^{\mu,\beta,g}(s,a,s') &= \sum_{y\in\mathscr{S}} (\mu_h^-(s,a,y) + \mu_h^+(x,a,y)) \varepsilon_h^k(s' \mid s,a) + \beta_{h+1}(s') - \beta_h(s) + g_h(s) \\
B_h^k(s,a,s' \mid v,e) &= e_h(s,a,s') + v_h(s,a,s') - \eta \hat{c}_h^k(s,a) - \sum_{y\in\mathscr{S}} \bar{P}_h^k(y \mid s,a) v_{h+1}(s,a,y).
\end{aligned}
$$

So the derivative becomes:

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial q_h(s,a,s')} &= \log \frac{q_h(s,a,s')}{\tilde{q}_h^{k+1}(s,a,s')} + \lambda_h - v_h^{\mu}(s,a,s') + \sum_{y\in\mathscr{S}} \bar{P}_h^k(y \mid s,a) v_{h+1}^{\mu}(s,a,y) \\ &= \log q_h(s,a,s') - \log \tilde{q}_h^{k+1}(s,a,s') + \lambda_h - \eta \hat{c}_h^k(s,a) - B_h^k(s,a,s' \mid v^{\mu}, e^{\mu,\beta,g}).\end{aligned}$$

Setting the gradient to zero and using the explicit form of $\tilde{q}_h^{k+1}(s,a,s')$ we obtain

$$\begin{aligned}q_h^{k+1}(s,a,s') &= \tilde{q}_h^{k+1}(s,a,s') e^{-\lambda_h + \eta \hat{c}_h^k(s,a) + B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})} \\ &= q_h^k(s,a,s') e^{-\eta \hat{c}_h^k(s,a)} e^{-\lambda_h + \eta \hat{c}_h^k(s,a) + B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})} \\ &= q_h^k(s,a,s') e^{-\lambda_h + B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})}.\end{aligned}$$

We can use the first constraint to discover that λ_h is a normalizer for every h, i.e.,

$$\begin{aligned}1 &= \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h^{k+1}(s,a,s') \\ 1 &= \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h^k(s,a,s') e^{-\lambda_h + B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})} \\ e^{\lambda_h} &= \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h^k(s,a,s') e^{B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})},\end{aligned}$$

so defining $Z_h^k(v,e) = \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} q_h^k(s,a,s') e^{B_h^k(s,a,s'|v,e)}$, we obtain

$$q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s') e^{B_h^k(s,a,s'|v^{\mu},e^{\mu,\beta,g})}}{Z_h^k(v^{\mu}, e^{\mu,\beta,g})}.$$

Now to find β and μ we consider the dual problem. Substituting q^{k+1} back into $\mathscr{L}$ we obtain the following dual problem.

$$\begin{aligned}\max_{\beta,\mu\geq 0,g\geq 0} \min_{q} \mathscr{L}(q) &= \max_{\beta,\mu\geq 0,g\geq 0} \mathscr{L}(q^{k+1}) \\ &= \max_{\beta,\mu\geq 0,g\geq 0} -\sum_{h=1}^{H} \log Z_h^k(v^{\mu}, e^{\mu,\beta,g}) - 1 + \sum_{h,s,a,s'} \tilde{q}_h^{k+1}(s,a,s') + \alpha \sum_{h,s} g_h(s).\end{aligned}$$

So after ignoring constants we observe that

$$\beta^k, \mu^k, g^k = \arg \min_{\beta,\mu\geq 0,g\geq 0} \sum_{h=1}^{H} \log Z_h^k(v^{\mu}, e^{\mu,\beta,g}) - \alpha \sum_{h,s} g_h(s).$$

B.2 Pseudo-code for "Bounded Bandit UC-O-REPS"

Algorithm 11 Bounded Bandit UC-O-REPS Algorithm

```
Parameters: state space $\mathscr{S}$, action space $\mathscr{A}$, number of episodes $K$, minimum reachability parameter $\alpha$, optimization parameter $\eta$ and confidence parameter $\delta$.
Initialization: $i(1) \leftarrow 1, N_h^1(s,a) \leftarrow 0, N_h^1(s,a,s') \leftarrow 0, n_h^1(s,a) \leftarrow 0, n_h^1(s,a,s') \leftarrow 0, \pi_h^1(a|s) \leftarrow 1/A, q_h^1(s,a,s') \leftarrow 1/(S^2A)\ \forall(s,a,s',h)$.
for $k = 1,\ldots,K$ do
  Traverse trajectory $U^k = (s_1^k, a_1^k, \ldots, s_H^k, a_H^k)$ using policy $\pi^k$.
  Observe costs $c^k(U^k) = \{c_h^k(s_h^k, a_h^k)\}_{h=1}^H$.
  Update in-epoch counters $\forall h = 1,\ldots,H$:
    $$n_h^{i(k)}(s_h^k,a_h^k) \leftarrow n_h^{i(k)}((s_h^k,a_h^k)+1$$
    $$n_h^{i(k)}(s_h^k,a_h^k,s_{h+1}^k) \leftarrow n_h^{i(k)}(s_h^k,a_h^k,s_{h+1}^k)+1.$$
  if $\exists(s,a,h) \in \mathscr{S}\times\mathscr{A}\times[H].\quad n_h^{i(k)}(s,a) \geq N_h^{i(k)}(s,a)$ then
    Start new epoch: $i(k+1) \leftarrow i(k)+1$.
    Initialize epoch counters $\forall(s,a,s',h) \in \mathscr{S}\times\mathscr{A}\times\mathscr{S}\times[H]$: $n_h^{i(k+1)}(s,a) \leftarrow 0, n_h^{i(k+1)}(s,a,s') \leftarrow 0$.
    Update total counters $\forall(s,a,s',h) \in \mathscr{S}\times\mathscr{A}\times\mathscr{S}\times[H]$:
      $$N_h^{i(k+1)}(s,a) \leftarrow N_h^{i(k)}(s,a) + n_h^{i(k)}(s,a)$$
      $$N_h^{i(k+1)}(s,a,s') \leftarrow N_h^{i(k)}(s,a,s') + n_h^{i(t)}(s,a,s').$$
    Compute transition estimate $\forall(s,a,s',h)$: $\bar{P}_h^{i(k+1)}(s'|s,a) \leftarrow \frac{N_h^{i(k+1)}(s,a,s')}{\max\{1,N_h^{i(k+1)}(s,a)\}}$.
  else
    Continue in the same epoch: $i(k+1) \leftarrow i(k)$.
  end if
  compute policy for next episode:
    $$q^{k+1},\pi^{k+1} \leftarrow \texttt{Comp-Policy}(\mathscr{S},\mathscr{A},K,\alpha,\eta,\delta,q^k,\bar{P}^{i(k+1)},N^{i(k+1)},c^k(U^k)).$$
end for
```

Algorithm 12 Comp-Policy Procedure

Input: state space $\mathscr{S}$, action space $\mathscr{A}$, number of episodes K, minimum reachability parameter α, optimization parameter η and confidence parameter δ, previous occupancy measure q^k, transition function estimate $\bar{P}^{i(k)}$, visit counters $N^{i(k)}$ and obtained losses $c^k(U^k)$.

Compute cost function estimate:

$$\hat{c}_h^k(s,a) = \begin{cases} \frac{c_h^k(s,a)}{q_h^k(s,a)}, & \text{if } s_h^k = s \text{ and } a_h^k = a \\ 0, & \text{otherwise} \end{cases}$$

Compute confidence set size parameter:

$$\varepsilon_h^{i(k)}(s,a) = \sqrt{\frac{2S \log \frac{KHSA}{\delta}}{\max\{1, N_h^{i(k)}(s,a)\}}}.$$

Define functions:

$$\begin{aligned} v_h^{\mu}(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\ e_h^{\mu,\beta,g}(s,a,s') &= \sum_{y \in \mathscr{S}} (\mu_h^-(s,a,y) + \mu_h^+(x,a,y))\varepsilon_h^k(s' \mid s,a) + \beta_{h+1}(s') - \beta_h(s) + g_h(s) \\ B_h^k(s,a,s' \mid v,e) &= e_h(s,a,s') + v_h(s,a,s') - \eta \hat{c}_h^k(s,a) - \sum_{y \in \mathscr{S}} \bar{P}_h^k(y \mid s,a) v_{h+1}(s,a,y) \\ Z_h^k(v,e) &= \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h^k(s,a,s') e^{B_h^k(s,a,s' \mid v,e)} \end{aligned}$$

Solve optimization problem:

$$\beta^k, \mu^k, g^k = \arg \min_{\beta, \mu \geq 0, g \geq 0} \sum_{h=1}^{H} \log Z_h^k(v^{\mu}, e^{\mu,\beta,g}) - \alpha \sum_{h,s} g_h(s).$$

Compute next occupancy measure $\forall (s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$:

$$q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s') e^{B_h^k(s,a,s' \mid v^{\mu}, e^{\mu,\beta,g})}}{Z_h^k(v^{\mu}, e^{\mu,\beta,g})}.$$

Compute next policy $\forall (s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H]$:

$$\pi_h^{k+1}(a \mid s) = \frac{\sum_{s' \in \mathscr{S}} q_h^{k+1}(s,a,s')}{\sum_{b \in \mathscr{A}} \sum_{s' \in \mathscr{S}} q_h^{k+1}(s,b,s')}.$$

B.3 Pseudo-code for "Shifted Bandit UC-O-REPS"

Algorithm 13 Shifted Bandit UC-O-REPS Algorithm

```
Parameters: state space $\mathscr{S}$, action space $\mathscr{A}$, number of episodes $K$, perturbation parameter $\rho$, optimization parameter $\eta$ and confidence parameter $\delta$.
Initialization: $i(1) \leftarrow 1, N_h^1(s,a) \leftarrow 0, N_h^1(s,a,s') \leftarrow 0, n_h^1(s,a) \leftarrow 0, n_h^1(s,a,s') \leftarrow 0, \pi_h^1(a|s) \leftarrow 1/A, q_h^1(s,a,s') \leftarrow 1/(S^2A)\ \forall(s,a,s',h)$.
for $k = 1,\dots,K$ do
    Traverse trajectory $U^k = (s_1^k, a_1^k, \dots, s_H^k, a_H^k)$ using policy $\pi^k$.
    Observe costs $c^k(U^k) = \{c_h^k(s_h^k, a_h^k)\}_{h=1}^H$.
    Update in-epoch counters $\forall h = 1,\dots,H$: $n_h^{i(k)}(s_h^k,a_h^k) \leftarrow n_h^{i(k)}((s_h^k,a_h^k)+1, n_h^{i(k)}(s_h^k,a_h^k,s_{h+1}^k) \leftarrow n_h^{i(k)}(s_h^k,a_h^k,s_{h+1}^k)+1$.
    if $\exists(s,a,h) \in \mathscr{S} \times \mathscr{A} \times [H].\ \ n_h^{i(k)}(s,a) \geq N_h^{i(k)}(s,a)$ then
        Start new epoch: $i(k+1) \leftarrow i(k)+1$.
        Initialize epoch counters $\forall(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$: $n_h^{i(k+1)}(s,a) \leftarrow 0, n_h^{i(k+1)}(s,a,s') \leftarrow 0$.
        Update total counters $\forall(s,a,s',h) \in \mathscr{S} \times \mathscr{A} \times \mathscr{S} \times [H]$: $N_h^{i(k+1)}(s,a) \leftarrow N_h^{i(k)}(s,a) + n_h^{i(k)}(s,a), N_h^{i(k+1)}(s,a,s') \leftarrow N_h^{i(k)}(s,a,s') + n_h^{i(t)}(s,a,s')$.
        Compute transition estimate $\forall(s,a,s',h)$: $\bar{P}_h^{i(k+1)}(s'|s,a) \leftarrow \frac{N_h^{i(k+1)}(s,a,s')}{\max\{1, N_h^{i(k+1)}(s,a)\}}$.
        Compute perturbed transition estimate $\forall(s,a,s',h)$:
            $$\widetilde{P}_h^{i(k+1)}(s'|s,a) \leftarrow (1-\rho)\bar{P}_h^{i(k+1)}(s'|s,a) + \frac{\rho}{S}.$$
    else
        Continue in the same epoch: $i(k+1) \leftarrow i(k)$.
    end if
    Compute next episode policy:
        $$q^{k+1}, \pi^{k+1} \leftarrow \texttt{Comp-Policy}(\mathscr{S}, \mathscr{A}, K, \frac{\rho}{S}, \eta, \delta, q^t, \widetilde{P}^{i(k+1)}, N^{i(k+1)}, c^k(U^k)).$$
end for
```

C Supplementary Material for Chapter 5

C.1 Algorithm

Algorithm 14 HOEFFDING-TYPE CONFIDENCE BOUNDS

```
input: state space $\mathscr{S}$, action space $\mathscr{A}$ and confidence parameter $\delta$.
initialization: arbitrary policy $\tilde{\pi}$, $m \leftarrow 1, \widetilde{B} \leftarrow c_{\min}, C_1 \leftarrow 0, \forall (s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$ :
$N(s,a,s') \leftarrow 0, N(s,a) \leftarrow 0$.
for $k = 1, 2, \dots$ do
    set $s \leftarrow s_{\text{init}}$.
    while $s \neq g$ do
        follow optimistic optimal policy: $a \leftarrow \tilde{\pi}(s)$, and suffer cost: $C_m \leftarrow C_m + c(s,a)$.
        observe next state $s' \sim P(\cdot \mid s,a)$.
        update visit counters: $N(s,a,s') \leftarrow N(s,a,s') + 1, N(s,a) \leftarrow N(s,a) + 1$.
        if $N(s', \tilde{\pi}(s')) \leq \frac{5000 \widetilde{B}^2 S}{c_{\min}^2} \log \frac{\widetilde{B} SA}{\delta c_{\min}}$ or $s' = g$ or $C_m \geq 24 \widetilde{B} \log \frac{4m}{\delta}$ then
            if $C_m \geq 24 \widetilde{B} \log \frac{4m}{\delta}$ then
                update $B_\star$ estimate: $\widetilde{B} \leftarrow 2\widetilde{B}$.
            end if
            advance intervals counter: $m \leftarrow m + 1$, and initialize suffered cost: $C_m \leftarrow 0$.
            compute empirical transition function $\bar{P}$ for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$ :

                $$\bar{P}(s' \mid s,a) = \frac{N(s,a,s')}{\max\{N(s,a),1\}}.$$

            compute policy $\tilde{\pi}$ that minimizes the expected cost w.r.t a transition function $\widetilde{P}$,
            such that for every $(s,a)$: $\left\|\widetilde{P}(\cdot \mid s,a) - \bar{P}(\cdot \mid s,a)\right\|_1 \leq 5\sqrt{\frac{S \log\left(SAN_+(s,a)/\delta\right)}{N_+(s,a)}}$.
        end if
        set $s \leftarrow s'$.
    end while
end for
```

C.2 Proofs

C.2.1 Proofs for Section 5.3.1

Proof of Theorem 5.3.3

Lemma (restatement of Theorem 5.3.3). With probability at least $1-\delta/2$, Ω^m holds and $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \leq 24 B_\star \log \frac{4m}{\delta}$ for all intervals m simultaneously. This implies that the total number of steps of the algorithm is

$$T = O\left(\frac{KB_\star}{c_{\min}} \log \frac{KB_\star SA}{\delta c_{\min}} + \frac{B_\star^3 S^2 A}{c_{\min}^3} \log^2 \frac{KB_\star SA}{\delta c_{\min}}\right).$$

Lemma C.2.1. *The event Ω^m holds for all intervals m simultaneously with probability at least* $1-\delta/4$.

Proof. Fix a state s and an action a. Consider an infinite sequence $\{Z_i\}_{i=1}^\infty$ of draws from the distribution $P(\cdot \mid s,a)$. By Theorem C.4.2 we get that for a prefix of length t of this sequence (that is $\{Z_i\}_{i=1}^t$)

$$\left\|P(\cdot \mid s,a) - \bar{P}_{\{Z_i\}_{i=1}^t}(\cdot \mid s,a)\right\|_1 \leq 2\sqrt{\frac{S \log(\delta_t^{-1})}{t}} \ ,$$

holds with probability $1-\delta_t$, where $\bar{P}_{\{Z_i\}_{i=1}^t}(\cdot \mid s,a)$ is the empirical distribution defined by the draws $\{Z_i\}_{i=1}^t$. We repeat this argument for every prefix $\{Z_i\}_{i=1}^t$ of $\{Z_i\}_{i=1}^\infty$ and for every state-action pair, with $\delta_t = \delta/8SAt^2$. Then from the union bound we get that Ω^m holds for all intervals m simultaneously with probability at least $1-\delta/4$. □

Lemma C.2.2. *Let m be an interval. If Ω^m holds then $\widetilde{V}^m(s) \leq V^\star(s) \leq B_\star$ for every $s \in \mathscr{S}$.*

Proof. [TGV$^+$20] show that all the transition functions in the confidence set of Equation (5.3) can be combined into a single augmented MDP. The optimal policy of the augmented MDP can be found efficiently, e.g., with Extended Value Iteration. The optimistic policy is the optimal policy in the augmented MDP. It minimizes $\widetilde{V}^m(s)$ over all policies and feasible transition functions, for all states $s \in \mathscr{S}$ simultaneously (following [BT91]). Since Ω^m holds, it follows that the real transition function is in the confidence set therefore it is also considered in the minimization. Thus $\widetilde{V}^m(s) \leq V^\star(s)$ for all $s \in \mathscr{S}$. Finally, $V^\star(s) \leq B_\star$ by the definition of $B_\star$. □

Lemma C.2.3. *Let m be an interval and (s,a) be a known state-action pair. If Ω^m holds then*

$$\|\widetilde{P}_m(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \le \frac{c(s,a)}{2B_\star} .$$

Proof. By the definition of the confidence set

$$\|\widetilde{P}_m(\cdot \mid s,a) - \bar{P}_m(\cdot \mid s,a)\|_1 \le 5\sqrt{\frac{S\log\bigl(SAN_+^m(s,a)/\delta\bigr)}{N_+^m(s,a)}} \le \frac{c(s,a)}{4B_\star},$$

where the last inequality follows because $\log(x)/x$ is decreasing, and since (s,a) is known $N_+^m(s,a) \ge \frac{5000B_\star^2 S}{c_{\min}^2}\log\frac{B_\star SA}{\delta c_{\min}}$. Similarly, since Ω^m holds we also have that

$$\|P(\cdot \mid s,a) - \bar{P}_m(\cdot \mid s,a)\|_1 \le 5\sqrt{\frac{S\log\bigl(SAN_+^m(s,a)/\delta\bigr)}{N_+^m(s,a)}} \le \frac{c(s,a)}{4B_\star},$$

and the lemma follows by the triangle inequality. □

Lemma C.2.4. *Let $\tilde{\pi}$ be a policy and $\widetilde{P}$ be a transition function. Denote the cost-to-go of $\tilde{\pi}$ with respect to $\widetilde{P}$ by $\widetilde{V}$. Assume that for every $s \in \mathscr{S}$, $\widetilde{V}(s) \le B_\star$ and that*

$$\left\|\widetilde{P}(\cdot \mid s,\tilde{\pi}(s)) - P(\cdot \mid s,\tilde{\pi}(s))\right\|_1 \le \frac{c(s,\tilde{\pi}(s))}{2B_\star}.$$

Then, $\tilde{\pi}$ is proper (with respect to P), and it holds that $V^{\tilde{\pi}}(s) \le 2B_\star$ for every $s \in \mathscr{S}$.

Proof. Consider the Bellman equations of $\tilde{\pi}$ with respect to transition function $\widetilde{P}$ at some state $s \in \mathscr{S}$ (see Theorem 2.2.2), defined as

$$\begin{aligned}
\widetilde{V}(s) &= c(s,\tilde{\pi}(s)) + \sum_{s' \in \mathscr{S}} \widetilde{P}(s' \mid s,\tilde{\pi}(s))\widetilde{V}(s') \\
&= c(s,\tilde{\pi}(s)) + \sum_{s' \in \mathscr{S}} P(s' \mid s,\tilde{\pi}(s))\widetilde{V}(s') \qquad \text{(C.1)} \\
&\quad + \sum_{s' \in \mathscr{S}} \widetilde{V}(s')\Bigl(\widetilde{P}(s' \mid s,\tilde{\pi}(s)) - P(s' \mid s,\tilde{\pi}(s))\Bigr) .
\end{aligned}$$

Notice that by our assumptions and using Hölder inequality,

$$\left|\sum_{s'\in\mathscr{S}} \widetilde{V}(s')\left(\widetilde{P}(s'\mid s,\tilde{\pi}(s)) - P(s'\mid s,\tilde{\pi}(s))\right)\right| \leq$$
$$\leq \|\widetilde{P}(\cdot\mid s,\tilde{\pi}(s)) - P(\cdot\mid s,\tilde{\pi}(s))\|_1 \cdot \|\widetilde{V}\|_\infty$$
$$\leq \frac{c(s,\tilde{\pi}(s))}{2B_\star}\cdot B_\star = \frac{c(s,\tilde{\pi}(s))}{2}\ .$$

Plugging this into Equation (C.1), we obtain

$$\begin{aligned}\widetilde{V}(s) &\geq c(s,\tilde{\pi}(s)) + \sum_{s'\in\mathscr{S}} P(s'\mid s,\tilde{\pi}(s))\widetilde{V}(s') - \frac{c(s,\tilde{\pi}(s))}{2}\\ &= \frac{c(s,\tilde{\pi}(s))}{2} + \sum_{s'\in\mathscr{S}} P(s'\mid s,\tilde{\pi}(s))\widetilde{V}(s').\end{aligned}$$

Therefore, defining $V' = 2\widetilde{V}$, then $V'(s) \geq c(s,\tilde{\pi}(s)) + \sum_{s'\in\mathscr{S}} P(s'\mid s,\tilde{\pi}(s))V'(s')$ for all $s\in\mathscr{S}$. The statement now follows by Theorem 2.2.2. □

Lemma C.2.5. *Let π be a proper policy such that for some $v>0$, $V^\pi(s)\leq v$ for every $s\in\mathscr{S}$. Then, the probability that the cost of π to reach the goal state from any state s is more than m, is at most $2e^{-m/4v}$ for all $m\geq 0$. Note that a cost of at most m implies that the number of steps is at most $\frac{m}{c_{min}}$.*

Proof. By Markov inequality, the probability that π accumulates cost of more than $2v$ before reaching the goal state is at most $1/2$. Iterating this argument, we get that the probability that π accumulates cost of more than $2kv$ before reaching the goal state is at most 2^{-k} for every integer $k\geq 0$. In general, for any $m\geq 0$, the probability that π suffers a cost of more than m is at most $2^{-\lfloor m/2v\rfloor} \leq 2\cdot 2^{-m/2v} \leq 2e^{-m/4v}$. □

For the next lemma we will need the following definitions. The trajectory visited in interval m is denoted by $U^m = (s^m_1, a^m_1, \dots, s^m_{H^m}, a^m_{H^m}, s^m_{H^m+1})$ where a^m_h is the action taken in s^m_h, and H^m is the length of the interval. In addition, the concatenation of the trajectories in the intervals up to and including interval m is denoted by $\bar{U}^m = \cup_{m'=1}^m U^{m'}$.

Lemma C.2.6. *Let m be an interval. For all $r\geq 0$, we have that*

$$\Pr\left[\sum_{h=1}^{H^m} c\left(s^m_h, a^m_h\right)\mathbb{I}\{\Omega^m\} > r \mid \bar{U}^{m-1}\right] \leq 3e^{-r/8B_\star}.$$

Proof. Note that Ω^m is determined given $\bar{U}^{m-1}$, and suppose that Ω^m holds otherwise $\sum_{h=1}^{H^m} c\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\}$ is 0. Also assume that $r \geq 8B_\star$ or else the statement holds trivially.

Define the MDP $M^{\text{know}} = (S^{\text{know}}, A, P^{\text{know}}, c, s_{\text{init}})$ in which every state $s \in \mathscr{S}$ such that $(s, \tilde{\pi}^m(s))$ is unknown is contracted into the goal state. Let P^{know} be the transition function induced in M^{know} by P, and let V_{know}^m be the cost-to-go of $\tilde{\pi}^m$ in M^{know} with respect to P^{know}. Similarly, define $\widetilde{P}_m^{\text{know}}$ as the transition function induced in M^{know} by $\widetilde{P}_m$, and $\widetilde{V}_{\text{know}}^m$ as the cost-to-go of $\tilde{\pi}^m$ in M^{know} with respect to $\widetilde{P}_m^{\text{know}}$. It is clear that $\widetilde{V}_{\text{know}}^m(s) \leq \widetilde{V}^m(s)$ for every $s \in \mathscr{S}$, so by Theorem C.2.2, $\widetilde{V}_{\text{know}}^m(s) \leq B_\star$. Moreover, since all the states $s \in \mathscr{S}$ for which $(s, \tilde{\pi}^m(s))$ is unknown were contracted to the goal state, we can use Theorem C.2.3 to obtain for all $s \in \mathscr{S}^{\text{know}}$:

$$\begin{aligned}\left\|\widetilde{P}_m^{\text{know}}(\cdot \mid s, \tilde{\pi}^m(s)) - P^{\text{know}}(\cdot \mid s, \tilde{\pi}^m(s))\right\|_1 &\leq \left\|\widetilde{P}_m(\cdot \mid s, \tilde{\pi}^m(s)) - P(\cdot \mid s, \tilde{\pi}^m(s))\right\|_1 \\ &\leq \frac{c(s, \tilde{\pi}^m(s))}{2B_\star}. \end{aligned} \tag{C.2}$$

We can apply Theorem C.2.4 in M^{know} and obtain that $V_{\text{know}}^m(s) \leq 2B_\star$ for every $s \in \mathscr{S}^{\text{know}}$. Notice that reaching the goal state in M^{know} is equivalent to reaching the goal state or an unknown state-action pair in M, and also recall that all state-action pairs in the interval are known except for the first one. Thus, from Theorem C.2.5,

$$\Pr\left[\sum_{h=2}^{H^m} c\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r - B_\star \mid \bar{U}^{m-1}\right] \leq 2e^{-(r-B_\star)/8B_\star} \leq 3e^{-r/8B_\star}.$$

Since $\widetilde{V}^m \leq B_\star$, our algorithm will never select an action whose instantaneous cost is larger than $B_\star$. Since the first state-action in the interval might not be known, its cost is at most $B_\star$, and therefore

$$\begin{aligned}\Pr\left[\sum_{h=1}^{H^m} c\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r \mid \bar{U}^{m-1}\right] &\leq \Pr\left[\sum_{h=2}^{H^m} c\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r - B_\star \mid \bar{U}^{m-1}\right] \\ &\leq 3e^{-r/8B_\star}.\end{aligned}$$

□

Proof of Theorem 5.3.3. From Theorem C.2.6, with probability at least $1 - \delta/16m^2$,

$$\sum_{h=1}^{H^m} c\big(s_h^m, a_h^m\big) \leq 24B_\star \log\frac{4m}{\delta},$$

and by the union bound this holds for all intervals m simultaneously with probability at least $1-\delta/4$. By Theorem C.2.1, with probability $1-\delta/4$, Ω^m holds for all intervals m. Combining these two facts again by a union bound, we get that both Ω^m holds and the cost of interval m is at most $24B_\star \log \frac{4m}{\delta}$ simultaneously to all intervals m with probability at least $1-\delta/2$.

If the cost of all intervals is bounded (and therefore so is the length of the interval), we can use the bound on the number of intervals in Observation 5.3.2 to conclude that

$$\begin{aligned} T &= O\left(\frac{B_\star}{c_{\min}} \log \frac{M}{\delta} \cdot \left(K + \frac{B_\star^2 S^2 A}{c_{\min}^2} \log \frac{B_\star SA}{\delta c_{\min}}\right)\right) \\ &= O\left(\frac{KB_\star}{c_{\min}} \log \frac{KB_\star SA}{\delta c_{\min}} + \frac{B_\star^3 S^2 A}{c_{\min}^3} \log^2 \frac{KB_\star SA}{\delta c_{\min}}\right). \end{aligned}$$

□

Proof of Theorem 5.3.4

Lemma (restatement of Theorem 5.3.4). With probability at least $1-\delta/2$, we have

$$\widetilde{R}_K \le \frac{5000 B_\star^3 S^2 A}{c_{\min}^2} \log \frac{B_\star SA}{c_{\min}\delta} + B_\star \sqrt{T \log \frac{4T}{\delta}} + 10 B_\star \sqrt{S \log \frac{SAT}{\delta}} \sum_{s,a} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.$$

To analyze $\widetilde{R}_K$, we begin by plugging in the Bellman optimality equation of $\tilde{\pi}^m$ with respect to $\widetilde{P}_m$ into $\widetilde{R}_K$. This allows us to decompose $\widetilde{R}_K$ into three terms as follows.

$$\begin{aligned} \widetilde{R}_K &= \sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{V}^m(s_h^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m) \widetilde{V}^m(s')\right) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}}) \\ &= \sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{V}^m(s_h^m) - \widetilde{V}^m(s_{h+1}^m)\right) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}}) && \text{(C.3)} \\ &\quad + \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{s' \in \mathscr{S}} \widetilde{V}^m(s') \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right) \mathbb{I}\{\Omega^m\} && \text{(C.4)} \\ &\quad + \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} P(s' \mid s_h^m, a_h^m) \widetilde{V}^m(s')\right) \mathbb{I}\{\Omega^m\}. && \text{(C.5)} \end{aligned}$$

Equation (C.3) is a bound on the cost suffered from switching policies each time we visit an unknown state-action pair and is bounded by the following lemma.

Lemma C.2.7. $\sum_{m=1}^{M}\sum_{h=1}^{H^m}\left(\widetilde{V}^m(s_h^m)-\widetilde{V}^m(s_{h+1}^m)\right)\mathbb{I}\{\Omega^m\}\leq B_\star SA\cdot\frac{5000B_\star^2S}{c_{min}^2}\log\frac{B_\star SA}{\delta c_{min}}+K\cdot V^\star(s_{init})$.

Proof. Note that per interval $\sum_{h=1}^{H^m}(\widetilde{V}^m(s_h^m)-\widetilde{V}^m(s_{h+1}^m))$ is a telescopic sum which equals $\widetilde{V}^m(s_1^m)-\widetilde{V}^m(s_{H^m+1}^m)$. Furthermore, for every two consecutive intervals $m,m+1$ one of the following occurs:

(i) If interval m ended in the goal state then $\widetilde{V}^m(s_{H^m+1}^m)=\widetilde{V}^m(g)=0$ and $\widetilde{V}^{m+1}(s_1^{m+1})=\widetilde{V}^{m+1}(s_{\text{init}})$. Thus, using Theorem C.2.2 for the last inequality,

$$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\}-\widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\}=\widetilde{V}^{m+1}(s_{\text{init}})\mathbb{I}\{\Omega^{m+1}\}\leq V^\star(s_{\text{init}}).$$

This happens at most K times.

(ii) If interval m ended in an unknown state then

$$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\}-\widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\}\leq\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\}\leq B_\star.$$

This happens at most $SA\cdot\frac{5000B_\star^2S}{c_{\text{min}}^2}\log\frac{B_\star SA}{\delta c_{\text{min}}}$ times.

□

Theorem C.2.8 bounds Equation (C.4) using techniques borrowed from [JOA10].

Lemma C.2.8. *It holds that*

$$\sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in\mathcal{S}}\widetilde{V}^m(s')\left(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\right)\mathbb{I}\{\Omega^m\}\leq 10B_\star\sqrt{S\log\frac{SAT}{\delta}}\sum_{s,a}\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.$$

Proof. Using the definition of the confidence sets we obtain

$$\begin{aligned}\sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in\mathcal{S}}\widetilde{V}^m(s')\left(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\right)\mathbb{I}\{\Omega^m\}\leq&\\ \leq B_\star\sum_{s\in\mathcal{S}}\sum_{a\in A}\sum_{m=1}^{M}n_m(s,a)\|P(\cdot\mid s,a)-\widetilde{P}_m(\cdot\mid s,a)\|_1\mathbb{I}\{\Omega^m\}&\\ \leq 10B_\star\sum_{s\in\mathcal{S}}\sum_{a\in A}\sum_{m=1}^{M}n_m(s,a)\sqrt{\frac{S\log\left(SAN_+^m(s,a)/\delta\right)}{N_+^m(s,a)}}&\\ \leq 10B_\star\sqrt{S\log\frac{SAT}{\delta}}\sum_{s\in\mathcal{S}}\sum_{a\in A}\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}.&\end{aligned}$$

where the first inequality follows from Hölder inequality and Theorem C.2.2, and the second because $\widetilde{P}_m$ and P are both in the confidence set of Equation (5.3) when Ω^m holds. The third inequality follows because $N_+^m(s,a) \leq T$. □

Theorem C.2.9 bounds the term in Equation (C.5) using Azuma's concentration inequality.

Lemma C.2.9. *With probability at least* $1-\delta/2$,

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}^m(s_{h+1}^m) - \sum_{s'\in\mathscr{S}} P(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\} \leq B_\star\sqrt{T\log\frac{4T}{\delta}}.$$

Proof. Consider the infinite sequence of random variables

$$X_t = \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{s'\in\mathscr{S}} P(s' \mid s_h^m, \tilde{\pi}^m(s_h^m))\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\},$$

where m is the interval containing time t, and h is the index of time step t within interval m. Notice that this is a martingale difference sequence, and $|X_t| \leq B_\star$ by Theorem C.2.2. Now, we apply anytime Azuma's inequality (Theorem C.4.1) to any prefix of the sequence $\{X_t\}_{t=1}^{\infty}$. Thus, with probability at least $1-\delta/2$, for every T:

$$\sum_{t=1}^{T} X_t \leq B_\star\sqrt{T\log\frac{4T}{\delta}}.$$

□

Proof of Theorem 5.3.1

Theorem (restatement of Theorem 5.3.1). Suppose that Theorem 5.2.1 holds. With probability at least $1-\delta$ the regret of Algorithm 6 is bounded as follows:

$$R_K = O\left(\sqrt{\frac{B_\star^3 S^2 AK}{c_{\min}}}\log\frac{KB_\star SA}{\delta c_{\min}} + \frac{B_\star^3 S^2 A}{c_{\min}^2}\log^{3/2}\frac{KB_\star SA}{\delta c_{\min}}\right).$$

Lemma C.2.10. *Assume that the number of steps in every interval is is at most* $\frac{24B_\star}{c_{min}}\log\frac{4m}{\delta}$. *Then for every* $s\in\mathscr{S}$ *and* $a\in A$,

$$\sum_{m=1}^{M}\frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}} \leq 3\sqrt{N_{M+1}(s,a)}\ .$$

Proof. We claim that, by the assumption of the lemma, for every interval m we have that $n_m(s,a) \leq N_+^m(s,a)$. Indeed, if (s,a) is unknown then $n_m(s,a) = 1$ and since $N_+^m(s,a) \geq 1$ the claim follows. If (s,a) is known then $N_+^m(s,a) \geq \frac{5000 B_\star^2 S}{c_{\min}^2} \log \frac{B_\star SA}{\delta c_{\min}}$ and by our assumption the length of the interval, and in particular $n_m(s,a)$, is at most $\frac{24 B_\star}{c_{\min}} \log \frac{4m}{\delta}$. Our statement then follows by [JOA10, Lemma 19]. □

Proof of Theorem 5.3.1. With probability at least $1-\delta$, both Theorems C.2.9 and 5.3.3 hold. Theorem 5.3.3 states that the length of every interval is at most $\frac{24 B_\star}{c_{\min}} \log \frac{4m}{\delta}$, and Theorem C.2.10 obtains

$$\sum_{s \in \mathcal{S}} \sum_{a \in A} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}} \leq 3 \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sqrt{N_{M+1}(s,a)} \leq 3\sqrt{SAT}, \tag{C.6}$$

where the last inequality follows from Jensen's inequality and the fact that

$$\sum_{s,a) \in \mathcal{S} \times \mathcal{A}} N_{M+1}(s,a) \leq T.$$

Next, we sum the bounds of Theorems C.2.7 to C.2.9 and use Equation (C.6) to obtain

$$R_K \leq 5000 \frac{B_\star^3 S^2 A}{c_{\min}^2} \log \frac{B_\star SA}{\delta c_{\min}} + 30 B_\star S \sqrt{AT \log \frac{SAT}{\delta}} + B_\star \sqrt{T \log \frac{4T}{\delta}}.$$

To finish the proof use Theorem 5.3.3 to bound T. □

C.2.2 Proofs for Section 5.4.1

Proof of Theorem 5.4.2

Lemma (restatement of Theorem 5.4.2). With probability at least $1-\delta/2$, Ω^m holds for all intervals m simultaneously.

Proof. Fix a triplet $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}^+$. Consider an infinite sequence $(Z_i)_{i=1}^\infty$ of draws from the distribution $P(\cdot \mid s,a)$ and let $X_i = \mathbb{I}\{Z_i = s'\}$. We apply Equation (C.18) of Theorem C.4.3 with $\delta_t = \frac{\delta}{4S^2At^2}$ to a prefix of length t of the sequence $(X_i)_{i=1}^\infty$. Then divide Equation (C.18) by t and obtain that, after simplifying using the assumptions that $S \geq 2$ and $A \geq 2$, Equation (5.5) holds with probability $1-\delta_t$. We repeat this argument for every prefix $(Z_i)_{i=1}^t$ of $(Z_i)_{i=1}^\infty$ and for every state-action-state triplet. Then from the union bound we get that Ω^m holds for all intervals m simultaneously with probability at least $1-\delta/2$. □

Proof of Theorem 5.4.3

Lemma (restatement of Theorem 5.4.3). It holds that

$$\widetilde{R}_M = \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_h^m) - \widetilde{V}^m(s_{h+1}^m) \right) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}})$$
$$+ \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}.$$

Lemma C.2.11. *Let m be an interval. If Ω^m holds then $\tilde{\pi}^m$ satisfies the Bellman equations in the optimistic model:*

$$\widetilde{V}^m(s) = c(s, \tilde{\pi}^m(s)) + \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s, \tilde{\pi}^i(s)) \widetilde{V}^m(s'), \quad \forall s \in \mathscr{S}.$$

Proof. Note that the Bellman equations hold in the optimistic model since as we defined this model, there is a nonzero probability of transition to the goal state by any action from every state. Thus in the optimistic model every policy is a proper policy and in particular Theorem 2.2.3 holds. □

Proof of Theorem 5.4.3. By Theorem C.2.11, we can use the Bellman equations in the optimistic model to have the following interpretation of the costs for every interval m and time h:

$$c(s_h^m, a_h^m)\mathbb{I}\{\Omega^m\} = \left(\widetilde{V}^m(s_h^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_i(s' \mid s_h^m, a_h^m) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}$$
$$= \left(\widetilde{V}^m(s_h^m) - \widetilde{V}^m(s_{h+1}^m) \right) \mathbb{I}\{\Omega^m\} + \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_i(s' \mid s_h^m, a_h^m) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}. \tag{C.7}$$

We now write $\widetilde{R}_M = \sum_{m=1}^{M} \sum_{h=1}^{H^m} c(s_h^m, a_h^m) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}})$, and substitute for each cost using Equation (C.7) to get the lemma. □

Proof of Theorem 5.4.4

Lemma (restatement of Theorem 5.4.4). It holds that

$$\sum_{m=1}^{M} \big(\sum_{h=1}^{H^m} \widetilde{V}^m(s_h^m) - \widetilde{V}^m(s_{h+1}^m) \big) \mathbb{I}\{\Omega^m\} - K \cdot V^\star(s_{\text{init}}) \leq 2B_\star SA \log T.$$

Lemma C.2.12. *Let m be an interval. If Ω^m holds then $\widetilde{V}^m(s) \le V^\star(s) \le B_\star$ for every $s \in \mathcal{S}$.*

Proof. Denote by $\widetilde{P}$ the transition function computed by Algorithm 7 at the beginning of epoch $i(m)$, and by $\widetilde{V}$ the cost-to-go with respect to $\widetilde{P}$. We claim that for every proper policy π and state $s \in \mathcal{S}$, $\widetilde{V}^\pi(s) \le V^\pi(s)$. Then, the lemma follows easily since $\widetilde{V}^m(s) \le \widetilde{V}^{\pi^\star}(s) \le V^{\pi^\star}(s) \le B_\star$.

Indeed, let $s \in \mathcal{S}$ and consider the Bellman equations of π with respect to P:

$$V^\pi(s) = c(s,\pi(s)) + \sum_{s' \in \mathcal{S}} P(s' \mid s,\pi(s)) V^\pi(s') \ge c(s,\pi(s)) + \sum_{s' \in \mathcal{S}} \widetilde{P}(s' \mid s,\pi(s)) V^\pi(s'),$$

where the inequality follows because $\widetilde{P}(s' \mid s,a) \le P(s' \mid s,a)$ for every $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. This holds since P is in the confidence set of Equation (5.5) (as Ω^m holds), and by the way $\widetilde{P}$ is computed in *Algorithm* 7. Therefore, by Theorem 2.2.2 we obtain that $V^\pi(s) \ge \widetilde{V}^\pi(s)$ for every $s \in \mathcal{S}$ as required. □

Proof of Theorem 5.4.4. For every two consecutive intervals $m, m+1$, denoting $i = i(m)$, we have one of the following:

(i) If interval m ended in the goal state then $\widetilde{V}^{i(m)}(s^m_{H^m+1}) = \widetilde{V}^{i(m)}(g) = 0$. Moreover, it holds that $\widetilde{V}^{i(m+1)}(s^{m+1}_1) = \widetilde{V}^{i(m+1)}(s_{\text{init}})$. Therefore, by Theorem C.2.12,

$$\widetilde{V}^{i(m+1)}(s^{m+1}_1)\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^{i(m)}(s^m_{H^m+1})\mathbb{I}\{\Omega^m\} = \widetilde{V}^{i(m+1)}(s_{\text{init}})\mathbb{I}\{\Omega^{m+1}\} \le V^\star(s_{\text{init}}).$$

This happens at most K times.

(ii) If interval m ended in an unknown state-action pair or since the cost reached $B_\star$, and we stay in the same epoch, then $i(m) = i(m+1) = i$ and $s^{m+1}_1 = s^m_{H^m+1}$. Thus

$$\begin{aligned}\widetilde{V}^{i(m+1)}(s^{m+1}_1)\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^{i(m)}(s^m_{H^m+1})\mathbb{I}\{\Omega^m\} &= \\ = \widetilde{V}^i(s^{m+1}_1)\mathbb{I}\{\Omega^m\} - \widetilde{V}^i(s^m_{H^m+1})\mathbb{I}\{\Omega^m\} &= 0.\end{aligned}$$

(iii) If interval m ended by doubling the visit count to some state-action pair, then we start a new epoch. Thus by Theorem C.2.12,

$$\widetilde{V}^{i(m+1)}(s^{m+1}_1)\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^{i(m)}(s^m_{H^m+1})\mathbb{I}\{\Omega^m\} \le \widetilde{V}^{i+1}(s^{m+1}_1)\mathbb{I}\{\Omega^{m+1}\} \le B_\star,$$

This happens at most $2SA\log T$ times.

To conclude, we have

$$\begin{aligned}\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}^{i(m)}(s_h^m)-\widetilde{V}^{i(m)}(s_{h+1}^m)\right)\mathbb{1}\{\Omega^m\}-KV^\star(s_{\text{init}}) &\leq KV^\star(s_{\text{init}})+2B_\star SA\log T-KV^\star(s_{\text{init}})\\ &= 2B_\star SA\log T.\end{aligned}$$

□

Proof of Theorem 5.4.5

Lemma (restatement of Theorem 5.4.5). With probability at least $1-\delta/4$, the following holds for all $M=1,2,\dots$ simultaneously.

$$\begin{aligned}&\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}^m(s_{h+1}^m)-\sum_{s'\in\mathscr{S}}\widetilde{P}_m(s'\mid s_h^m,a_h^m)\widetilde{V}^m(s')\right)\mathbb{1}\{\Omega^m\}\\ &\quad\leq\sum_{m=1}^{M}\mathbb{E}\left[\left(\sum_{h=1}^{H^m}\widetilde{V}^m(s_{h+1}^m)-\sum_{s'\in\mathscr{S}}\widetilde{P}_m(s'\mid s_h^m,a_h^m)\widetilde{V}^m(s')\right)\mathbb{1}\{\Omega^m\}\mid\bar{U}^{m-1}\right]+3B_\star\sqrt{M\log\frac{8M}{\delta}}.\end{aligned}$$

Proof. Consider the following martingale difference sequence $(X^m)_{m=1}^\infty$ defined by

$$X^m=\sum_{h=1}^{H^m}\left(\widetilde{V}^m(s_{h+1}^m)-\sum_{s'\in\mathscr{S}}\widetilde{P}_m(s'\mid s_h^m,a_h^m)\widetilde{V}^m(s')\right)\mathbb{1}\{\Omega^m\}.$$

The Bellman optimality equations of $\tilde{\pi}^m$ with respect to $\widetilde{P}_m$ (Theorem C.2.11) obtain

$$|X^m|=\left|\left(\underbrace{\widetilde{V}^m(s_{H^m+1}^m)-\widetilde{V}^m(s_1^m)}_{\leq B_\star}+\underbrace{\sum_{h=1}^{H^m}c(s_h^m,a_h^m)}_{\leq 2B_\star}\right)\mathbb{1}\{\Omega^m\}\right|\leq 3B_\star,$$

where the inequality follows from Theorem C.2.12 and the fact that the total cost within each interval at most $2B_\star$ by construction. Therefore, we use anytime Azuma's inequality (Theorem C.4.1) to obtain that with probability at least $1-\delta/4$:

$$\sum_{m=1}^{M}X^m\leq\sum_{m=1}^{M}\mathbb{E}\left[X^m\mid\bar{U}^{m-1}\right]+3B_\star\sqrt{M\log\frac{8M}{\delta}}.$$

□

Proof of Theorem 5.4.6

Lemma (restatement of Theorem 5.4.6). For every interval m and time h, denote $A_h^m = \frac{\log(SAN_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)}$. Then,

$$\mathbb{E}\left[\left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathcal{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\leq 16 \cdot \mathbb{E}\left[\sum_{h=1}^{H^m} \sqrt{S\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] + 272 \cdot \mathbb{E}\left[\sum_{h=1}^{H^m} B_\star S A_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right],$$

where $\mathbb{V}_h^m$ is the empirical variance defined as

$$\mathbb{V}_h^m = \sum_{s' \in \mathcal{S}^+} P(s' \mid s_h^m, a_h^m)\left(\widetilde{V}^m(s') - \sum_{s'' \in \mathcal{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)^2.$$

The next lemma gives a different interpretation to the confidence bounds of Equation (5.5), and will be useful in the proofs that follow.

Lemma C.2.13. *Denote $A_h^m = \log(SAN_+^m(s,a)/\delta)/N_+^m(s,a)$. When Ω^m holds we have for any $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}^+$:*

$$\left|P(s' \mid s,a) - \widetilde{P}_m(s' \mid s,a)\right| \leq 8\sqrt{P(s' \mid s,a)A_h^m} + 136A_h^m.$$

Proof. Since Ω^m holds we have for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}^+$ that

$$\bar{P}_m(s' \mid s,a) - P(s' \mid s,a) \leq 4\sqrt{\bar{P}_m(s' \mid s,a)A_h^m} + 28A_h^m.$$

This is a quadratic inequality in $\sqrt{\bar{P}_m(s' \mid s,a)}$. Using the fact that $x^2 \leq a \cdot x + b$ implies $x \leq a + \sqrt{b}$ with $a = 4\sqrt{A_h^m}$ and $b = P(s' \mid s,a) + 28A_h^m$, we have

$$\sqrt{\bar{P}_m(s' \mid s,a)} \leq 4\sqrt{A_h^m} + \sqrt{P(s' \mid s,a) + 28A_h^m} \leq \sqrt{P(s' \mid s,a)} + 10\sqrt{A_h^m},$$

where we used the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ that holds for any $x \geq 0$ and $y \geq 0$. Substituting back into Equation (5.5) obtains

$$\left|P(s' \mid s,a) - \bar{P}_m(s' \mid s,a)\right| \leq 4\sqrt{P(s' \mid s,a)A_h^m} + 68A_h^m.$$

From a similar argument

$$\left|\widetilde{P}_m(s' \mid s,a) - \bar{P}_m(s' \mid s,a)\right| \le 4\sqrt{P(s' \mid s,a)A_h^m} + 68A_h^m.$$

Using the triangle inequality finishes the proof. □

Proof of Theorem 5.4.6. Denote $X^m = \left(\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\}$, and $Z_h^m = \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} P(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\}$. Think of the interval as an infinite stochastic process, and note that, conditioned on $\bar{U}^{m-1}$, $\left(Z_h^m\right)_{h=1}^{\infty}$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^{\infty}$, where U^h is the trajectory of the learner from the beginning of the interval and up to and including time h. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval. Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^{\infty}$ which is bounded by $2B_\star/c_{\min}$. Hence by the optional stopping theorem $\mathbb{E}[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1}] = 0$, which gets us

$$\begin{aligned}
\mathbb{E}[X^m \mid \bar{U}^{m-1}] &= \mathbb{E}\left[\sum_{h=1}^{H^m}\left(\widetilde{V}^m(s_{h+1}^m) - \sum_{s' \in \mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1}\right] + \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}} \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\widetilde{V}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}} \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\widetilde{V}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right].
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}} \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\widetilde{V}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad = \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}^+} \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\left(\widetilde{V}^m(s') - \sum_{s'' \in \mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad \le \mathbb{E}\left[8\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}^+} \sqrt{A_h^m P(s' \mid s_h^m, a_h^m)\left(\widetilde{V}^m(s') - \sum_{s'' \in \mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)^2}\,\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\qquad + \mathbb{E}\left[136\sum_{h=1}^{H^m}\sum_{s' \in \mathscr{S}^+} A_h^m\left|\widetilde{V}^m(s') - \sum_{s'' \in \mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right|\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \\
&\quad \le \mathbb{E}\left[16\sum_{h=1}^{H^m}\sqrt{S\mathbb{V}_h^m A_h^m}\,\mathbb{I}\{\Omega^m\} + 272SB_\star A_h^m\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right],
\end{aligned}$$

where the first equality follows since $\widetilde{V}^m(g) = 0$, and $P(\cdot \mid s_h^m, a_h^m)$ and $\widetilde{P}_i(\cdot \mid s_h^m, a_h^m)$ are probability distributions over S^+ whence $\sum_{s''\in\mathcal{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')$ does not depend on s'. The first inequality follows from Theorem C.2.13, and the second inequality from Jensen's inequality, Theorem C.2.12, $|S^+| \le 2S$, and the definition of $\mathbb{V}_h^m$. □

Proof of Theorem 5.4.7

Lemma (restatement of Theorem 5.4.7). For any interval m, $\mathbb{E}\big[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \le 44B_\star^2$.

Lemma C.2.14. *Let m be an interval and (s,a) be a known state-action pair. If Ω^m holds then for every $s' \in \mathcal{S}^+$*

$$\big|\widetilde{P}_m(s' \mid s,a) - P(s' \mid s,a)\big| \le \frac{1}{8}\sqrt{\frac{c_{min} \cdot P(s' \mid s,a)}{SB_\star}} + \frac{c_{min}}{4SB_\star}.$$

Proof. By Theorem C.2.13 we have that

$$\big|\widetilde{P}_m(s' \mid s,a) - P(s' \mid s,a)\big| \le 8\sqrt{\frac{P(s' \mid s,a)\log\big(SAN_+^m(s,a)/\delta\big)}{N_+^m(s,a)}} + \frac{136\log\big(SAN_+^m(s,a)/\delta\big)}{N_+^m(s,a)}$$

which gives the required bound because $\log(x)/x$ is decreasing, and (s,a) is a known state-action pair so $N_+^m(s,a) \ge 30000 \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$. □

Proof of Theorem 5.4.7. Note that the first state-action pair in the subinterval, (s_1^m, a_1^m), might be unknown and that all state-action pairs that appear afterwards are known. Thus, we bound

$$\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mid \bar{U}^{m-1}\right] = \mathbb{E}\left[\mathbb{V}_1^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] + \mathbb{E}\left[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right].$$

The first summand is trivially bounded by $B_\star^2$ (Theorem C.2.12). We now upper bound $\mathbb{E}\big[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big]$. Denote $Z_h^m = \big(\widetilde{V}^m(s_{h+1}^m) - \sum_{s'\in\mathcal{S}} P(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\big)\mathbb{I}\{\Omega^m\}$, and think of the interval as an infinite stochastic process. Note that, conditioned on $\bar{U}^{m-1}$, $\big(Z_h^m\big)_{h=1}^\infty$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^\infty$, where U^h is the trajectory of the learner from the beginning of the interval and up to time h and including. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval. Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^\infty$

which is bounded by $2B_\star/c_{\min}$. Therefore, applying Theorem C.2.15 found below obtains

$$\mathbb{E}\left[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] = \mathbb{E}\left[\left(\sum_{h=2}^{H^m} Z_h^m \mathbb{I}\{\Omega^m\}\right)^2 \mid \bar{U}^{m-1}\right]. \tag{C.8}$$

We now proceed by bounding $|\sum_{h=1}^{H^m} Z_h^m|$ when Ω^m occurs. Therefore,

$$\begin{aligned}\left|\sum_{h=2}^{H^m} Z_h^m\right| &= \left|\sum_{h=2}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{s'\in\mathscr{S}} P(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right| \\ &\leq \left|\sum_{h=2}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \widetilde{V}^m(s_h^m)\right| \qquad \text{(C.9)} \\ &\quad + \left|\sum_{h=2}^{H^m} \widetilde{V}^m(s_h^m) - \sum_{s'\in\mathscr{S}} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{V}^m(s')\right| \qquad \text{(C.10)} \\ &\quad + \left|\sum_{h=2}^{H^m}\sum_{s'\in\mathscr{S}^+} \left(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\right)\left(\widetilde{V}^m(s') - \sum_{s''\in\mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)\right|, \qquad \text{(C.11)}\end{aligned}$$

where Equation (C.11) is given as $P(\cdot \mid s_h^m, a_h^m)$ and $\widetilde{P}_i(\cdot \mid s_h^m, a_h^m)$ are probability distributions over S^+, $\sum_{s''\in\mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')$ is constant w.r.t s', and $\widetilde{V}^m(g) = 0$.

We now bound each of the three terms above individually. Equation (C.9) is a telescopic sum that is at most $B_\star$ on Ω^m (Theorem C.2.12). For Equation (C.10), we use the Bellman equations for $\tilde{\pi}^m$ on the optimistic model defined by the transitions $\widetilde{P}_m$ (Theorem C.2.11) thus it is at most $\sum_{h=2}^{H^m} c(s_h^m, a_h^m) \leq 2B_\star$ (see text following Theorem 5.4.5). For Equation (C.11), recall that all states-action pairs at times $h = 2, \ldots, H^m$ are known by definition of H^m. Hence by Theorem C.2.14,

$$\begin{aligned}&\left|\sum_{s'\in\mathscr{S}^+}\left(\widetilde{V}^m(s') - \sum_{s''\in\mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)\left(\widetilde{P}_m(s' \mid s_h^m, a_h^m) - P(s' \mid s_h^m, a_h^m)\right)\right| \\ &\quad \leq \frac{1}{8}\sum_{s'\in\mathscr{S}^+}\sqrt{\frac{c_{\min}\cdot P(s' \mid s_h^m, a_h^m)\left(\widetilde{V}^m(s') - \sum_{s''\in\mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)^2}{SB_\star}} \\ &\qquad + \sum_{s'\in\mathscr{S}^+}\frac{c_{\min}}{4SB_\star}\cdot\underbrace{\left|\widetilde{V}^m(s') - \sum_{s''\in\mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right|}_{\leq B_\star \text{ by Theorem C.2.12}} \\ &\quad \leq \frac{1}{4}\sqrt{\frac{c_{\min}\cdot\mathbb{V}_h^m}{B_\star}} + \frac{c(s_h^m, a_h^m)}{2}, \quad \text{(by Jensen's inequality, } c_{\min} \leq c(s_h^m, a_h^m),\ |S^+| \leq 2S)\end{aligned}$$

and again by Jensen's inequality and that the total cost throughout the interval is at most $2B_\star$, we have on Ω^m

$$\sum_{h=2}^{H^m}\frac{1}{4}\sqrt{\frac{c_{\min}\cdot\mathbb{V}_h^m}{B_\star}}+\frac{c\left(s_h^m,a_h^m\right)}{2}\le\frac{1}{4}\sqrt{\underbrace{H^m}_{\le 2B_\star/c_{\min}}\cdot\sum_{h=2}^{H^m}\frac{c_{\min}\cdot\mathbb{V}_h^m}{B_\star}}+\frac{1}{2}\underbrace{\sum_{h=2}^{H^m}c\left(s_h^m,a_h^m\right)}_{\le 2B_\star} \tag*{(Jensen's inequality)}$$

$$\le\frac{1}{4}\sqrt{2\sum_{h=2}^{H^m}\mathbb{V}_h^m}+B_\star.$$

Plugging these bounds back into Equation (C.8) gets us

$$\begin{aligned}\mathbb{E}\left[\sum_{h=2}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right]&\le\mathbb{E}\left[\left(4B_\star+\frac{1}{4}\sqrt{2\sum_{h=1}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\}}\right)^2\;\middle|\;\bar{U}^{m-1}\right]\\&\le 32B_\star^2+\frac{1}{4}\mathbb{E}\left[\sum_{h=2}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right],\end{aligned}$$

where the last inequality is by the elementary inequality $(a+b)^2\le 2(a^2+b^2)$. Rearranging gets us $\mathbb{E}\left[\sum_{h=2}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\}\mid\bar{U}^{m-1}\right]\le 43B_\star^2$, and the lemma follows. □

Lemma C.2.15. *Let $(X_t)_{t=1}^\infty$ be a martingale difference sequence adapted to the filtration $(\mathscr{F}_t)_{t=0}^\infty$. Let $Y_n=(\sum_{t=1}^n X_t)^2-\sum_{t=1}^n\mathbb{E}[X_t^2\mid\mathscr{F}_{t-1}]$. Then $(Y_n)_{n=0}^\infty$ is a martingale, and in particular if τ is a stopping time such that $\tau\le c$ almost surely, then $\mathbb{E}[Y_\tau]=0$.*

Proof. We first show that $(Y_n)_{n=1}^\infty$ is a martingale. Indeed,

$$\begin{aligned}\mathbb{E}[Y_n\mid\mathscr{F}_{n-1}]&=\mathbb{E}\left[\left(\sum_{t=1}^n X_t\right)^2-\sum_{t=1}^n\mathbb{E}[X_t^2\mid\mathscr{F}_{t-1}]\mid\mathscr{F}_{n-1}\right]\\&=\mathbb{E}\left[\left(\sum_{t=1}^{n-1}X_t\right)^2-2\left(\sum_{t=1}^{n-1}X_t\right)X_n+X_n^2-\sum_{t=1}^n\mathbb{E}[X_t^2\mid\mathscr{F}_{t-1}]\mid\mathscr{F}_{n-1}\right]\\&=\left(\sum_{t=1}^{n-1}X_t\right)^2-2\left(\sum_{t=1}^{n-1}X_t\right)\cdot 0+\mathbb{E}[X_n^2\mid\mathscr{F}_{n-1}]-\sum_{t=1}^n\mathbb{E}[X_t^2\mid\mathscr{F}_{t-1}]\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(\mathbb{E}[X_n\mid\mathscr{F}_{n-1}]=0)\\&=\left(\sum_{t=1}^{n-1}X_t\right)^2-\sum_{t=1}^{n-1}\mathbb{E}[X_t^2\mid\mathscr{F}_{t-1}]=Y_{n-1}.\end{aligned}$$

We would now like to show that $\mathbb{E}[Y_\tau] = \mathbb{E}[Y_1] = 0$ using the optional stopping theorem. The latter holds since $\tau \le c$ almost surely and $\mathbb{E}[Y_1] = \mathbb{E}[X_1^2 - \mathbb{E}[X_1^2 \mid \mathscr{F}_0]] = 0$. □

Proof of Theorem 5.4.8

Lemma (restatement of Theorem 5.4.8). With probability at least $1 - \delta/4$,

$$\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in \mathscr{S}} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{V}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\le 614 B_\star \sqrt{MS^2 A \log^2 \frac{TSA}{\delta}} + 8160 B_\star S^2 A \log^2 \frac{TSA}{\delta}.$$

Proof. Recall the following definitions:

$$A_h^m = \frac{\log(SAN_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)}.$$
$$\mathbb{V}_h^m = \sum_{s' \in \mathscr{S}^+} P(s' \mid s_h^m, a_h^m)\left(\widetilde{V}^m(s') - \sum_{s'' \in \mathscr{S}^+} P(s'' \mid s_h^m, a_h^m)\widetilde{V}^m(s'')\right)^2.$$

From Theorem 5.4.6 we have that

$$\mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in \mathscr{S}} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big)\widetilde{V}^m(s')\mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$
$$\le \mathbb{E}\left[16\sqrt{S}\sum_{h=1}^{H^m} \sqrt{\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} + 272 B_\star S A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right].$$

Moreover, by applying the Cauchy-Schwartz inequality twice, we get that

$$\mathbb{E}\left[\sum_{h=1}^{H^m} \sqrt{\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] \le \mathbb{E}\left[\sqrt{\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\}} \cdot \sqrt{\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\}} \,\middle|\, \bar{U}^{m-1}\right]$$
$$\le \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]}$$
$$\le 7B_\star \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]}. \qquad \text{(Theorem 5.4.7)}$$

We sum over all intervals to obtain

$$\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{s' \in \mathscr{S}} \big(P(s' \mid s_h^m, a_h^m) - \widetilde{P}_m(s' \mid s_h^m, a_h^m)\big) \widetilde{V}^m(s') \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \leq$$

$$\leq 112 B_\star \sum_{m=1}^{M} \sqrt{S \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]} + 272 B_\star S \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]$$

$$\leq 112 B_\star \sqrt{MS \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right]} + 272 B_\star S \sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right],$$

where the last inequality follows from Jensen's inequality. We finish the proof using Theorem C.2.16 below. □

Lemma C.2.16. *With probability at least* $1 - \delta/4$, *the following holds for* $M = 1, 2, \ldots$ *simultaneously.*

$$\sum_{m=1}^{M} \mathbb{E}\left[\sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \leq O\left(SA \log^2 \frac{TSA}{\delta}\right).$$

Proof. Define the infinite sequence of random variables: $X^m = \sum_{h=1}^{H^m} A_h^m \mathbb{I}\{\Omega^m\}$ for which $|X^m| \leq 3 \log(SA/\delta)$ due to Theorem C.2.17 below. We apply Equation (C.19) of Theorem C.4.4 to obtain with probability at least $1 - \delta/4$, for all $M = 1, 2, \ldots$ simultaneously

$$\sum_{m=1}^{M} \mathbb{E}\big[X^m \mid \bar{U}^{m-1}\big] \leq 2 \sum_{m=1}^{M} X^m + 12 \log\left(\frac{SA}{\delta}\right) \log\left(\frac{8M}{\delta}\right).$$

Now, we bound the sum over X^m by rewriting it as a sum over epochs:

$$\sum_{m=1}^{M} X^m \leq \sum_{m=1}^{M} \sum_{h=1}^{H^m} \frac{\log(SAN_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} \leq \log \frac{SAT}{\delta} \sum_{s \in \mathscr{S}} \sum_{a \in A} \sum_{i=1}^{E} \frac{n_i(s,a)}{N_+^i(s,a)},$$

where E is the last epoch. Finally, from Theorem C.2.18 below we have that for every $(s,a) \in \mathscr{S} \times \mathscr{A}$,

$$\sum_{i=1}^{E} \frac{n_i(s,a)}{N_+^i(s,a)} \leq 2 \log N_{E+1}(s,a) \leq 2 \log T.$$

We now plugin the resulting bound for $\sum_{m=1}^{M} X^m$ and simplify the acquired expression by using $M \leq T$. □

Lemma C.2.17. *For any interval* m, $|\sum_{h=1}^{H^m} A_h^m| \leq 3 \log(SA/\delta)$.

Proof. Note that all state-action pairs (s_h^m, a_h^m) (except the first one (s_1^m, a_1^m)) are known. Hence, for $h \geq 2$, $N_+^m(s_h^m, a_h^m) \geq 30000 \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$. Therefore, since $\log(x)/x$ is decreasing and since $S \geq 2$ and $A \geq 2$ by assumption,

$$\begin{aligned}
\sum_{h=1}^{H^m} \frac{\log(SAN_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} &\leq \frac{\log(SAN_+^m(s_1^m, a_1^m)/\delta)}{N_+^m(s_1^m, a_1^m)} + \sum_{h=2}^{H^m} \frac{\log(SAN_+^m(s_h^m, a_h^m)/\delta)}{N_+^m(s_h^m, a_h^m)} \\
&\leq \log(SA/\delta) + \frac{c_{\min} H^m}{B_\star} \\
&\leq \log(SA/\delta) + 2 \qquad\qquad (H^m \leq \tfrac{2B_\star}{c_{\min}} \text{ by definition.}) \\
&\leq 3\log(SA/\delta).
\end{aligned}$$

□

Lemma C.2.18. *For any sequence of integers $z_1, \dots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ and $Z_0 = 1$, it holds that*

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq 2\log Z_n.$$

Proof. We use the inequality $x \leq 2\log(1+x)$ for every $0 \leq x \leq 1$ to obtain

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq 2\sum_{k=1}^{n} \log\left(1 + \frac{z_k}{Z_{k-1}}\right) = 2\sum_{k=1}^{n} \log\frac{Z_{k-1} + z_k}{Z_{k-1}} = 2\sum_{k=1}^{n} \log\frac{Z_k}{Z_{k-1}} = 2\log\prod_{k=1}^{n} \frac{Z_k}{Z_{k-1}} = 2\log Z_n.$$

□

Proof of Theorem 5.2.3

Theorem (restatement of Theorem 5.2.3). Assume that Theorem 5.2.1 holds. With probability at least $1-\delta$ the regret of Algorithm 7 is bounded as follows:

$$R_K = O\left(B_\star S\sqrt{AK} \log\frac{KB_\star SA}{\delta c_{\min}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}}} \log^2\frac{KB_\star SA}{\delta c_{\min}}\right).$$

Proof. Let C_M denote the cost of the learner after M intervals. First, with probability at least $1-\delta$, we have Theorems 5.4.2, 5.4.5 and 5.4.8 via a union bound. Now, as Ω^m holds for all intervals, we have $\widetilde{R}_M = R_M$ for any number of intervals M. Plugging in the bounds of Theorems 5.4.4, 5.4.5 and 5.4.8 into Theorem 5.4.3, we have that for any number of intervals M:

$$C_M = O\left(K \cdot V^\star(s_{\text{init}}) + B_\star\sqrt{MS^2A\log^2\frac{TSA}{\delta}} + B_\star S^2 A \log^2\frac{TSA}{\delta}\right).$$

We now plug in the bounds on M and T from Theorem 5.4.1 into the bound above. First, we plug in the bound on M. As long as the K episodes have not elapsed we have that $M \le O\big(C_M/B_\star + K + 2SA\log T + \frac{B_\star S^2 A}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}\big)$. This gets after using the subadditivity of the square root to simplify the resulting expression,

$$\begin{aligned} C_M = O\Bigg(& K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} \\ & + \sqrt{B_\star C_M S^2 A \log^2 \frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}} \Bigg). \end{aligned}$$

From which, by solving for C_M (using that $x \le a\sqrt{x} + b$ implies $x \le (a+\sqrt{b})^2$ for $a \ge 0$ and $b \ge 0$), and simplifying the resulting expression by applying $V^\star(s_{\text{init}}) \le B_\star$ and our assumptions that $K \ge S^2A$, $S \ge 2$, $A \ge 2$, we get that

$$\begin{aligned} C_M &= O\Bigg(\Bigg(\sqrt{B_\star S^2 A \log^2 \frac{TSA}{\delta}} \\ &\quad + \sqrt{K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}}} \Bigg)^2 \Bigg) \\ &= O\Bigg(B_\star S^2 A \log^2 \frac{TSA}{\delta} \\ &\quad + \sqrt{B_\star S^2 A \log^2 \frac{TSA}{\delta}} \cdot \sqrt{K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}}} \\ &\quad + K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}} \Bigg) \\ &= O\Bigg(B_\star S^2 A \log^2 \frac{TSA}{\delta} + B_\star \sqrt{K^{1/4} S^3 A^{3/2} \log^3 \frac{TSA}{\delta}} + \sqrt{\frac{B_\star^{5/2} S^4 A^2}{c_{\min}^{1/2}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}} \\ &\quad + K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}} \Bigg) \end{aligned}$$

$$= O\Bigg(K \cdot V^\star(s_{\text{init}}) + B_\star \sqrt{KS^2A\log^2\frac{TSA}{\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}} \log^4 \frac{TB_\star SA}{c_{\min}\delta}} \Bigg). \tag{C.12}$$

Note that in particular, by simplifying the bound above, we have $C_M = O\Big(\sqrt{B_\star^3 S^4 A^2 KT / c_{\min}\delta}\Big)$.

Next we combine this with the fact, stated in Theorem 5.4.1 that $T \leq C_M/c_{\min}$. Isolating T gets $T = O\Big(\frac{B_\star^3 S^4 A^2 K}{c_{\min}^3 \delta}\Big)$, and plugging this bound back into Equation (C.12) and simplifying gets us

$$C_M = O\bigg(K \cdot V^\star(s_{\text{init}}) + B_\star S\sqrt{AK\log^2\frac{KB_\star SA}{c_{\min}\delta}} + \sqrt{\frac{B_\star^3 S^4 A^2}{c_{\min}}\log^4\frac{KB_\star SA}{c_{\min}\delta}}\bigg).$$

Finally, we note that the bound above holds for any number of intervals M as long as K episodes do not elapse. As the instantaneous costs in the model are positive, this means that the learner must eventually finish the K episodes from which we derive the bound for R_K claimed by the theroem. □

C.3 Lower Bound

In this section we prove Theorem 5.2.6. At first glance, it is tempting to try and use the lower bound of [JOA10, Theorem 5] on the regret suffered against learning average-reward MDPs by reducing any problem instance from an average-reward MDP to an instance of SSP. However, it is unclear to us if such a reduction is possible, and if it is, how to perform it.[1] We consequently prove the theorem here directly.

By Yao's minimax principle, in order to derive a lower bound on the learner's regret, it suffices to show a distribution over MDP instances that forces any deterministic learner to suffer a regret of $\Omega(B_\star\sqrt{SAK})$ in expectation.

To simplify our arguments, let us first consider the following simpler problem before considering the problem in its full generality. Think of a simple MDP with two states: the initial state and a goal state. The set of actions A has a special action $a^\star$ chosen uniformly at random a-priori. Upon choosing the special action, the learner transitions to the goal state with probability $\approx 1/B_\star$ and remains at s_{init} with the remaining probability. Concretely $P(g \mid a^\star) = 1/B_\star$ and $P(s_{\text{init}} \mid a^\star) = 1 - 1/B_\star$, and for any other action $a \neq a^\star$ we have $P(g \mid a) = (1-\varepsilon)/B_\star$ and $P(s_{\text{init}} \mid a) = 1-(1-\varepsilon)/B_\star$ for some $\varepsilon \in (0, 1/8)$.[2] The costs of all actions equal 1; i.e., $c(s_{\text{init}}, a) = 1$ for all $a \in A$. Clearly, the optimal policy constantly plays $a^\star$ and therefore $V^\star(s_{\text{init}}) = B_\star$.

Fix any deterministic learning algorithm, we shall now quantify the regret of the learner during a single episode in terms of the number of times that it chooses $a^\star$. Let

[1] Even though a reduction in the reverse direction is fairly straight-forward in the unit-cost case [TGV$^+$20].

[2] For ease of notation and since there is only one state other than g, we do not write this state as the origin state in the definition of the transition function.

N_k denote the number of steps that the learner spends in s_{init} during episode k, and let $N_k^\star$ be the number of times the learner plays $a^\star$ at s_{init} during the episode. Note that N_k is also the total cost that the learning algorithm suffered during episode k. We have the following lemma.

Lemma C.3.1. $\mathbb{E}\big[N_k\big] - V^\star(s_{init}) = \varepsilon \cdot \mathbb{E}\big[N_k - N_k^\star\big]$.

Proof. Let us denote by $s_1, s_2, \dots$ and $a_1, a_2, \dots$ the sequences of states and actions observed by the learner during the episode. We have,

$$\begin{aligned}
\mathbb{E}[N_k] &= \sum_{t=1}^{\infty} \Pr[s_t = s_{\text{init}}] \\
&= 1 + \sum_{t=2}^{\infty} \Pr[s_t = s_{\text{init}}] \\
&= 1 + \sum_{t=2}^{\infty} \Pr[s_t = s_{\text{init}} \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] \Pr[s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] \\
&\qquad + \sum_{t=2}^{\infty} \Pr[s_t = s_{\text{init}} \mid s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star] \Pr[s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star] \\
&= 1 + \sum_{t=2}^{\infty} \left(1 - \frac{1}{B_\star}\right) \Pr[s_{t-1} = s_{\text{init}}, a_{t-1} = a^\star] + \sum_{t=2}^{\infty} \left(1 - \frac{1-\varepsilon}{B_\star}\right) \Pr[s_{t-1} = s_{\text{init}}, a_{t-1} \neq a^\star] \\
&= 1 + \left(1 - \frac{1}{B_\star}\right) \sum_{t=1}^{\infty} \Pr[s_t = s_{\text{init}}, a_t = a^\star] + \left(1 - \frac{1-\varepsilon}{B_\star}\right) \sum_{t=1}^{\infty} \Pr[s_t = s_{\text{init}}, a_t \neq a^\star] \\
&= 1 + \left(1 - \frac{1}{B_\star}\right) \mathbb{E}[N_k^\star] + \left(1 - \frac{1-\varepsilon}{B_\star}\right) \mathbb{E}[N_k - N_k^\star].
\end{aligned}$$

Rearranging using $V^\star(s_{\text{init}}) = B_\star$ gives the Lemma's statement. □

By Theorem C.3.1 the overall regret of the learner over K episodes is: $\mathbb{E}[R_K] = \varepsilon \cdot \mathbb{E}\big[N - N^\star\big]$, where $N = \sum_{k=1}^K N_k$ and $N^\star = \sum_{k=1}^K N_k^\star$. In words, the regret of the learner is ε times the expected number of visits to s_{init} in which the learner did not play $a^\star$.

In the remainder of the proof we lower bound N in expectation and upper bound the expected value of $N^\star$. To upper bound $N^\star$, we use standard techniques from lower bounds of multi-armed bandits [ACBFS02] that bound the total variation distance between the distribution of the sequence of states traversed by the learner in the original MDP and that generated in a "uniform MDP" in which all actions are identical. However, we cannot apply this argument directly since it requires $N^\star$ to be bounded almost surely, yet here $N^\star$ depends on the total length of all K episodes which is unbounded in general. We fix this issue by looking only on the first T steps (where T is to be determined) and showing that the regret is large even in these T steps.

Formally, we view the run of the K episodes as a continuous process in which when the learner reaches the goal state we transfer it to s_{init} (at no cost) and let it restart from there. Furthermore, we *cap* the learning process to consist of exactly T steps as follows. If the K episodes are completed before T steps are elapsed, the learner remains in g (until completing T steps) without suffering any additional cost, and otherwise we stop the learner after T steps before it completes its K episodes. In this capped process, we denote the number of visits in s_{init} by N_- and the number of times the learner played $a^\star$ in s_{init} by $N_-^\star$. We have

$$\mathbb{E}[R_K] \geq \varepsilon \cdot \big(\mathbb{E}\big[N_-\big] - \mathbb{E}\big[N_-^\star\big]\big). \tag{C.13}$$

The number of visits to s_{init} under this capping is lower bounded by the following lemma.

Lemma C.3.2. *For any deterministic learner, if $T \geq 2KB_\star$ then we have that $\mathbb{E}\big[N_-\big] \geq KB_\star/4$.*

Proof. If the capped learner finished its K episodes then $N_- = N$. Otherwise, it visits the goal state less than K times and therefore $N_- \geq T - K$. Hence $\mathbb{E}\big[N_-\big] \geq \mathbb{E}\big[\min\{T - K, N\}\big] \geq \sum_{k=1}^{K} \mathbb{E}\big[\min\{T/K - 1, N_k\}\big]$. Since $T \geq 2KB_\star$, the lemma will follow if we show that $N_k \geq B_\star$ with probability at least $1/4$. We lower bound the probability that $N_k \geq B_\star$ by the probability of staying at s_{init} for $B_\star$ steps and picking $a^\star$ in the first $B_\star - 1$ steps. Indeed, using $(1-1/x)^{x-1} \geq 1/e$ for $x \geq 1$, we get that $\Pr[N_k \geq B_\star] \geq \big(1 - \frac{1}{B_\star}\big)^{B_\star - 1} \geq \frac{1}{4}$. □

We now introduce an additional distribution of the transitions which call $\Pr_{\text{unif}}$. $\Pr_{\text{unif}}$ is identical to Pr as defined above, except that $P(g \mid a) = (1-\varepsilon)/B_\star$ for all actions a. We denote expectations over $\Pr_{\text{unif}}$ by $\mathbb{E}_{\text{unif}}$. The following lemma uses standard lower bound techniques used for multi-armed bandits (see, e.g., [JOA10, Theorem 13]) to bound the difference in the expectation of $N_-^\star$ when the learner plays in Pr compared to when it plays in $\Pr_{\text{unif}}$.

Lemma C.3.3. *For any deterministic learner we have that $\mathbb{E}\big[N_-^\star\big] \leq \mathbb{E}_{unif}\big[N_-^\star\big] + \varepsilon T\sqrt{\mathbb{E}_{unif}[N_-^\star]/B}$.*

Proof. Fix any deterministic learner. Let us denote by $s^{(t)}$ the sequence of states observed by the learner up to time t and including. Now, as $N_-^\star \leq T$ and the fact that $N_-^\star$ is a function of $s^{(T)}$, $\mathbb{E}\left[N_{-}^{\star}\right] \leq \mathbb{E}_{\text{unif}}\left[N_{-}^{\star}\right] + T \cdot \mathrm{TV}(\Pr_{\text{unif}}[s^{(T)}], \Pr[s^{(T)}])$. Pinsker's inequality yields

$$\mathrm{TV}(\Pr_{\text{unif}}[s^{(T)}], \Pr[s^{(T)}]) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\Pr_{\text{unif}}[s^{(T)}] \parallel \Pr[s^{(T)}])}. \tag{C.14}$$

Next, the chain rule of the KL divergence obtains

$$\mathrm{KL}(\Pr_{\text{unif}}[s^{(T)}] \parallel \Pr[s^{(T)}]) = \sum_{t=1}^{T} \sum_{s^{(t-1)}} \Pr_{\text{unif}}[s^{(t-1)}] \cdot \mathrm{KL}(\Pr_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \Pr[s_t \mid s^{(t-1)}]).$$

Observe that at any time, since the learning algorithm is deterministic, the learner chooses an action given $s^{(t-1)}$ regardless of whether $s^{(t-1)}$ was generated under Pr or under $\Pr_{\text{unif}}$. Thus, the $\mathrm{KL}(\Pr_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \Pr[s_t \mid s^{(t-1)}])$ is zero if $a_{t-1} \neq a_{\star}$, and otherwise

$$\begin{aligned}
&\mathrm{KL}(\Pr_{\text{unif}}[s_t \mid s^{(t-1)}] \parallel \Pr[s_t \mid s^{(t-1)}]) = \\
&= \sum_{s \in \mathscr{S}} \Pr_{\text{unif}}[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^{\star}] \log \frac{\Pr_{\text{unif}}[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^{\star}]}{\Pr[s_t \mid s_{t-1} = s_{\text{init}}, a_{t-1} = a^{\star}]} \\
&= \frac{1-\varepsilon}{B_{\star}} \cdot \log(1-\varepsilon) + \left(1 - \frac{1-\varepsilon}{B_{\star}}\right) \log\left(1 + \frac{\varepsilon}{B_{\star}-1}\right) \\
&\leq \frac{\varepsilon^2}{B_{\star}-1}. \qquad \text{(using } \log(1+x) \leq x \text{ for all } x > 0\text{)}
\end{aligned}$$

Plugging the above back into Equation (C.14) and using $B_{\star} \geq 2$ gives the lemma. □

In the following result, we combine the lemma above with standard techniques from lower bounds of multi-armed bandits (see [JOA10, Thm. 5] for example).

Theorem C.3.4. *Suppose that $B_{\star} \geq 2$, $\varepsilon \in (0, \frac{1}{8})$ and $A \geq 16$. For the problem described above we have that*

$$\mathbb{E}[R_K] \geq \varepsilon K B_{\star}\left(\frac{1}{8} - 2\varepsilon\sqrt{\frac{2K}{A}}\right).$$

Proof of Theorem C.3.4. Note that as under $\Pr_{\text{unif}}$ the transition distributions are identical for all actions, we have that

$$\sum_{a \in A | a^{\star} = a} \mathbb{E}_{\text{unif}}\left[N_{-}^{\star}\right] = \mathbb{E}_{\text{unif}}\left[\sum_{a \in A | a^{\star} = a} N_{-}^{\star}\right] = \mathbb{E}_{\text{unif}}\left[N_{-}\right] \leq T. \tag{C.15}$$

Suppose that $a^{\star}$ is sampled uniformly at random before the game starts. Denote the probability and expectation with respect to the distribution induced by a specific choice of

$a^\star = a$ by $\Pr_a$ and $\mathbb{E}_a$ respectively. Then for $T = 2KB_\star$,

$$
\begin{aligned}
\mathbb{E}[R_K] &= \frac{1}{A}\sum_{a\in A}\mathbb{E}_a[R_K] \\
&\geq \frac{1}{A}\sum_{a\in A}\mathbb{E}_a[N_- - N_-^\star] && \text{(Equation (C.13))} \\
&\geq \frac{1}{A}\sum_{a\in A|a_\star=a}\left(\frac{KB_\star}{4} - \mathbb{E}_{\text{unif}}[N_-^\star] - \varepsilon T\sqrt{\frac{\mathbb{E}_{\text{unif}}[N_-^\star]}{B_\star}}\right) && \text{(Theorems C.3.2 and C.3.3)} \\
&\geq \frac{KB_\star}{4} - \frac{1}{A}\sum_{a\in A|a_\star=a}\mathbb{E}_{\text{unif}}[N_-^\star] - \varepsilon T\sqrt{\frac{1}{B_\star}\cdot\frac{1}{A}\sum_{a\in A|a_\star=a}\mathbb{E}_{\text{unif}}[N_-^\star]} \\
& && \text{(Jensen's inequality)} \\
&\geq \frac{KB_\star}{4} - \frac{T}{A} - \varepsilon T\sqrt{\frac{T}{B_\star A}} && \text{(Equation (C.15))} \\
&= \varepsilon\left(\frac{KB_\star}{4} - \frac{2KB_\star}{A} - 2\varepsilon KB_\star\sqrt{\frac{2KB_\star}{AB_\star}}\right) \\
&= \varepsilon KB_\star\left(\frac{1}{4} - \frac{2}{A} - 2\varepsilon\sqrt{\frac{2K}{A}}\right).
\end{aligned}
$$

The theorem follows from $A \geq 16$ and by rearranging. □

Proof of Theorem 5.2.6. Consider the following MDP. Let S be the set of states disregarding g. The initial state is sampled uniformly at random from S. Each $s \in \mathscr{S}$ has its own special action $a_s^\star$. The transition distributions are defined $P(g \mid a_s^\star, s) = 1/B_\star$, $P(s \mid a_s^\star, s) = 1 - 1/B_\star$, and $P(g \mid a, s) = (1-\varepsilon)/B_\star$, $P(s \mid a, s) = 1 - (1-\varepsilon)/B_\star$ for any other action $a \in A\setminus\{a_s^\star\}$.

Note that for each $s \in \mathscr{S}$, the learner is faced with a simple problem as the one described above from which it cannot learn about from other states $s' \neq s$. Therefore, we can apply Theorem C.3.4 for each $s \in \mathscr{S}$ separately and lower bound the learner's expected regret the sum of the regrets suffered at each $s \in \mathscr{S}$, which would depend on the number of times $s \in \mathscr{S}$ is drawn as the initial state. Since the states are chosen uniformly at random there are many states (constant fraction) that are chosen $\Theta(K/S)$ times. Summing the regret bounds of Theorem C.3.4 over only these states and choosing ε appropriately gives the sought-after bound.

Denote by K_s the number of episodes that start in each state $s \in \mathscr{S}$.

$$\mathbb{E}[R_K] \geq \sum_{s\in\mathscr{S}} \mathbb{E}\left[\varepsilon K_s B_\star \left(\frac{1}{8} - 2\varepsilon\sqrt{\frac{2K_s}{A}}\right)\right] = \frac{\varepsilon K B_\star}{8} - 2\varepsilon^2 B_\star \sqrt{\frac{2}{A}} \sum_{s\in\mathscr{S}} \mathbb{E}[K_s^{3/2}]. \tag{C.16}$$

Taking expectation over the initial states and applying Cauchy-Schwartz inequality gives

$$\begin{aligned}\sum_{s\in\mathscr{S}} \mathbb{E}\left[K_s^{3/2}\right] &\leq \sum_{s\in\mathscr{S}} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s^2]} = \sum_{s\in\mathscr{S}} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s]^2 + \mathbb{V}[K_s]} \\ &= \sum_{s\in\mathscr{S}} \sqrt{\frac{K}{S}}\sqrt{\frac{K^2}{S^2} + \frac{K(S-1)}{S^2}} \leq K\sqrt{\frac{2K}{S}},\end{aligned}$$

where we have used the expectation and variance formulas of the Binomial distribution. The lower bound is now given by applying the inequality above in Equation (C.16) and choosing $\varepsilon = \frac{1}{64}\sqrt{AS/K}$. □

C.4 Concentration inequalities

Theorem C.4.1 (Anytime Azuma)**.** *Let $(X_n)_{n=1}^\infty$ be a martingale difference sequence with respect to the filtration $(\mathscr{F}_n)_{n=0}^\infty$ such that $|X_n| \leq B$ almost surely. Then with probability at least $1-\delta$,*

$$\left|\sum_{i=1}^n X_i\right| \leq B\sqrt{n \log\frac{2n}{\delta}}, \qquad \forall n \geq 1.$$

Theorem C.4.2 ([WOS$^+$03])**.** *Let $p(\cdot)$ be a distribution over m elements, and let $\bar{p}_t(\cdot)$ be the empirical distribution defined by t iid samples from $p(\cdot)$. Then, with probability at least $1-\delta$,*

$$\left\|\bar{p}_t(\cdot) - p(\cdot)\right\|_1 \leq 2\sqrt{\frac{m\log\frac{1}{\delta}}{t}}.$$

Theorem C.4.3 (Anytime Bernstein)**.** *Let $(X_n)_{n=1}^\infty$ be a sequence of i.i.d. random variables with expectation μ. Suppose that $0 \leq X_n \leq B$ almost surely. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\mu n\log\frac{2n}{\delta}} + B\log\frac{2n}{\delta}. \tag{C.17}$$

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\sum_{i=1}^n X_i \log\frac{2n}{\delta}} + 7B\log\frac{2n}{\delta}. \tag{C.18}$$

Proof. Fix some $n \geq 1$. By Bernstein's concentration inequality (see for example, [CBL06, Corollary A.3]), we have with probability at least $1 - \frac{\delta}{2n^2}$ that Equation (C.17) holds. By a union bound, the inequality holds with probability at least $1 - \delta$ for all $n \geq 1$ simultaneously.

To show Equation (C.18), note that in particular we have

$$\mu \cdot n - \sum_{i=1}^{n} X_i \leq 2\sqrt{B\mu n \log \frac{2n}{\delta}} + B \log \frac{2n}{\delta}$$

that is a quadratic inequality in μ. This implies that $\sqrt{\mu} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i} + 3\sqrt{\frac{B \log \frac{2n}{\delta}}{n}}$. Plugging this inequality back into the RHS of Equation (C.17) gets us Equation (C.18). □

Lemma C.4.4. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables with expectation adapted to the filtration $(\mathscr{F}_n)_{n=0}^{\infty}$. Suppose that $0 \leq X_n \leq B$ almost surely. Then with probability at least $1 - \delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathscr{F}_{i-1}] \leq 2\sum_{i=1}^{n} X_i + 4B \log \frac{2n}{\delta}. \tag{C.19}$$

Proof. For all $n \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[e^{-X_n/B} \mid \mathscr{F}_{n-1}] &\leq \mathbb{E}\left[1 - \frac{X_n}{B} + \frac{X_n^2}{2B^2} \;\middle|\; \mathscr{F}_{n-1}\right] && (e^{-x} \leq 1 - x + \tfrac{x^2}{2} \text{ for all } x \geq 0) \\
&\leq 1 - \frac{\mathbb{E}[X_n \mid \mathscr{F}_{n-1}]}{B} + \frac{\mathbb{E}[X_n \mid \mathscr{F}_{n-1}]}{2B} && (X_n \leq B) \\
&= 1 - \frac{\mathbb{E}[X_n \mid \mathscr{F}_{n-1}]}{2B} \leq e^{-\mathbb{E}[X_n \mid \mathscr{F}_{n-1}]/2B}. && (1 - x \leq e^{-x} \text{ for all } x)
\end{aligned}$$

Hence, fix some $n \geq 1$, then

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right)\right)\right] \\
&\quad = \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n-1}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right)\right) \cdot \underbrace{\mathbb{E}\left[\exp\left(\frac{1}{B}\left(\frac{1}{2}\mathbb{E}[X_n \mid \mathscr{F}_{n-1}] - X_n\right) \;\middle|\; \mathscr{F}_{n-1}\right)\right]}_{\leq 1}\right] \\
&\quad \leq \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n-1}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right)\right)\right] \\
&\quad \leq 1. && \text{(by repeating the last argument inductively.)}
\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr\left[\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right) > 2B\log\frac{2n}{\delta}\right] &\leq \Pr\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right)\right) > \frac{2n^2}{\delta}\right] \\ &\leq \mathbb{E}\left[\exp\left(\frac{1}{B}\sum_{i=1}^{n}\left(\frac{1}{2}\mathbb{E}[X_i \mid \mathscr{F}_{i-1}] - X_i\right)\right)\right]\cdot\frac{\delta}{2n^2} \\ &\qquad\qquad \text{(Markov inequality)} \\ &\leq \frac{\delta}{2n^2}.\end{aligned}$$

Hence the above holds for all $n \geq 1$ via a union bound which provides the lemma. □

D Supplementary Material for Chapter 6

D.1 Proofs for Section 6.3

D.1.1 Proof of Lemma 6.3.1

In this section we relate the SSP regret and the finite-horizon regret, which relies on Theorems D.1.1 and D.1.2 below that compare the cost-to-go function in the SSP $\mathscr{M}$ to the value function in the finite-horizon $\widehat{\mathscr{M}}$. To that end, we define a cost-to-go function with respect to the finite-horizon MDP $\widehat{\mathscr{M}}$ as: $\widehat{V}_h^\pi(s) = \mathbb{E}\big[\sum_{h'=h}^H c(s_{h'},a_{h'}) \mid s_h = s\big]$, for any deterministic finite-horizon policy $\pi : \mathscr{S} \times [H] \to \mathscr{A}$.

Lemma D.1.1. *Let π be a stationary policy. For every $s \in \widehat{\mathscr{S}}$ and $h = 1,\dots,H+1$ it holds that*

$$\widehat{V}_h^\pi(s) \le V^\pi(s) + 8B_\star \Pr[s_{H+1} \ne g \mid s_h = s, \widehat{P}, \pi].$$

Proof.

$$\begin{aligned}
\widehat{V}_h^\pi(s) &= \sum_{h'=h}^H \sum_{s' \in \widehat{\mathscr{S}}} \Pr[s_{h'} = s' \mid s_h = s, \widehat{P}, \pi]\ \hat{c}\big(s', \pi(s')\big) + \sum_{s' \in \widehat{\mathscr{S}}} \Pr[s_{H+1} = s' \mid s_h = s, \widehat{P}, \pi]\ \hat{c}_f(s') \\
&= \sum_{h'=h}^H \sum_{s' \in \mathscr{S}} \Pr[s_{h'} = s' \mid s_h = s, P, \pi]\ c\big(s', \pi(s')\big) + 8B_\star\ \Pr[s_{H+1} \ne g \mid s_h = s, \widehat{P}, \pi] \\
&\le \sum_{h'=h}^\infty \sum_{s' \in \mathscr{S}} \Pr[s_{h'} = s' \mid s_h = s, P, \pi]\ c\big(s', \pi(s')\big) + 8B_\star\ \Pr[s_{H+1} \ne g \mid s_h = s, \widehat{P}, \pi] \\
&= V^\pi(s) + 8B_\star\ \Pr[s_{H+1} \ne g \mid s_h = s, \widehat{P}, \pi].
\end{aligned}$$

□

Lemma D.1.2. *For every $s \in \widehat{\mathscr{S}}$, it holds that $V^\star(s) \ge \widehat{V}_1^{\pi^\star}(s) - \frac{B_\star}{K}$.*

Proof. The probability that $\pi^\star$ does not reach the goal in H steps is at most $1/(8K)$ due to [CLW21, Lemma 7]. Plugging that into Theorem D.1.1 yields the desired result. □

Proof of Theorem 6.3.1. Consider the first interval of the first episode. If it ends in the goal state then

$$\sum_{i=1}^{I^1} C_i^1 = \sum_{h=1}^{H} C_h^1 + \hat{c}_f(g) = \sum_{h=1}^{H} C_h^1 + \hat{c}_f(s_{H+1}^1).$$

If the agent did not reach g in the first interval, then the agent also suffered the $8B_\star$ terminal cost and thus

$$\begin{aligned}\sum_{i=1}^{I^1} C_i^1 &= \sum_{h=1}^{H} C_h^1 + \hat{c}_f(s_{H+1}^1) + \sum_{i=H+1}^{I^1} C_i^1 - \hat{c}_f(s_{H+1}^1) \\ &= \sum_{h=1}^{H} C_h^1 + \hat{c}_f(s_{H+1}^1) + \sum_{i=H+1}^{I^1} C_i^1 - 8B_\star \\ &\le \sum_{h=1}^{H} C_h^1 + \hat{c}_f(s_{H+1}^1) + \sum_{i=H+1}^{I^1} C_i^1 - \widehat{V}_1^{\pi^\star}(s_{H+1}^1),\end{aligned}$$

where the last inequality follows by combining Theorem D.1.2 with our assumption that $V^\star(s) \le B_\star$.

Repeating this argument iteratively we get, for every episode k,

$$\begin{aligned}\sum_{i=1}^{I^k} C_i^k - V^\star(s_{\text{init}}) &\le \sum_{i=1}^{I^k} C_i^k - \widehat{V}_1^{\pi^\star}(s_1^m) + \frac{B_\star}{K} \\ &\le \sum_{m\in M_k} \sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}_1^{\pi^\star}(s_1^m) + \frac{B_\star}{K} \\ &= \sum_{m\in M_k} \left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}^{\pi^m}(s_1^m) \right) + \sum_{m\in M_k} \left(\widehat{V}^{\pi^m}(s_1^m) - \widehat{V}_1^{\pi^\star}(s_1^m) \right) + \frac{B_\star}{K},\end{aligned}$$

where M_k is the set of intervals that are contained in episode k, and the first inequality follows from Theorem D.1.2. Summing over all episodes obtains

$$R_K \le \sum_{m=1}^{M} \left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}^{\pi^m}(s_1^m) \right) + \sum_{m=1}^{M} \left(\widehat{V}^{\pi^m}(s_1^m) - \widehat{V}_1^{\pi^\star}(s_1^m) \right) + \frac{B_\star}{K}.$$

Notice that the second summand in the bound above is exactly the expected finite-horizon regret over the M intervals. We finish the proof of the lemma by using the regret guarantees of `ALG` (Definition 6.2.1). □

D.1.2 Proof of Lemma 6.3.2

In this section we bound the deviation of the actual cost in each interval from its expected value. To do that, we apply Theorem D.1.3 below to bound the second moment of the cumulative cost in an interval up until an unknown state-action pair or the goal state were reached. Here $\bar{U}^m$ denotes the union of all information prior to the m^{th} interval together with the first state of the m^{th} interval (more formally, $\{\bar{U}^m\}_{m\geq 1}$ is a filtration). Moreover, we denote by h_m the last time step before an unknown state-action pair or the goal state were reached in interval m (or H if they were not reached).

Lemma D.1.3. *Let m be an interval and assume that the reduction is performed using an admissible algorithm* `ALG`*. If the good event of* `ALG` *holds until the beginning of interval m, then the agent reaches the goal state or an unknown state-action pair with probability at least $\frac{1}{2}$. Moreover, denote by $C^m = \sum_{h=1}^{h_m} C_h^m + \hat{c}_f(s_{H+1}^m)\mathbb{I}\{h_m = H\}$ the cumulative cost in the interval until time h_m. Then, $\mathbb{E}[(C^m)^2 \mid \bar{U}^m] \leq 2\cdot 10^5 B_\star^2 + 4B_\star$.*

Proof. The result is given by bounding the total expected cost suffered by the agent in another MDP (defined below) where all unknown state-action pairs are contracted with the goal state. The cost in this MDP is exactly C^m by definition.

Let π^m be the optimistic policy chosen by the algorithm for interval m. Consider the following finite-horizon MDP $\widehat{\mathscr{M}}^m = (\widehat{\mathscr{S}}, A, \widehat{P}^m, H, \hat{c}, \hat{c}_f)$ that contracts unknown state-action pairs with the goal:

$$\widehat{P}_h^m(s' \mid s,a) = \begin{cases} 0, & (s', \pi_{h+1}^m(s')) \text{ is unknown;} \\ P(s' \mid s,a), & s' \neq g \text{ and } (s', \pi_{h+1}^m(s')) \text{ is known;} \\ 1 - \sum_{s'' \in \widehat{\mathscr{S}} \setminus \{g\}} \widehat{P}_h^m(s'' \mid s,a), & s' = g. \end{cases}$$

Denote by V^m the cost-to-go function of π^m in the finite-horizon MDP $\widehat{\mathscr{M}^m}$. Further, let $\widetilde{P}'^m$ be the transition function induced by $\widetilde{P}^m$ in the MDP $\widehat{\mathscr{M}^m}$ similarly to $\widehat{P}^m$, and $\widetilde{V}^m$ the cost-to-go function of π^m with respect to $\widetilde{P}'^m$ (and with cost function $\tilde{c}^m$). Notice that π^m can only reach the goal state quicker in $\widehat{\mathscr{M}^m}$ than in $\widehat{\mathscr{M}}$, so that $\widetilde{V}_h^m(s) \leq \underline{\mathrm{V}}_h^m(s) \leq \widehat{V}_h^{\pi^\star}(s)$ for any $s \in \widehat{\mathscr{S}}$. By the value difference lemma (see, e.g., [SERM20]), for every s, h

such that $(s,\pi^m_h(s))$ is known,

$$
\begin{aligned}
V^m_h(s) &= \widetilde{V}^m_h(s) + \sum_{h'=h}^{H} \mathbb{E}\Big[\hat{c}(s_{h'},a_{h'}) - \tilde{c}^m_{h'}(s_{h'},a_{h'}) \mid s_h = s, \widehat{P}^m, \pi^m\Big] \\
&\quad + \sum_{h'=h}^{H} \mathbb{E}\Big[\big(\widehat{P}^m_{h'}(\cdot \mid s_{h'},a_{h'}) - \widetilde{P}'^m_{h'}(\cdot \mid s_{h'},a_{h'})\big) \cdot \widetilde{V}^m \mid s_h = s, \widehat{P}^m, \pi^m\Big] \\
&\le \widetilde{V}^m_h(s) + H \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} |c(s,\pi^m_{h'}(s)) - \tilde{c}^m_{h'}(s,\pi^m_{h'}(s))| + H\|\widetilde{V}^m\|_\infty \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} \|\widehat{P}^m_{h'}(\cdot|s,\pi^m_{h'}(s)) - \widetilde{P}'^m_{h'}(\cdot|s,\pi^m_{h'}(s))\|_1 \\
&\overset{(a)}{\le} \widehat{V}^{\pi^\star}_h(s) + H \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} |c(s,\pi^m_{h'}(s)) - \tilde{c}^m_{h'}(s,\pi^m_{h'}(s))| \\
&\quad + H\|\widehat{V}^{\pi^\star}_h(s)\|_\infty \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} \|\widehat{P}(\cdot|s,\pi^m_{h'}(s)) - \widetilde{P}^m(\cdot|s,\pi^m_{h'}(s))\|_1 \\
&\le \widehat{V}^{\pi^\star}_h(s) + H \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} |c(s,\pi^m_{h'}(s)) - \tilde{c}^m_{h'}(s,\pi^m_{h'}(s))| + 9HB_\star \max_{\substack{(s,\pi^m_{h'}(s)) \\ \text{known}}} \|\widehat{P}(\cdot|s,\pi^m_{h'}(s)) - \widetilde{P}^m(\cdot|s,\pi^m_{h'}(s))\|_1,
\end{aligned}
$$

where the last inequality follows by optimism and since $\widehat{V}^{\pi^\star}_h(s) \le 9B_\star$ (Theorem D.1.1), and (a) follows because

$$
\begin{aligned}
\|\widehat{P}^m_h(\cdot|s,a) - \widetilde{P}'^m_h(\cdot|s,a)\|_1 &= \sum_{\substack{(s',\pi^m_{h+1}(s')) \\ \text{known}}} |\widehat{P}^m_h(s'|s,a) - \widetilde{P}'^m_h(s'|s,a)| + |\widehat{P}^m_h(g|s,a) - \widetilde{P}'^m_h(g|s,a)| \\
&= \sum_{\substack{(s',\pi^m_{h+1}(s')) \\ \text{known}}} |\widehat{P}(s'|s,a) - \widetilde{P}^m(s'|s,a)| + \Big| \sum_{\substack{(s',\pi^m_{h+1}(s')) \\ \text{unknown}}} \widehat{P}(s'|s,a) + \widehat{P}(g|s,a) - \widetilde{P}^m(s'|s,a) - \widetilde{P}^m(g|s,a) \Big| \\
&\le \|\widehat{P}(\cdot|s,a) - \widetilde{P}^m(\cdot|s,a)\|_1.
\end{aligned}
$$

Thus $V^m_h(s) \le \widehat{V}^{\pi^\star}_h(s) + 2B_\star$ since the number of visits to each known state-action pair is at least $\omega_{\texttt{ALG}} \log \frac{MHSA}{\delta}$ and by property (iv) of admissible algorithms (Definition 6.2.1). Also note that $V^m_h(s) \le 11B_\star$ by Theorem D.1.1, and for $h = 1$ in particular we use Theorem D.1.2 to obtain $V^m_1(s) \le 4B_\star$.

By Markov inequality, the probability that the agent suffers a cost of more than $8B_\star$ in $\widehat{\mathcal{M}^m}$ is at most $\frac{1}{2}$. Notice that all costs are non-negative and there is a terminal cost of $8B_\star$ in all states but the goal, therefore the agent cannot suffer a cost of less than $8B_\star$ unless she reaches the goal. So the probability to reach the goal is at least $\frac{1}{2}$. Moreover, note that the probability to reach the goal in $\widehat{\mathcal{M}^m}$ is equal to the probability to reach the goal or an unknown state-action pair in $\widehat{\mathcal{M}}$.

Similarly, we notice that $\mathbb{E}[(C^m)^2 \mid \bar{U}^m] = \mathbb{E}[(\widehat{C})^2]$, where $\widehat{C}$ is the cumulative cost in

$\widehat{\mathcal{M}}^m$, and we override notation by denoting $\widehat{C} = \sum_{h=1}^{H} C_h + \hat{c}_f(s_{H+1})$. We have that,

$$\begin{aligned}
\mathbb{E}[(\widehat{C})^2] &= \mathbb{E}\left[\left(\sum_{h=1}^{H} C_h + \hat{c}_f(s_{H+1})\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{h=1}^{H-1} C_h + \hat{c}(s_H,a_H) + \hat{c}_f(s_{H+1})\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\left(\sum_{h=1}^{H-1} C_h + \hat{c}(s_H,a_H) + \hat{c}_f(s_{H+1})\right)(C_H - \hat{c}(s_H,a_H))\right] + \mathbb{E}[(C_H - \hat{c}(s_H,a_H))^2].
\end{aligned}$$

The second summand is zero since the realization of C_H is independent of all other randomness given s_H. Also, since $C_H \in [0,1]$, the third summand satisfies

$$\mathbb{E}[(C_H - \hat{c}(s_H,a_H))^2] \le \mathbb{E}[(C_H)^2] \le \mathbb{E}[C_H] = \mathbb{E}[\hat{c}(s_H,a_H)].$$

Thus we arrived at

$$\mathbb{E}[(\widehat{C})^2] \le \mathbb{E}\left[\left(\sum_{h=1}^{H-1} C_h + \hat{c}(s_H,a_H) + \hat{c}_f(s_{H+1})\right)^2\right] + \mathbb{E}[\hat{c}(s_H,a_H)],$$

and iterating this argument yields

$$\mathbb{E}[(\widehat{C})^2] \le \mathbb{E}\left[\left(\sum_{h=1}^{H} \hat{c}(s_h,a_h) + \hat{c}_f(s_{H+1})\right)^2\right] + \mathbb{E}\left[\sum_{h=1}^{H} \hat{c}(s_h,a_h)\right].$$

Here, the second summand equals $V_1^m(s_1)$ which is at most $4B_\star$.

Next, for the first summand, we split the time steps into Q blocks as follows. We denote by t_1 the first time step in which we accumulated a total cost of at least $11B_\star$ (or $H+1$ if it did not occur), by t_2 the first time step in which we accumulated a total cost of at least $11B_\star$ after t_1, and so on up until $t_Q = H+1$. Then, the first block consists of time steps $t_0 = 1,\dots,t_1-1$, the second block consists of time steps $t_1,\dots,t_2-1$, and so on. Since $V_h^m(s) \le 11B_\star$ we must have $\hat{c}(s_h,a_h) \le 11B_\star$ for all $h = 1,\dots,H$ and thus in every such block the total cost is between $11B_\star$ and $22B_\star$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right)^2\right] &\geq \mathbb{E}\left[\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right]^2 \\
&= \mathbb{E}\left[\sum_{i=0}^{Q-1}\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right]^2 \\
&\geq \mathbb{E}[11B_\star Q]^2 = 121B_\star^2\mathbb{E}[Q]^2,
\end{aligned}$$

by Jensen's inequality. On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right)^2\right] = \mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})-V_1^m(s_1)+V_1^m(s_1)\right)^2\right] \\
&\quad\leq 2\mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})-V_1^m(s_1)\right)^2\right]+2V_1^m(s_1)^2 \\
&\quad\leq 2\mathbb{E}\left[\left(\sum_{i=0}^{Q-1}\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)-V_{t_i}^m(s_{t_i})+V_{t_{i+1}}^m(s_{t_{i+1}})\right)^2\right]+32B_\star^2 \\
&\quad\overset{(a)}{=} 4\mathbb{E}\left[\sum_{i=0}^{Q-1}\left(\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)-V_{t_i}^m(s_{t_i})+V_{t_{i+1}}^m(s_{t_{i+1}})\right)^2\right]+32B_\star^2 \\
&\quad\leq 4\mathbb{E}[Q\cdot(33B_\star)^2]+32B_\star^2 \leq 4356B_\star^2\mathbb{E}[Q]+32B_\star^2.
\end{aligned}$$

For (a) we used the fact that $\mathbb{E}[\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)-V_{t_i}(s_{t_i})+V_{t_{i+1}}(s_{t_{i+1}})]=0$ using the Bellman optimality equations and conditioned on all past randomness up until time t_i, and the fact that t_{i+1} is a (bounded) stopping time by the optional stopping theorem, in the following manner,

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)-V_{t_i}^m(s_{t_i})+V_{t_{i+1}}^m(s_{t_{i+1}})\right] &= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)-V_h^m(s_h)+V_{h+1}^m(s_{h+1})\right] \\
&= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1}\mathbb{E}\Big[\hat{c}(s_h,a_h)-V_h^m(s_h)+V_{h+1}^m(s_{h+1})\;\Big|\;s_1,\ldots,s_h\Big]\right] \\
&= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1}\hat{c}(s_h,a_h)+\mathbb{E}\Big[V_{h+1}^m(s_{h+1})\mid s_h\Big]-V_h^m(s_h)\right]=0.
\end{aligned}$$

Thus, we have $121B_\star^2\mathbb{E}[Q]^2\leq 4356B_\star^2\mathbb{E}[Q]+32B_\star^2$, and solving for $\mathbb{E}[Q]$ we obtain $\mathbb{E}[Q]\leq$

37, so

$$\mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right)^2\right]\leq 2\cdot 10^5 B_\star^2,$$

and therefore

$$\mathbb{E}[(\widehat{C})^2]\leq \mathbb{E}\left[\left(\sum_{h=1}^{H}\hat{c}(s_h,a_h)+\hat{c}_f(s_{H+1})\right)^2\right]+\mathbb{E}\left[\sum_{h=1}^{H}\hat{c}(s_h,a_h)\right]\leq 2\cdot 10^5 B_\star^2+4B_\star.$$

□

Proof of Theorem 6.3.2. Recall that h_m is the last time step before an unknown state-action pair or the goal state were reached (or H if they were not reached) in interval m, and let Ω^m be the event that the good event of algorithm ALG holds up to the beginning of interval m. We start by decomposing the sum as follows

$$\begin{aligned}\sum_{m=1}^{M}\left(\sum_{h=1}^{H}C_h^m+\hat{c}_f(s_{H+1}^m)-\widehat{V}_1^{\pi^m}(s_1^m)\right)\mathbb{I}\{\Omega^m\} &= \sum_{m=1}^{M}\left(\sum_{h=1}^{h_m}C_h^m+c_f(s_{H+1}^m)\mathbb{I}\{h_m=H\}-\widehat{V}_1^{\pi^m}(s_1^m)\right)\mathbb{I}\{\Omega^m\}\\ &\quad+\sum_{m=1}^{M}\left(\sum_{h=h_m+1}^{H}C_h^m+\hat{c}_f(s_{H+1}^m)\mathbb{I}\{h_m\neq H\}\right)\mathbb{I}\{\Omega^m\}.\end{aligned}$$

The second term is trivially bounded by $(H+8B_\star)SA\omega_{\texttt{ALG}}\log\frac{MHSA}{\delta}$ since every state-action pair becomes known after $\omega_{\texttt{ALG}}\log\frac{MHSA}{\delta}$ visits. Next, since

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{h=1}^{h_m}C_h^m+c_f(s_{H+1}^m)\mathbb{I}\{h_m=H\}\right)\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^m\right] &= \mathbb{E}\left[\sum_{h=1}^{h_m}C_h^m+c_f(s_{H+1}^m)\mathbb{I}\{h_m=H\}\;\middle|\;\bar{U}^m\right]\mathbb{I}\{\Omega^m\}\\ &\leq \widehat{V}_1^{\pi^m}(s_1^m)\mathbb{I}\{\Omega^m\},\end{aligned}$$

the first term is bounded by $\sum_{m=1}^{M}X^m$ where

$$X^m=\left(\sum_{h=1}^{h_m}C_h^m+c_f(s_{H+1}^m)\mathbb{I}\{h_m=H\}-\mathbb{E}\left[\sum_{h=1}^{h_m}C_h^m+c_f(s_{H+1}^m)\mathbb{I}\{h_m=H\}\;\middle|\;\bar{U}^m\right]\right)\mathbb{I}\{\Omega^m\}$$

is a martingale difference sequence bounded by $H+8B_\star$ with probability 1. For any fixed $M=m$, by Freedman's inequality (Theorem D.5.1, we have with probability at least $1-\frac{\delta}{8m(m+1)}$,

$$\sum_{m'=1}^{m}X^{m'}\leq \eta\sum_{m'=1}^{m}\mathbb{E}[(X^{m'})^2\mid \bar{U}^{m'}]+\frac{\log(8m(m+1)/\delta)}{\eta}$$

for any $\eta \in (0, 1/(H+8B_\star))$. By Theorem D.1.3, for some universal constant $\alpha > 0$, that

$$\sum_{m'=1}^{m} \mathbb{E}[(X^{m'})^2 \mid \bar{U}^{m'}] \leq \alpha m(B_\star^2 + B_\star),$$

and setting $\eta = \min\left\{\sqrt{\frac{\log(8m(m+1)/\delta)}{(B_\star^2+B_\star)m}}, \frac{1}{H+8B_\star}\right\}$ obtains

$$\sum_{m'=1}^{m} X^{m'} \leq O\Big(\sqrt{(B_\star^2+B_\star)m\log\frac{m}{\delta}} + (H+B_\star)\log\frac{m}{\delta}\Big).$$

Taking a union bound on all values of $m = 1, 2, \ldots$ that the inequality above holds for all such values of m simultaneously with probability at least $1 - \delta/8$. In particular, with probability at least $1 - \delta/8$, we have

$$\sum_{m=1}^{M} X^{m} \leq O\Big(\sqrt{(B_\star^2+B_\star)M\log\frac{M}{\delta}} + (H+B_\star)\log\frac{M}{\delta}\Big).$$

The proof is concluded via a union bound—both Freedman inequality and the good event of ALG hold with probability at least $1 - \frac{3}{8}\delta$, and this implies that $\mathbb{I}\{\Omega^m\} = 1$ for every m. □

D.1.3 Proof of Lemma 6.3.3

In this section we bound the number of intervals M with high probability for any admissible algorithm. To that end, we first define the notion of unknown state-action pairs. A state-action pair is defined as *unknown* if the number of times it was visited is at most $\omega_{\texttt{ALG}} \log\frac{MHSA}{\delta}$ (and otherwise *known*).

Proof of Theorem 6.3.3. Let Ω^m be the event that the good event of algorithm ALG holds up to the beginning of interval m, and define X^m to be 1 if an unknown state-action pair or the goal state were reached during interval m (and 0 otherwise). Notice that $\mathbb{E}[X^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^m] = \mathbb{E}[X^m \mid \bar{U}^m]\mathbb{I}\{\Omega^m\} \geq \mathbb{I}\{\Omega^m\}/2$ by Theorem D.1.3. Moreover, note that every state-action pair becomes known after $\omega_{\texttt{ALG}} \log\frac{MHSA}{\delta}$ visits and therefore $\sum_{m=1}^{M} X^m \mathbb{I}\{\Omega^m\} \leq \sum_{m=1}^{M} X^m \leq K + SA\omega_{\texttt{ALG}} \log\frac{MHSA}{\delta}$. By Theorem D.5.2, which is a consequence of Freedman's inequality for bounded positive random variables, we have with probability at least

$1-\frac{\delta}{8}$ for all $M \geq 1$ simultaneously

$$\sum_{m=1}^{M} \mathbb{E}[X^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^m] \leq 2\sum_{m=1}^{M} X^m \mathbb{I}\{\Omega^m\} + 108\log\frac{M}{\delta} \leq 2K + 110SA\omega_{\texttt{ALG}}\log\frac{MHSA}{\delta}.$$

Using a union bound, this inequality and the good event of ALG both hold with probability at least $1-\frac{3}{8}\delta$. Then, $\mathbb{I}\{\Omega^m\} = 1$ for all m, and therefore

$$\frac{M}{2} \leq 2K + 110SA\omega_{\texttt{ALG}}\log\frac{MHSA}{\delta}.$$

Using the fact that $x \leq a\log(bx) + c \to x \leq 6a\log(abc) + c$ for $a,b,c \geq 1$, this implies

$$M \leq 4K + 4\cdot 10^4 SA\omega_{\texttt{ALG}}\log\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta}.$$

□

D.2 Proofs for Section 6.4

Since all the proofs in this section refer to the finite-horizon setting (without a connection to SSP), we use the simpler notations $\mathscr{M} = (\mathscr{S}, \mathscr{A}, P, H, c, c_f)$ for the MDP, $V_h^\pi(s)$ for the value function of policy π, and $B_\star \geq \max_{s,h} V_h^\star(s)$ for the upper bound on the value function of the optimal policy.

We define a state-action pair (s,a) to be *known* if it was visited at least $\alpha H^4 B_\star^{-2} S$ times (for some universal constant $\alpha > 0$ to be determined later), and otherwise *unknown*. In addition, we denote by h_m the last time step before an unknown state-action pair was reached (or H if they were not reached).

D.2.1 The good event, optimism and pessimism

Throughout this section we use the notation $a \vee 1$ defined as $\max\{a, 1\}$. In addition, we define the logarithmic factor $L_m = 3\log(6SAHm/\delta)$. Define the following events:

$$E^c(m) = \left\{\forall (s,a) :\ |\bar{c}^{m-1}(s,a) - c(s,a)| \leq b_c^m(s,a)\right\}$$

$$E^{cv}(m) = \left\{\forall (s,a) :\ \left|\sqrt{\overline{\mathrm{Var}}_{s,a}^{m-1}(C)} - \sqrt{\mathrm{Var}_{s,a}(c)}\right| \leq \sqrt{\frac{12L_m}{n^{m-1}(s,a) \vee 1}}\right\}$$

$$E^p(m) = \left\{\forall (s,a,s') :\ |P(s'|s,a) - \bar{P}^{m-1}(s'|s,a)| \leq \sqrt{\frac{2P(s'|s,a)L_m}{n^{m-1}(s,a) \vee 1}} + \frac{2L_m}{n^{m-1}(s,a) \vee 1}\right\}$$

$$E^{pv1}(m) = \left\{\forall (s,a,h) :\ \left|\left(\bar{P}^{m-1}(\cdot|s,a) - P(\cdot|s,a)\right) \cdot V_{h+1}^*\right| \leq \sqrt{\frac{2\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)L_m}{n^{m-1}(s,a) \vee 1}} + \frac{5B_\star L_m}{n^{m-1}(s,a) \vee 1}\right\}$$

$$E^{pv2}(m) = \left\{\forall (s,a,h) :\ \left|\sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)} - \sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(V_{h+1}^*)}\right| \leq \sqrt{\frac{12B_\star^2 L_m}{n^{m-1}(s,a) \vee 1}}\right\}$$

For brevity, we denote $b_{pv1,h}^m(s,a) = \sqrt{\frac{2\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)L_m}{n^{m-1}(s,a)\vee 1}} + \frac{5B_\star L_m}{n^{m-1}(s,a)\vee 1}$. This good event, which is the intersection of the above events, is the one used in [EMSM21]. The following lemma establishes that the good event holds with high probability. The proof is supplied in [EMSM21, Lemma 13] by applying standard concentration results.

Lemma D.2.1 (The First Good Event)**.** *Let* $\mathbb{G}_1 = \cap_{m\geq 1} E^c(m) \cap_{m\geq 1} E^{cv}(m) \cap_{m\geq 1} E^p(m) \cap_{m\geq 1} E^{pv1}(m) \cap_{m\geq 1} E^{pv2}(m)$ *be the basic good event. It holds that* $\Pr(\mathbb{G}_1) \geq 1 - \frac{1}{4}\delta$.

Under the first good event, we can prove that the value is optimistic using standard techniques.

Lemma D.2.2 (Upper Value Function is Optimistic, Lower Value Function is Pessimistic)**.** *Conditioned on the first good event* $\mathbb{G}_1$*, it holds that* $\underline{V}_h^m(s) \le V_h^*(s) \le V_h^{\pi^m}(s) \le \bar{V}_h^m(s)$ *for every* $m = 1,2,\ldots$*,* $s \in \mathcal{S}$ *and* $h = 1,\ldots,H+1$*.*

Proof. Since $V_h^*(s) \le V_h^{\pi}(s)$ for any policy π, we only need to prove the leftmost and rightmost inequalities of the claim. We prove this result via induction.

Base case, the claim holds for $h = H+1$**.** Since we assume the terminal costs are known, for any $s \in \mathcal{S}$,

$$\underline{V}_{H+1}^m(s) = V_{H+1}^*(s) = V_{H+1}^{\pi^m}(s) = \bar{V}_{H+1}^m(s) = c_f(s).$$

Induction step, prove for $h \in [H]$ **assuming the claim holds for all** $h+1 \le h' \le H+1$**.**

Leftmost inequality, optimism. Let $a^*(s) \in \arg\min_{a\in A} Q_h^*(s,a)$, then

$$V_h^*(s) - \underline{V}_h^m(s) = Q_h^*(s,a^*(s)) - \max\left\{\min_{a\in A} \underline{Q}_h^m(s,a), 0\right\}. \tag{D.1}$$

Assume that $\min_a \bar{Q}_h^m(s,a) > 0$ (otherwise, the inequality is satisfied). Then,

$$\begin{aligned}
\text{(D.1)} &\ge Q_h^*(s,a^*(s)) - \underline{Q}_h^m(s,a^*(s)) \\
&= c(s,a^*(s)) - \bar{c}^{m-1}(s,a^*(s)) + b_c^m(s,a^*(s)) + b_p^m(s,a^*(s)) \\
&\quad + (P - \bar{P}^{m-1})(\cdot \mid s,a^*(s)) \cdot V_{h+1}^* + \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a^*(s))}\underbrace{[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')]}_{\ge 0 \text{ Induction hypothesis}} \\
&\ge -b_{pv1,h}^m(s,a^*(s)) + b_p^m(s,a^*(s)),
\end{aligned} \tag{D.2}$$

where the last relation holds since the events $\cap_m E^{pv1}(m)$ and $\cap_m E^{c}(m)$ hold. We now analyze this term.

$$
\begin{aligned}
(\text{D.2}) &= -b^m_{pv1,h}(s,a^*(s)) + b^m_p(s,a^*(s)) \\
&\overset{(a)}{\geq} -\sqrt{\frac{2\mathrm{Var}_{P(\cdot|s,a^*(s))}(V^*_{h+1})L_m}{n^{m-1}(s,a^*(s))\vee 1}} - \frac{5B_\star L_m}{n^{m-1}(s,a^*(s))\vee 1} \\
&\quad + \sqrt{\frac{2\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a^*(s))}(\underline{V}^m_{h+1})L_m}{n^{m-1}(s,a^*(s))\vee 1}} + \frac{17H^3B_\star^{-1}L_m}{n^{m-1}(s,a^*(s))\vee 1} + \frac{B_\star}{16H^2}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] \\
&\geq -\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s,a^*(s))}(V^*_{h+1})} - \sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a^*(s))}(\underline{V}^m_{h+1})}}{\sqrt{n^{m-1}(s,a^*(s))\vee 1}} \\
&\quad + \frac{B_\star}{16H^2}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] + \frac{13H^3B_\star^{-1}L_m}{n^{m-1}(s,a^*(s))\vee 1} \\
&\overset{(b)}{\geq} -\frac{B_\star}{16H^2}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] - \frac{13H^2L_m}{n^{m-1}(s,a^*(s))\vee 1} \\
&\quad + \frac{B_\star}{16H^2}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] + \frac{13H^3B_\star^{-1}L_m}{n^{m-1}(s,a)\vee 1} \geq 0,
\end{aligned}
$$

where (a) holds by plugging the definition of the bonuses $b^m_{pv1,h}$ and b^m_p (recall Equation (6.2)), as $S \geq 1$ by assumption, and by the induction hypothesis ($\bar{V}^m_{h+1}(s) \geq V^*_{h+1}(s)$). (b) holds by Lemma D.2.11 while setting $\alpha = 16H^2B_\star^{-1}$ and bounding $(5+\alpha/2)B_\star \leq 13H^2$. Combining all the above we conclude the proof of the rightmost inequality since $V^*_h(s) - \underline{V}^m_h(s) \geq (\text{D.1}) \geq (\text{D.2}) \geq 0$.

Rightmost inequality, pessimism. The following relations hold.

$$
V^{\pi^m}_h(s) - \bar{V}^m_h(s) = Q^{\pi^m}_h(s,\pi^m_h(s)) - \min\left\{\bar{Q}^m_h(s,\pi^m_h(s)), H\right\}. \tag{D.3}
$$

Assume that $\bar{Q}^m_h(s,\pi^m_h(s)) < H$ (otherwise, the claim holds). Then,

$$
\begin{aligned}
(\text{D.3}) &= Q^{\pi^m}_h(s,\pi^m_h(s)) - \bar{Q}^m_h(s,\pi^m_h(s)) \\
&= c(s,\pi^m_h(s)) - \bar{c}^{m-1}(s,\pi^m_h(s)) - b^m_c(s,\pi^m_h(s)) - b^m_p(s,\pi^m_h(s)) \\
&\quad + (P - \bar{P}^{m-1})(\cdot \mid s,\pi^m_h(s)) \cdot V^{\pi^m}_{h+1} + \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi^m_h(s))}[\underbrace{V^{\pi^m}_{h+1}(s') - \bar{V}^m_{h+1}(s')}_{\leq 0 \text{ Induction hypothesis}}] \\
&\leq -b^m_p(s,\pi^m_h(s)) + (P - \bar{P}^{m-1})(\cdot \mid s,\pi^m_h(s)) \cdot V^{\pi^m}_{h+1}.
\end{aligned} \tag{D.4}
$$

We now focus on the last term. Observe that

$$
\begin{aligned}
(P - & \bar{P}^{m-1})(\cdot \mid s, \pi_h^m(s)) \cdot V_{h+1}^{\pi^m} = \\
&= (P - \bar{P}^{m-1})(\cdot \mid s, \pi_h^m(s)) \cdot V_{h+1}^{*} + (P - \bar{P}^{m-1})(\cdot \mid s, \pi_h^m(s)) \cdot (V_{h+1}^{\pi^m} - V_{h+1}^{*}) \\
&\leq b_{pv1,h}^m(s, \pi_h^m(s)) + (P - \bar{P}^{m-1})(\cdot \mid s, \pi_h^m(s)) \cdot (V_{h+1}^{\pi^m} - V_{h+1}^{*}) \qquad (\cap_m E^{pv1}(m) \text{ holds}) \\
&\overset{(a)}{\leq} b_{pv1,h}^m(s, \pi_h^m(s)) + \frac{36 H^3 B_\star^{-1} S L_m}{n^{m-1}(s, \pi_h^m(s)) \vee 1} + \frac{B_\star}{32 H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\Big[(V_{h+1}^{\pi^m} - V_{h+1}^{*})(s')\Big] \\
&\overset{(b)}{\leq} b_{pv1,h}^m(s, \pi_h^m(s)) + \frac{36 H^3 B_\star^{-1} S L_m}{n^{m-1}(s, \pi_h^m(s)) \vee 1} + \frac{B_\star}{32 H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\big[(\bar{V}_{h+1}^{m} - \underline{V}_{h+1}^{m})(s')\big] \\
&\overset{(c)}{\leq} \sqrt{\frac{2 \mathrm{Var}_{P(\cdot|s,\pi_h^m(s))}(V_{h+1}^{*}) L_m}{n^{m-1}(s, \pi_h^m(s)) \vee 1}} + \frac{41 H^3 B_\star^{-1} S L_m}{n^{m-1}(s, \pi_h^m(s)) \vee 1} \\
&\qquad + \frac{B_\star}{32 H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\big[(\bar{V}_{t-1,h+1} - \underline{V}_{h+1}^{m})(s')\big],
\end{aligned}
$$

where (a) holds by applying Theorem D.2.13 while setting $\alpha = 32H^2B_\star^{-1}, C_1 = 2, C_2 = 2$ and bounding $2C_2 + \alpha S C_1/2 \leq 36H^2B_\star^{-1}S$ (assumption holds since $\cap_m E^p(m)$ holds), (b) holds by the induction hypothesis, and (c) holds by plugging in $b_{pv1,h}^m$. Plugging this back into (D.4) and plugging the explicit form of the bonus $b_p^m(s,a)$ we get

$$
\begin{aligned}
(\text{D.4}) \leq & -\sqrt{2L_m} \frac{\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}(\underline{V}_{h+1}^m)} - \sqrt{\mathrm{Var}_{P(\cdot|s,\pi_h^m(s))}(V_{h+1}^*)}}{\sqrt{n^{m-1}(s, \pi_h^m(s)) \vee 1}} \\
& - \frac{21 H^3 B_\star^{-1} S L_m}{n^{m-1}(s, \pi_h^m(s)) \vee 1} - \frac{B_\star}{32H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\big[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')\big] \\
\leq & \frac{B_\star}{32H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\big[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')\big] + \frac{21 H^3 B_\star^{-1} L_m}{n^{m-1}(s, \pi_h^m(s))} \\
& - \frac{B_\star}{32H^2} \mathbb{E}_{\bar{P}^{m-1}(\cdot|s,\pi_h^m(s))}\big[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')\big] - \frac{21 H^3 B_\star^{-1} S L_m}{n^{m-1}(s, \pi_h^m(s))} = 0,
\end{aligned}
$$

where the last inequality holds by Lemma D.2.11 while setting $\alpha = 32H^2B_\star^{-1}$ and bounding $(5 + \alpha/2)B_\star \leq 21H^3B_\star^{-1}$. Combining all the above we concludes the proof as

$$
V_h^{\pi^m}(s) - \bar{V}_h^m(s) \leq (\text{D.3}) \leq (\text{D.4}) \leq 0.
$$

□

Finally, using similar techniques to [EMSM21], we can prove an additional high probability bounds which hold alongside the basic good event $\mathbb{G}_1$.

Lemma D.2.3 (The Good Event)**.** *Let* $\mathbb{G}_1$ *be the event defined in Theorem D.2.1, and define the following random variables.*

$$
\begin{aligned}
Y_{1,h}^m &= \bar{V}_h^m(s_h^m) - \underline{V}_h^m(s_h^m)\\
Y_{2,h}^m &= \mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})\\
Y_3^m &= \left(\sum_{h=1}^{H} c(s_h^m,a_h^m) + c_f(s_{h+1}^m)\right)^2\\
Y_4^m &= \left(\sum_{h=1}^{h_m} c(s_h^m,a_h^m) + c_f(s_{h+1}^m)\mathbb{I}\{h_m = H\}\right)^2\\
Y_5^m &= \sum_{h=1}^{h_m} c(s_h^m,a_h^m) + c_f(s_{h+1}^m)\mathbb{I}\{h_m = H\}.
\end{aligned}
$$

The second good event is the intersection of two events $\mathbb{G}_2 = E^{OP} \cap E^{\mathrm{Var}} \cap E^{Sec1} \cap E^{Sec2} \cap E^{cost}$ *defined as follows.*

$$
\begin{aligned}
E^{OP} &= \left\{\forall h \in [H], M \geq 1: \ \sum_{m=1}^{M} \mathbb{E}[Y_{1,h}^m \mid \bar{U}_h^m] \leq 68H^2L_M + \left(1+\frac{1}{4H}\right)\sum_{m=1}^{M} Y_{1,h}^m\right\}\\
E^{\mathrm{Var}} &= \left\{\forall M \geq 1: \ \sum_{m=1}^{M}\sum_{h=1}^{H} Y_{2,h}^m \leq 16H^3L_M + 2\sum_{m=1}^{M}\sum_{h=1}^{H} \mathbb{E}[Y_{2,h}^m | \bar{U}^m]\right\}\\
E^{Sec1} &= \left\{\forall M \geq 1: \ \sum_{m=1}^{M} \mathbb{E}[Y_3^m \mid \bar{U}^m] \leq 68H^4L_M + 2\sum_{m=1}^{M} Y_3^m\right\}\\
E^{Sec2} &= \left\{\forall M \geq 1: \ \sum_{m=1}^{M} Y_4^m \leq 16H^4L_M + 2\sum_{m=1}^{M} \mathbb{E}[Y_4^m \mid \bar{U}^m]\right\}\\
E^{cost} &= \left\{\forall M \geq 1: \ \sum_{m=1}^{M} Y_5^m \leq 8HL_M + 2\sum_{m=1}^{M} \mathbb{E}[Y_5^m \mid \bar{U}^m]\right\}.
\end{aligned}
$$

Then, the good event $\mathbb{G} = \mathbb{G}_1 \cap \mathbb{G}_2$ *holds with probability at least* $1-\delta$.

Proof. **Event** E^{OP}**.** Fix h and M. We start by defining the random variable $W^m = \mathbb{I}\{\bar{V}_h^m(s) - \underline{V}_h^m(s) \geq 0 \ \forall h \in [H], s \in \mathcal{S}\}$. Observe that Y_h^m is $\bar{U}_h^m$ measurable and also notice that W^m is $\bar{U}^m$ measurable, as both π^m and $\bar{V}_h^m$ are $\bar{U}^m$-measurable. Finally, define $\tilde{Y}^m = W^m Y_h^m$. Importantly, notice that $\tilde{Y}^m \in [0,2H]$ almost surely, by definition of W^m and since $\bar{V}_h^m(s), \underline{V}_h^m(s) \in [0,2H]$ by the update rule. Thus, using Theorem D.5.2 with $C = 2H \geq 1$, we get

$$
\sum_{m=1}^{M} \mathbb{E}[\tilde{Y}_h^m \mid \bar{U}_h^m] \leq \left(1+\frac{1}{4H}\right)\sum_{m=1}^{M} \tilde{Y}_h^m + 68H^2 \log\frac{2HM(M+1)}{\delta},
$$

with probability greater than $1-\delta$, and since W^m is $\bar{U}^m$-measurable, we can write

$$\sum_{m=1}^{M} W^m \mathbb{E}[Y_h^m | \bar{U}_h^m] \le \left(1+\frac{1}{4H}\right) \sum_{m=1}^{M} W^m Y_h^m + 68H^2 \log \frac{2HM(M+1)}{\delta}. \qquad \text{(D.5)}$$

Importantly, notice that under $\mathbb{G}_1$, it holds that $W^m \equiv 1$ (by Theorem D.2.2). Therefore, applying the union bound and setting $\delta = \delta/(2HM(M+1))$ we get

$$\begin{aligned}
&\Pr(\overline{E^O} \cap \mathbb{G}_1) \le \\
&\le \sum_{h=1}^{H} \sum_{M=1}^{\infty} \Pr\left(\left\{ \sum_{m=1}^{M} \mathbb{E}[Y_h^m | \bar{U}_h^m] \ge \left(1+\frac{1}{4H}\right) \sum_{m=1}^{M} Y_h^m + 68H^2 \log \frac{2HM(M+1)}{\delta} \right\} \cap \mathbb{G}_1 \right) \\
&= \sum_{h=1}^{H} \sum_{M=1}^{\infty} \Pr\left(\left\{ \sum_{m=1}^{M} W^m \mathbb{E}[Y_h^m | \bar{U}_h^m] \ge \left(1+\frac{1}{4H}\right) \sum_{m=1}^{M} W^m Y_h^m + 68H^2 \log \frac{2HM(M+1)}{\delta} \right\} \cap \mathbb{G}_1 \right) \\
&\le \sum_{h=1}^{H} \sum_{M=1}^{\infty} \Pr\left(\sum_{m=1}^{M} W^m \mathbb{E}[Y_h^m | \bar{U}_h^m] \ge \left(1+\frac{1}{4H}\right) \sum_{m=1}^{M} W^m Y_h^m + 68H^2 \log \frac{2HM(M+1)}{\delta} \right) \\
&\le \sum_{h=1}^{H} \sum_{M=1}^{\infty} \frac{\delta}{2HM(M+1)} = \delta/2,
\end{aligned}$$

where the first relation is by a union bound, the second relation follows because $W^m \equiv 1$ under $\mathbb{G}_1$, and the last relation is by (D.5). Finally, we have

$$\Pr(\overline{\mathbb{G}}) \le \Pr(\overline{\mathbb{G}_2} \cap \mathbb{G}_1) + 2\Pr(\overline{\mathbb{G}_1}) \le \frac{\delta}{2} + \frac{2\delta}{4} = \delta.$$

Replacing $\delta \to \delta/5$ implies that $\Pr(\overline{E^{OP}} \cap \mathbb{G}_1) \le \frac{\delta}{10}$.

Event E^{Var}. Fix $h \in [H]$. Observe that $Y_{2,h}^m$ is $\bar{U}^m$ measurable and that $0 \le Y_{2,h}^m \le 4H^2$. Applying the second statement of Theorem D.5.2 we get that

$$\sum_{m=1}^{M} Y_{2,h}^m \le 2 \sum_{m=1}^{M} \mathbb{E}[Y_{2,h}^m | \bar{U}^m] + 16H^2 \log \frac{1}{\delta}.$$

By taking union bound, as in the proof of the first statement of the lemma on all $h \in [H]$ and summing over $h \in [H]$, we get that with probability at least $1-\delta/10$ for all $M \ge 1$ it holds that

$$\sum_{m=1}^{M} \sum_{h=1}^{H} Y_{2,h}^m \le 2 \sum_{m=1}^{M} \sum_{h=1}^{H} \mathbb{E}[Y_{2,h}^m | \bar{U}^m] + 16H^3 L_M.$$

Event E^{Sec1}. Observe that Y_3^m is $\bar{U}^m$ measurable and that $0 \le Y_3^m \le 4H^2$. Applying the first statement of Theorem D.5.2 we get that

$$\sum_{m=1}^{M} \mathbb{E}[Y_3^m | \bar{U}^m] \leq 2 \sum_{m=1}^{M} Y_3^m + 50H^4 \log \frac{1}{\delta}.$$

By taking union bound we get that with probability at least $1 - \delta/10$ the event holds.

Event E^{Sec2}. Observe that Y_4^m is $\bar{U}^m$ measurable and that $0 \leq Y_4^m \leq 4H^2$. Applying the second statement of Theorem D.5.2 we get that

$$\sum_{m=1}^{M} Y_4^m \leq 2 \sum_{m=1}^{M} \mathbb{E}[Y_4^m | \bar{U}^m] + 16H^2 \log \frac{1}{\delta}.$$

By taking union bound we get that with probability at least $1 - \delta/10$ the event holds.

Event E^{cost}. Observe that Y_5^m is $\bar{U}^m$ measurable and that $0 \leq Y_5^m \leq 2H$. Applying the second statement of Theorem D.5.2 we get that

$$\sum_{m=1}^{M} Y_5^m \leq 2 \sum_{m=1}^{M} \mathbb{E}[Y_5^m | \bar{U}^m] + 8H \log \frac{1}{\delta}.$$

By taking union bound we get that with probability at least $1 - \delta/10$ the event holds.

Combining all the above. We bound the probability of $\overline{G}$ as follows:

$$\Pr(\overline{\mathbb{G}}) \leq \Pr(\overline{\mathbb{G}_1}) + \Pr(\overline{E^{OP}} \cap \mathbb{G}_1) + \Pr(\overline{E^{\mathrm{Var}}}) + \Pr(\overline{E^{Sec1}}) + \Pr(\overline{E^{Sec2}}) + \Pr(\overline{E^{cost}}) \leq \frac{\delta}{2} + 5 \cdot \frac{\delta}{10} = \delta.$$

□

D.2.2 ULCVI is admissible

By the definition of the algorithm and its regret bound in Theorem 6.4.1, it is clear that properties 1,2,3 of the admissible algorithm definition hold. Thus, it remains to show property 4 by bounding ω_{ULCVI}. In order to show that $\omega_{\mathrm{ULCVI}} = O(H^4 B_\star^{-2} S)$, we need to show that if the number of visits to (s,a) is at least $\alpha H^4 B_\star^{-2} S \log \frac{MHSA}{\delta}$ (for a large enough universal constant $\alpha > 0$) then $\|P(\cdot \mid s,a) - \widetilde{P}_t(\cdot \mid s,a)\|_1 \leq 1/(18H)$ and $|c(s,a) - \tilde{c}_h^t(s,a)| \leq B_\star/H$ (under the good event), where $\widetilde{P}, \tilde{c}$ are the estimations used by the algorithm to compute its optimistic Q-function (i.e., these are the empirical transition estimate and the empirical cost estimate plus the bonus).

Indeed, by event $\cap_{m>0} E^p(m)$,

$$\begin{aligned}
\|P(\cdot \mid s,a) - \widetilde{P}(\cdot \mid s,a)\|_1 &= \|P(\cdot \mid s,a) - \bar{P}(\cdot \mid s,a)\|_1 \\
&\leq \sqrt{\frac{2S \log \frac{16M^3HS^2A}{\delta}}{n(s,a)}} + \frac{2S \log \frac{16M^3HS^2A}{\delta}}{n(s,a)} \\
&\leq \frac{4B_\star}{\sqrt{\alpha}H^2} + \frac{16B_\star^2}{\alpha H^4} \leq \frac{1}{18H},
\end{aligned}$$

for $\alpha > 5800$, where the first inequality holds by Jensen inequality and since event $\cap_{m>0} E^p(m)$ holds. By the definition of the exploration bonuses we have

$$\begin{aligned}
&|c(s,a) - \tilde{c}_h(s,a)| \leq |c(s,a) - \bar{c}(s,a)| + b_c(s,a) + b_p(s,a) \\
&\quad \leq 3\sqrt{\frac{2B_\star^2 \log \frac{16M^3HS^2A}{\delta}}{n(s,a)}} + \frac{72H^3B_\star^{-1}S \log \frac{16M^3HS^2A}{\delta}}{n(s,a)} + \frac{B_\star \max_{s'} \bar{V}_{h+1}(s') - \underline{V}_{h+1}(s')}{16H^2} \\
&\quad \leq \frac{12B_\star^2}{\sqrt{\alpha}H^2} + \frac{800B_\star}{\alpha H} + \frac{B_\star}{16H} \leq \frac{B_\star}{H},
\end{aligned}$$

for $\alpha > 5800$.

Finally, note that although our algorithm does not update the policy in the beginning of every episode (only when the number of visits to some state-action pair is doubled), this only implies that the constant α needs to be doubled.

D.2.3 Proof of Theorem 6.4.1

As in the proof of UCBVI, before establishing the proof of Theorem 6.4.1 we establish the following key lemma that bounds the on-policy errors at time step h by the on-policy errors at time step $h+1$ and additional additive terms. Given this result, the analysis follows with relative ease.

Lemma D.2.4 (ULCBVI, Key Recursion Bound)**.** *Conditioning on the good event $\mathbb{G}$, the following bound holds for all $h \in [H]$.*

$$\begin{aligned}
\sum_{m=1}^M \bar{V}_h^m(s_h^m) - \underline{V}_h^m(s_h^m) &\leq 68H^2 L_M + \sum_{m=1}^M \frac{310H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m) \vee 1} + \sum_{m=1}^M 4\sqrt{L_m} \frac{\sqrt{c(s_h^m,a_h^m)}}{\sqrt{n^{m-1}(s_h^m,a_h^m) \vee 1}} \\
&+ \sum_{m=1}^M 2\sqrt{2L_m} \frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m) \vee 1}} + \left(1 + \frac{1}{2H}\right)^2 \sum_{m=1}^M \left(\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)\right).
\end{aligned}$$

Proof. We bound each of the terms in the sum as follows.

$$\begin{aligned}
\bar{V}_h^m(s_h^m) - \underline{V}_h^m(s_h^m) &= 2b_c^m(s_h^m,a_h^m) + 2b_p^m(s_h^m,a_h^m) + \mathbb{E}_{\bar{P}^{m-1}(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)] \\
&= 2b_c^m(s_h^m,a_h^m) + 2b_p^m(s_h^m,a_h^m) \\
&\quad + \mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)] + (\bar{P}^{m-1} - P)(\cdot|s_h^m,a_h^m) \cdot \left(\bar{V}_{h+1}^m - \underline{V}_{h+1}^m\right) \\
&\le 2b_c^m(s_h^m,a_h^m) + 2b_p^m(s_h^m,a_h^m) \\
&\quad + \frac{8H^2SL_m}{n^{m-1}(s_h^m,a_h^m) \vee 1} + \left(1 + \frac{1}{4H}\right)\mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)],
\end{aligned} \tag{D.6}$$

where the last relation holds by Theorem D.2.13 which upper bounds

$$(\bar{P}^{m-1} - P)(\cdot|s_h^m,a_h^m) \cdot \left(\bar{V}_{h+1}^m - \underline{V}_{h+1}^m\right) \le \frac{8H^2SL_m}{n^{m-1}(s_h^m,a_h^m) \vee 1} + \frac{1}{4H}\mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)]$$

by setting $\alpha = 4H, C_1 = C_2 = 2$ and bounding $HL_m(2C_2 + \alpha SC_1/2) \le 8H^2SL_m$ (the assumption of the lemma holds since the event $\cap_m E^p(m)$ holds). Taking the sum over $m \in [M]$ we get that

$$\begin{aligned}
\sum_{m=1}^M \bar{V}_h^m(s_h^m) - \underline{V}_h^m(s_h^m) &\le \sum_{m=1}^M 2b_c^m(s_h^m,a_h^m) + \sum_{m=1}^M 2b_p^m(s_h^m,a_h^m) \\
&\quad + \sum_{m=1}^M \frac{8H^2SL_m}{n^{m-1}(s_h^m,a_h^m) \vee 1} + \sum_{m=1}^M \left(1 + \frac{1}{4H}\right)\mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)].
\end{aligned} \tag{D.7}$$

The first sum is bounded in Theorem D.2.5 by

$$\sum_{m=1}^M b_c^m(s_h^m,a_h^m) \le \sum_{m=1}^M \sqrt{\frac{2c(s_h^m,a_h^m)L_m}{n^{m-1}(s_h^m,a_h^m) \vee 1}} + \sum_{m=1}^M \frac{10L_m}{n^{m-1}(s_h^m,a_h^m) \vee 1},$$

and the second sum is bounded in Theorem D.2.6 by

$$\begin{aligned}
\sum_{m=1}^M b_p^m(s_h^m,a_h^m) &\le \sum_{m=1}^M \frac{139H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m) \vee 1} + \sum_{m=1}^M \sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m) \vee 1}} \\
&\quad + \frac{1}{8H}\sum_{m=1}^M \mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)].
\end{aligned}$$

Plugging this into (D.7) and rearranging the terms we get

$$
\begin{aligned}
\sum_{m=1}^{M} \bar{V}_h^m(s_h^m) - \underline{V}_h^m(s_h^m) &\le \sum_{m=1}^{M} \frac{2\sqrt{2c(s_h^m,a_h^m)L_m}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} + \sum_{m=1}^{M} 2\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} \\
&\quad + \sum_{m=1}^{M} \frac{286H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m)\vee 1} + \left(1+\frac{1}{2H}\right)\sum_{m=1}^{M} \mathbb{E}_{P(\cdot|s_h^m,a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)] \\
&\le 68H^2L_M + \sum_{m=1}^{M} \frac{2\sqrt{2L_m}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} + \sum_{m=1}^{M} \frac{286H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m)\vee 1} \\
&\quad + \sum_{m=1}^{M} 2\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} + \left(1+\frac{1}{2H}\right)^2 \sum_{m=1}^{M} \bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m),
\end{aligned}
$$

where the last inequality follows since the second good event holds. □

Proof of Theorem 6.4.1. Start by conditioning on the good event which holds with probability greater than $1-\delta$. Applying the optimism-pessimism of the upper and lower value function we get

$$
\sum_{m=1}^{M} V_1^{\pi^m}(s_1^m) - V_1^*(s_1^m) \le \sum_{m=1}^{M} \bar{V}_1^m(s_1^m) - \underline{V}_1^m(s_1^m). \tag{D.8}
$$

Iteratively applying Theorem D.2.4 and bounding the exponential growth by $\left(1+\frac{1}{2H}\right)^{2H} \le e \le 3$, the following upper bound on the cumulative regret is obtained.

$$
\begin{aligned}
\text{(D.8)} &\le 204H^3B_\star^{-1}L_M + \sum_{m=1}^{M}\sum_{h=1}^{H} \frac{930H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m)\vee 1} \\
&\quad + \sum_{m=1}^{M}\sum_{h=1}^{H} \frac{12\sqrt{c(s_h^m,a_h^m)L_m}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} + 9\sum_{m=1}^{M}\sum_{h=1}^{H} \frac{\sqrt{L_m\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)}}.
\end{aligned} \tag{D.9}
$$

We now bound each of the three sums in Equation (D.9). We bound the first sum in

Equation (D.9) via standard analysis as follows:

$$\begin{aligned}
\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m,a_h^m)\vee 1} &\le H^3B_\star^{-1}SL_M\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{1}{n^{m-1}(s_h^m,a_h^m)\vee 1}\\
&= H^3B_\star^{-1}SL_M\sum_{m=1}^{M}\sum_{s,a}\frac{\sum_{h=1}^{H}\mathbb{I}\{s_h^m=s,a_h^m=a\}}{n^{m-1}(s,a)\vee 1}\\
&\le H^3B_\star^{-1}SL_M\sum_{m=1}^{M}\sum_{s,a}\mathbb{I}\{n^{m-1}(s,a)\ge H\}\frac{\sum_{h=1}^{H}\mathbb{I}\{s_h^m=s,a_h^m=a\}}{n^{m-1}(s,a)\vee 1}+2H^4B_\star^{-1}S^2AL_M\\
&\le 3H^3B_\star^{-1}S^2AL_M\log(MH)+2H^4B_\star^{-1}S^2AL_M,
\end{aligned}$$

where the last inequality is by Theorem D.2.12 that bounds $\sum_{m,s,a}\mathbb{I}\{n^{m-1}(s,a)\ge H\}\frac{\sum_{h=1}^{H}\mathbb{I}\{s_h^m=s,a_h^m=a\}}{n^{m-1}(s,a)\vee 1}\le 3SA\log(MH)$.

The second sum in Equation (D.9) is bounded as follows.

$$\begin{aligned}
\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{\sqrt{c(s_h^m,a_h^m)L_m}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} &\le \sum_{m=1}^{M}\sum_{h=1}^{H}\frac{\sqrt{c(s_h^m,a_h^m)L_m}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}}\mathbb{I}\{n^{m-1}(s_h^m,a_h^m)\ge H\}+2HSAL_M\\
&\overset{(a)}{\le} \sqrt{L_M}\sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}c(s_h^m,a_h^m)}\cdot\sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{\mathbb{I}\{n^{m-1}(s_h^m,a_h^m)\ge H\}}{n^{m-1}(s_h^m,a_h^m)\vee 1}}+2HSAL_M\\
&\overset{(b)}{\le} \sqrt{L_M}\sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}c(s_h^m,a_h^m)}\cdot\sqrt{3SA\log(MH)}+2HSAL_M\\
&\le \sqrt{3SA}L_M\sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}c(s_h^m,a_h^m)+c_f(s_{H+1}^m)}+2HSAL_M\\
&\le O\Big(\sqrt{B_\star SAM}L_M+H^3B_\star^{-1}S^2A\log^{3/2}\frac{MHSA}{\delta}\Big).
\end{aligned}$$

where (a) is by Cauchy-Schwartz, (b) is by Theorem D.2.12, and the last inequality is by Theorem D.2.7. The third sum in Equation (D.9) is bounded in Theorem D.2.8 by

$$\begin{aligned}
\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{\sqrt{L_m\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)}} &\le \sqrt{L_M}\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)}}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(L_m\text{ increasing in }m)\\
&\le \sqrt{L_m}\cdot O\Big(\sqrt{B_\star^2SAM\log(MH)}+H^3B_\star^{-1}S^2A\log\frac{MHSA}{\delta}\Big).\quad\text{(Lemma D.2.8)}
\end{aligned}$$

□

D.2.4 Bounds on the cumulative bonuses

Lemma D.2.5 (Bound on the Cumulative Cost Function Bonus)**.** *Conditioning on the good event the following bound holds for all $h \in [H]$.*

$$\sum_{m=1}^{M} b_c^m(s_h^m, a_h^m) \le \sum_{m=1}^{M} \sqrt{\frac{2c(s_h^m, a_h^m)L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1}} + \sum_{m=1}^{M} \frac{10L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1}.$$

Proof. By definition of b_c^m and since the event $\cap_m E^{cv}(m)$ holds, we have

$$\begin{aligned}
\sum_{m=1}^{M} b_c^m(s_h^m, a_h^m) &= \sum_{m=1}^{M} \sqrt{\frac{2\overline{\mathrm{Var}}^{m-1}_{s_h^m, a_h^m}(c)L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1}} + \frac{5L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1} \\
&\le \sum_{m=1}^{M} \sqrt{\frac{2\mathrm{Var}_{s_h^m, a_h^m}(c)L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1}} + \sqrt{\frac{2L_m \left|\mathrm{Var}_{s_h^m, a_h^m}(c) - \overline{\mathrm{Var}}^{m-1}_{s_h^m, a_h^m, t-1}(c)\right|}{n^{m-1}(s_h^m, a_h^m) \vee 1}} + \frac{5L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1} \\
&\le \sum_{m=1}^{M} \sqrt{\frac{2\mathrm{Var}_{s_h^m, a_h^m}(c)L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1}} + \frac{10L_m}{n^{m-1}(s_h^m, a_h^m) \vee 1},
\end{aligned}$$

where the first inequality holds since $\sqrt{a+b} \le \sqrt{A} + \sqrt{|b|}$. Finally, notice that for every $(s,a) \in \mathscr{S} \times \mathscr{A}$ the variance of the cost is bounded by the second moment, which is bounded by the expected value $c(s,a)$ since the random cost value is bounded in $[0,1]$. □

Lemma D.2.6 (Bound on the Cumulative Transition Model Bonus)**.** *Conditioning on the good event the following bound holds for all $h \in [H]$.*

$$\begin{aligned}
\sum_{m=1}^{M} b_p^m(s_h^m, a_h^m) \le \sum_{m=1}^{M} \frac{139H^3B_\star^{-1}SL_m}{n^{m-1}(s_h^m, a_h^m) \vee 1} + \sum_{m=1}^{M} \sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m, a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m, a_h^m) \vee 1}} \\
+ \frac{1}{8H}\sum_{m=1}^{M} \mathbb{E}_{P(\cdot|s_h^m, a_h^m)}[\bar{V}_{h+1}^m(s_{h+1}^m) - \underline{V}_{h+1}^m(s_{h+1}^m)].
\end{aligned}$$

Proof. First, by applying Lemma D.2.13 with $\alpha = 8H, C_1 = C_2 = 2$ and $HL_m(2C_2 + \alpha SC_1/2) \le 12H^2SL_m$, we have

$$\begin{aligned}
\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')] &= \mathbb{E}_{P(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')] + (\bar{P}^{m-1} - P)(\cdot \mid s,a) \cdot (\bar{V}_{h+1}^m - \underline{V}_{h+1}^m) \\
&\le \frac{9}{8}\mathbb{E}_{P(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')] + \frac{12H^2SL_m}{n^{m-1}(s,a) \vee 1}.
\end{aligned} \tag{D.10}$$

Thus, the bonus $b_t^p(s,a)$ can be upper bounded as follows.

$$
\begin{aligned}
b_p^m(s,a) &\le \sqrt{2}\sqrt{\frac{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)L_m}{n^{m-1}(s,a)\vee 1}} + \frac{1}{16H}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')] + \frac{62H^3B_\star^{-1}SL_m}{n^{m-1}(s,a)\vee 1} \\
&\le \sqrt{2}\sqrt{\frac{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)L_m}{n^{m-1}(s,a)\vee 1}} + \frac{9}{128H}\mathbb{E}_{P(\cdot|s,a)}[\bar{V}_{h+1}^m(s') - \underline{V}_{h+1}^m(s')] + \frac{74H^3B_\star^{-1}SL_m}{n^{m-1}(s,a)\vee 1}.
\end{aligned} \tag{D.11}
$$

We bound the first term of (D.11) to establish the lemma. It holds that

$$
\begin{aligned}
\sqrt{2L_m}\sqrt{\frac{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)}{n^{m-1}(s,a)\vee 1}} &= \underbrace{\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)}}{\sqrt{n^{m-1}(s,a)\vee 1}}}_{(i)} \\
&\quad + \underbrace{\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s,a)\vee 1}}}_{(ii)} \\
&\quad + \frac{\sqrt{2L_m}\sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s,a)\vee 1}}.
\end{aligned}
$$

Term (i) is bounded by Theorem D.2.11 (by setting $\alpha = 32H$ and $(5+\alpha/2)B_\star \le 21H^2$),

$$
\sqrt{2L_m}\frac{\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}_{h+1}^m)} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V_{h+1}^*)}}{\sqrt{n^{m-1}(s,a)\vee 1}} \le \frac{1}{32H}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')\right] + \frac{21H^2L_m}{n^{m-1}(s,a)}
$$

Following the same steps as in (D.10), we get

$$
\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')\right] \le \frac{9}{8}\mathbb{E}_{P(\cdot|s,a)}\left[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')\right] + \frac{12H^2SL_m}{n^{m-1}(s,a)\vee 1},
$$

and thus,

$$
(i) \le \frac{9}{256H}\mathbb{E}_{P(\cdot|s,a)}\left[V_{h+1}^*(s') - \underline{V}_{h+1}^m(s')\right] + \frac{33H^2SL_m}{n^{m-1}(s,a)\vee 1}.
$$

Term (ii) is bounded as follows.

$$
\begin{aligned}
(ii) &\le \frac{\sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V^*_{h+1} - V^{\pi^m}_{h+1})}}{\sqrt{n^{m-1}(s,a) \vee 1}} && \text{(By Lemma D.5.3)} \\
&\le \frac{\sqrt{\mathbb{E}_{P(\cdot|s,a)}[(V^*_{h+1}(s') - V^{\pi^m}_{h+1}(s'))^2]}}{\sqrt{n^{m-1}(s,a) \vee 1}} \\
&\le \frac{\sqrt{2H\mathbb{E}_{P(\cdot|s,a)}[(V^*_{h+1}(s') - V^{\pi^m}_{h+1}(s'))]}}{\sqrt{n^{m-1}(s,a) \vee 1}} && (0 \le V^*_h(s') - V^{\pi^m}_h(s') \le 2H) \\
&\le \frac{1}{64H}\mathbb{E}_{P(\cdot|s,a)}[(V^{\pi^m}_{h+1}(s') - V^*_{h+1}(s'))] + \frac{32H^2}{n^{m-1}(s,a) \vee 1}. \\
& && (ab \le \tfrac{1}{\alpha}a^2 + \tfrac{\alpha}{4}b^2 \text{ for } \alpha = 64H)
\end{aligned}
$$

Thus, applying $\bar{V}^m_h \ge V^{\pi^m}_h \ge V^*_h \ge \underline{V}^m_h$ (Lemma D.2.2) in the bounds of (i) and (ii) we get

$$
b^m_p(s,a) \le \frac{1}{8H}\mathbb{E}_{P(\cdot|s,a)}[(\bar{V}^m_h(s') - \underline{V}^m_h(s'))] + \frac{139H^3B_\star^{-1}SL_m}{n^{m-1}(s,a) \vee 1} + \frac{\sqrt{2L_m}\sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V^{\pi^m}_{h+1})}}{\sqrt{n^{m-1}(s,a) \vee 1}},
$$

and summing over m concludes the proof. □

Lemma D.2.7 (Bound on Cost Term)**.** *Conditioning on the good event, it holds that*

$$
\sum_{m=1}^M \sum_{h=1}^H c(s^m_h, a^m_h) + c_f(s^m_{H+1}) \le O\Big(B_\star M + H^5 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta}\Big).
$$

Proof. Denote by h_m the last time step before reaching an unknown state-action pair (or H if it was not reached). By the event E^{cost} we have

$$
\begin{aligned}
\sum_{m=1}^M \sum_{h=1}^H c(s^m_h, a^m_h) + c_f(s^m_{H+1}) &= \sum_{m=1}^M \left(\sum_{h=h_m+1}^H c(s^m_h, a^m_h) + c_f(s^m_{h+1})\mathbb{I}\{h_m \ne H\} \right) \\
&\quad + \sum_{m=1}^M \left(\sum_{h=1}^{h_m} c(s^m_h, a^m_h) + c_f(s^m_{h+1})\mathbb{I}\{h_m = H\} \right) \\
&\le 2\alpha H^5 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + \sum_{m=1}^M \left(\sum_{h=1}^{h_m} c(s^m_h, a^m_h) + c_f(s^m_{h+1})\mathbb{I}\{h_m = H\} \right) \\
&\le 10\alpha H^5 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + 2\sum_{m=1}^M \mathbb{E}\left[\sum_{h=1}^{h_m} c(s^m_h, a^m_h) + c_f(s^m_{h+1})\mathbb{I}\{h_m = H\} \;\middle|\; \bar{U}^m \right] \\
&\le O\Big(H^5 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + B_\star M\Big),
\end{aligned}
$$

where the second inequality follows since every state-action pair becomes known after the number of visits is $\alpha H^4 B_\star^{-2} S \log \frac{MHSA}{\delta}$, and the last one by Theorem D.2.10. □

Lemma D.2.8 (Bound on Variance Term)**.** *Conditioning on the good event, it holds that*

$$\sum_{m=1}^{M}\sum_{h=1}^{H} \frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)}} \le O\Big(\sqrt{B_\star^2 SAM\log(MH)} + H^3 B_\star^{-1} S^{3/2} A \log\frac{MHSA}{\delta}\Big).$$

Proof. Applying Cauchy-Schwartz inequality we get

$$\begin{aligned}
\sum_{m=1}^{M}\sum_{h=1}^{H} \frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}} &\le \sum_{m=1}^{M}\sum_{h=1}^{H} \frac{\sqrt{\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}}{\sqrt{n^{m-1}(s_h^m,a_h^m)\vee 1}}\mathbb{I}\{n^{m-1}(s_h^m,a_h^m)\ge H\} + 2H^2SA \\
&\le \sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}\sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}\frac{1}{n^{m-1}(s_h^m,a_h^m)\vee 1}\mathbb{I}\{n^{m-1}(s_h^m,a_h^m)\ge H\}} + 2H^2SA \\
&\le \sqrt{\sum_{m=1}^{M}\sum_{h=1}^{H}\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m})}\sqrt{3SA\log(MH)} + 2H^2SA \qquad \text{(Lemma D.2.12)} \\
&\le \sqrt{2\sum_{m=1}^{M}\mathbb{E}\left[\sum_{h=1}^{H}\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m}) \mid \bar{U}^m\right] + 16H^3L_M}\sqrt{3SA\log(MH)} + 2H^2SA \\
&\qquad\qquad \text{(Event } E^{\mathrm{Var}} \text{ holds)} \\
&\le 3\sqrt{\sum_{m=1}^{M}\mathbb{E}\left[\sum_{h=1}^{H}\mathrm{Var}_{P(\cdot|s_h^m,a_h^m)}(V_{h+1}^{\pi^m}) \mid \bar{U}^m\right]}\sqrt{SA\log(MH)} \\
&\quad + 7\sqrt{SAH^3\log(MH)L_M} + 2H^2SA \qquad (\sqrt{a+b}\le\sqrt{a}+\sqrt{b}) \\
&\overset{(a)}{=} 3\sqrt{\sum_{m=1}^{M}\mathbb{E}\left[\left(\sum_{h=1}^{H}c(s_h^m,a_h^m)+c_f(s_{h+1}^m)-V_1^{\pi^m}(s_1)\right)^2 \;\middle|\; \bar{U}^m\right]}\sqrt{SA\log(MH)} \\
&\quad + 7\sqrt{SAH^3\log(MH))L_m} + 2H^2SA \\
&\overset{(b)}{\le} 3\sqrt{\sum_{m=1}^{M}\mathbb{E}\left[\left(\sum_{h=1}^{H}c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\right)^2 \;\middle|\; \bar{U}^m\right]}\sqrt{SA\log(MH)} + 9H^2SAL_M \\
&\le O\Big(\sqrt{B_\star^2 SAM\log(MH)} + H^3B_\star^{-1}S^{3/2}A\log\frac{MHSA}{\delta}\Big),
\end{aligned}$$

where (a) is by law of total variance [AOM17], see Lemma D.2.14, (b) is because the variance is bounded by the second moment, and the last inequality is by Theorem D.2.9.

□

D.2.5 Bounds on the second moment

Lemma D.2.9. *Conditioning on the good event, it holds that*

$$\sum_{m=1}^{M} \mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\right)^2 \;\middle|\; \bar{U}^m\right] \le O\Big(B_\star^2 M + H^6 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta}\Big).$$

Proof. Denote by h_m the last time step before reaching an unknown state-action pair (or H if it was not reached). By the event E^{Sec1} we have

$$\begin{aligned}
\sum_{m=1}^{M} \mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\right)^2 \;\middle|\; \bar{U}^m\right] &\le 2\sum_{m=1}^{M}\left(\sum_{h=1}^{H} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\right)^2 + 62H^4 L_M \\
&\le 4\sum_{m=1}^{M}\left(\sum_{h=h_m+1}^{H} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\mathbb{I}\{h_m\neq H\}\right)^2 + 62H^4 L_M \\
&\quad + 4\sum_{m=1}^{M}\left(\sum_{h=1}^{h_m} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\mathbb{I}\{h_m = H\}\right)^2 \\
&\le 300\alpha H^6 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + 4\sum_{m=1}^{M}\left(\sum_{h=1}^{h_m} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\mathbb{I}\{h_m = H\}\right)^2 \\
&\le 400\alpha H^6 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + 4\sum_{m=1}^{M}\mathbb{E}\left[\left(\sum_{h=1}^{h_m} c(s_h^m,a_h^m)+c_f(s_{h+1}^m)\mathbb{I}\{h_m = H\}\right)^2 \;\middle|\; \bar{U}^m\right] \\
&\le O\Big(H^6 B_\star^{-2} S^2 A \log\frac{MHSA}{\delta} + B_\star^2 M\Big),
\end{aligned}$$

where the third inequality follows since every state-action pair becomes known after the number of visits is $\alpha H^4 B_\star^{-2} S \log\frac{MHSA}{\delta}$, the forth inequality by event E^{Sec2}, and the last one by Theorem D.2.10. □

Lemma D.2.10. *Let m be an episode and h_m be the last time step before an unknown state-action pair was reached (or H if they were not reached). Further, denote by $C^m = \sum_{h=1}^{h_m} c(s_h^m,a_h^m)+c_f(s_{H+1}^m)\mathbb{I}\{h_m = H\}$ the cumulative cost in the episode until time h_m. Then, under the good event, $\mathbb{E}[C^m \mid \bar{U}^m] \le 3B_\star$ and $\mathbb{E}[(C^m)^2 \mid \bar{U}^m] \le 2\cdot 10^4 B_\star^2$.*

Proof. Consider the following finite-horizon MDP $\mathscr{M}^m = (S\cup\{g\},A,P^m,H,c^m,c_f^m)$ that contracts unknown state-action pairs with a new goal state, i.e., $c^m(s,a) = c(s,a)\mathbb{I}\{s\neq g\}$

and $c_f^m(s) = c_f(s)\mathbb{I}\{s \neq g\}$ and

$$P_h^m(s' \mid s,a) = \begin{cases} 0, & (s', \pi_{h+1}^m(s')) \text{ is unknown;} \\ P(s' \mid s,a), & s' \neq g \text{ and } (s', \pi_{h+1}^m(s')) \text{ is known;} \\ 1 - \sum_{s'' \in \mathcal{S}} P_h^m(s'' \mid s,a), & s' = g. \end{cases}$$

Denote by V^m the cost-to-go function of π^m in the MDP $\mathcal{M}^m$. Moreover, we slightly abuse notation to let $\widetilde{P}^m$ be the transition function induced by $\bar{P}^{m-1}$ in the MDP $\mathcal{M}^m$ similarly to P^m, and $\widetilde{V}^m$ the cost-to-go function of π^m with respect to $\bar{P}^{m-1}$ (and cost function $\tilde{c}^m = \tilde{c}^{m-1} - b_c^m - b_p^m$). By the value difference lemma (see, e.g., [SERM20]), for every s,h such that $(s, \pi_h^m(s))$ is known,

$$\begin{aligned} V_h^m(s) &= \widetilde{V}_h^m(s) + \sum_{h'=h}^{H} \mathbb{E}\Big[c^m(s_{h'}, a_{h'}) - \tilde{c}_{h'}^m(s_{h'}, a_{h'}) \mid s_h = s, P^m, \pi^m\Big] \\ &\quad + \sum_{h'=h}^{H} \mathbb{E}\Big[\big(P_{h'}^m(\cdot \mid s_{h'}, a_{h'}) - \widetilde{P}_{h'}^m(\cdot \mid s_{h'}, a_{h'})\big) \cdot \widetilde{V}^m \mid s_h = s, P^m, \pi^m\Big] \\ &\le \widetilde{V}_h^m(s) + H \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} |c(s, \pi_{h'}^m(s)) - \tilde{c}_{h'}^m(s, \pi_{h'}^m(s))| + H \|\widetilde{V}^m\|_\infty \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} \|P_{h'}^m(\cdot|s, \pi_{h'}^m(s)) - \widetilde{P}_{h'}^m(\cdot|s, \pi_{h'}^m(s))\|_1 \\ &\le \widetilde{V}_h^m(s) + H \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} |c(s, \pi_{h'}^m(s)) - \tilde{c}_{h'}^m(s, \pi_{h'}^m(s))| \\ &\quad + 2H \|\widetilde{V}^m\|_\infty \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} \|P(\cdot|s, \pi_{h'}^m(s)) - \bar{P}^{m-1}(\cdot|s, \pi_{h'}^m(s))\|_1 \\ &\le V_h^*(s) + H \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} |c(s, \pi_{h'}^m(s)) - \tilde{c}_{h'}^m(s, \pi_{h'}^m(s))| + 2HB_\star \max_{\substack{(s,\pi_{h'}^m(s)) \\ \text{known}}} \|P(\cdot|s, \pi_{h'}^m(s)) - \bar{P}^{m-1}(\cdot|s, \pi_{h'}^m(s))\|_1, \end{aligned}$$

where the last inequality follows by optimism and since $V_h^\star(s) \le B_\star$. Thus, by Section D.2.2 (since all state-action pairs in $\mathcal{M}^m$ are known), we have that $V_h^m(s) \le V_h^*(s) + 2B_\star \le 3B_\star$. Notice that C^m is exactly the cost in the MDP $\mathcal{M}^m$, so $\mathbb{E}[C^m \mid \bar{U}^m] \le 3B_\star$.

Similarly, we notice that $\mathbb{E}[(C^m)^2 \mid \bar{U}^m] = \mathbb{E}[(\widehat{C})^2]$, where $\widehat{C}$ is the cumulative cost in $\mathcal{M}^m$, and we override notation by denoting $\widehat{C} = \sum_{h=1}^{H} c(s_h, a_h) + c_f(s_{H+1})$. We split the time steps into Q blocks as follows. We denote by t_1 the first time step in which we accumulated a total cost of at least $3B_\star$ (or $H+1$ if it did not occur), by t_2 the first time step in which we accumulated a total cost of at least $3B_\star$ after t_1, and so on up until $t_Q = H+1$. Then, the first block consists of time steps $t_0 = 1, \dots, t_1 - 1$, the second block consists of time steps $t_1, \dots, t_2 - 1$, and so on. Since $V_h^m(s) \le 3B_\star$ we must have $c(s_h, a_h) \le 3B_\star$ for all $h = 1, \dots, H$ and thus in every such block the total cost is between $3B_\star$ and $6B_\star$. Thus,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1})\right)^2\right] &\geq \mathbb{E}\left[\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1})\right]^2 \\ &= \mathbb{E}\left[\sum_{i=0}^{Q-1}\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) + c_f(s_{H+1})\right]^2 \\ &\geq \mathbb{E}[3B_\star Q]^2 = 9B_\star^2\mathbb{E}[Q]^2,\end{aligned}$$

by Jensen's inequality. On the other hand,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1})\right)^2\right] &= \mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1}) - V_1^m(s_1) + V_1^m(s_1)\right)^2\right] \\ &\leq 2\mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1}) - V_1^m(s_1)\right)^2\right] + 2V_1^m(s_1)^2 \\ &\leq 2\mathbb{E}\left[\left(\sum_{i=0}^{Q-1}\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) - V_{t_i}^m(s_{t_i}) + V_{t_{i+1}}^m(s_{t_{i+1}})\right)^2\right] + 18B_\star^2 \\ &\overset{(a)}{=} 4\mathbb{E}\left[\sum_{i=0}^{Q-1}\left(\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) - V_{t_i}^m(s_{t_i}) + V_{t_{i+1}}^m(s_{t_{i+1}})\right)^2\right] + 18B_\star^2 \\ &\leq 4\mathbb{E}[Q\cdot(9B_\star)^2] + 18B_\star^2 \leq 324B_\star^2\mathbb{E}[Q] + 18B_\star^2.\end{aligned}$$

For (a) we used the fact that $\mathbb{E}[\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) - V_{t_i}(s_{t_i}) + V_{t_{i+1}}(s_{t_{i+1}})] = 0$ using the Bellman optimality equations and conditioned on all past randomness up until time t_i, and the fact that t_{i+1} is a stopping time, in the following manner,

$$\begin{aligned}\mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) - V_{t_i}^m(s_{t_i}) + V_{t_{i+1}}^m(s_{t_{i+1}})\right] &= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) - V_h^m(s_h) + V_{h+1}^m(s_{h+1})\right] \\ &= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1} \mathbb{E}\Big[c(s_h,a_h) - V_h^m(s_h) + V_{h+1}^m(s_{h+1}) \mid s_h\Big]\right] \\ &= \mathbb{E}\left[\sum_{h=t_i}^{t_{i+1}-1} c(s_h,a_h) + \mathbb{E}\Big[V_{h+1}^m(s_{h+1}) \mid s_h\Big] - V_h^m(s_h)\right] = 0.\end{aligned}$$

Thus, we have

$$9B_\star^2\mathbb{E}[Q]^2 \leq 324B_\star^2\mathbb{E}[Q] + 18B_\star^2,$$

and solving for $\mathbb{E}[Q]$ we obtain $\mathbb{E}[Q] \leq 37$, so

$$\mathbb{E}[(C^m)^2 \mid \bar{U}^m] = \mathbb{E}\left[\left(\sum_{h=1}^{H} \hat{c}(s_h, a_h) + \hat{c}_f(s_{H+1})\right)^2\right] \leq 2 \cdot 10^4 B_\star^2.$$

□

Lemma D.2.11 (Variance Difference is Upper Bounded by Value Difference)**.** *Assume that the value at time step $h+1$ is optimistic, i.e., $\underline{V}^m_{h+1}(s) \leq V^*_{h+1}(s)$ for all $s \in \mathscr{S}$. Conditioning on the event $\cap_m E^{pv2}(m)$ it holds for all $(s,a) \in \mathscr{S} \times \mathscr{A}$ that*

$$\sqrt{2L_m}\frac{\left|\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}^m_{h+1})} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V^*_{h+1})}\right|}{\sqrt{n^{m-1}(s,a) \vee 1}} \leq \frac{1}{\alpha}\mathbb{E}_{\bar{P}^{m-1}(\cdot|s,a)}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] + \frac{(5+\alpha/2)B_\star L_m}{n^{m-1}(s,a) \vee 1},$$

for any $\alpha > 0$.

Proof. Conditioning on $\cap_m E^{pv2}(m)$, the following relations hold.

$$\begin{aligned}\left|\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}^m_{h+1})} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V^*_{h+1})}\right| &\leq \left|\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}^m_{h+1})} - \sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(V^*_{h+1})}\right| \\ &\quad + \sqrt{\frac{12B_\star^2 L_m}{n^{m-1}(s,a) \vee 1}} \\ &\leq \sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(V^*_{h+1} - \underline{V}^m_{h+1})} + \sqrt{\frac{12B_\star^2 L_m}{n^{m-1}(s,a) \vee 1}} \\ &\leq \sqrt{\mathbb{E}_{\bar{P}^{m-1}}\left[(V^*_{h+1}(s') - \underline{V}^m_{h+1}(s'))^2\right]} + \sqrt{\frac{12B_\star^2 L_m}{n^{m-1}(s,a) \vee 1}} \\ &\leq \sqrt{B_\star \mathbb{E}_{\bar{P}^{m-1}}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right]} + \sqrt{\frac{12B_\star^2 L_m}{n^{m-1}(s,a) \vee 1}},\end{aligned}$$

where the second inequality is by Theorem D.5.3, and the last relation holds since $V^*_{h+1}(s'), \underline{V}^m_{h+1}(s') \in [0, B_\star]$ (the first, by model assumption, and the second, by the update rule) and since

$V^*_{h+1}(s') \geq \underline{V}^m_{h+1}(s')$ by the assumption the value is optimistic. Thus,

$$\begin{aligned}
\sqrt{2L_m}\frac{\left|\sqrt{\mathrm{Var}_{\bar{P}^{m-1}(\cdot|s,a)}(\underline{V}^m_{h+1})} - \sqrt{\mathrm{Var}_{P(\cdot|s,a)}(V^*_{h+1})}\right|}{\sqrt{n^{m-1}(s,a)}} &\leq \\
&\leq \sqrt{\mathbb{E}_{\bar{P}^{m-1}}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right]}\sqrt{\frac{2B_\star L_m}{n^{m-1}(s,a) \vee 1}} \\
&\quad + \frac{\sqrt{24}B_\star L_m}{n^{m-1}(s,a) \vee 1} \\
&\leq \frac{1}{\alpha}\mathbb{E}_{\bar{P}^{m-1}}\left[V^*_{h+1}(s') - \underline{V}^m_{h+1}(s')\right] + \frac{(5+\alpha/2)B_\star L_m}{n^{m-1}(s,a) \vee 1},
\end{aligned}$$

where the last inequality is by Young's inequality, $ab \leq \frac{1}{\alpha}a^2 + \frac{\alpha}{4}b^2$. □

D.2.6 Useful results for reinforcement learning analysis

Lemma D.2.12 (Cumulative Visitation Bound for Stationary MDP, e.g., [EMM21], Lemma 23)**.** *It holds that*

$$\sum_{m=1}^{M}\sum_{s,a}\mathbb{I}\{n^{m-1}(s,a) \geq H\}\frac{\sum_{h=1}^{H}\mathbb{I}\{s^m_h = s, a^m_h = a\}}{n^{m-1}(s,a) \vee 1} \leq 3SA\log(MH).$$

Proof. Recall that we recompute the optimistic policy only in the end of episodes in which the number of visits to some state-action pair was doubled. In this proof we refer to a sequence of consecutive episodes in which we did not perform a recomputation of the optimistic policy by the name of *epoch*. Let E be the number of epochs and note that $E \leq SA\log(MH)$ because the number of visits to each state-action pair (s,a) can be doubled at most $\log(MH)$ times. Next, denote by $\tilde{n}^e(s,a)$ the number of visits to (s,a) until the end of epoch e and by $\widetilde{N}^e(s,a)$ the number of visits to (s,a) during epoch e. The following relations hold for any fixed (s,a) pair.

$$
\begin{aligned}
\sum_{m=1}^{M} \mathbb{I}\{n^{m-1}(s,a) \geq H\} \frac{\sum_{h=1}^{H} \mathbb{I}\{s_h^m = s, a_h^m = a\}}{n^{m-1}(s,a) \vee 1} &= \\
&= \sum_{e=1}^{E} \mathbb{I}\{\tilde{n}^{e-1}(s,a) \geq H\} \frac{\widetilde{N}^e(s,a)}{\tilde{n}^{e-1}(s,a)} \\
&= \sum_{e=1}^{E} \mathbb{I}\{\tilde{n}^{e-1}(s,a) \geq H\} \frac{\widetilde{N}^e(s,a)}{\tilde{n}^{e}(s,a)} \frac{\tilde{n}^e(s,a)}{\tilde{n}^{e-1}(s,a)} \\
&\leq 3 \sum_{e=1}^{E} \mathbb{I}\{\tilde{n}^{e-1}(s,a) \geq H\} \frac{\widetilde{N}^e(s,a)}{\tilde{n}^{e}(s,a)} \\
&= 3 \sum_{e=1}^{E} \mathbb{I}\{\tilde{n}^{e-1}(s,a) \geq H\} \frac{\tilde{n}^e(s,a) - \tilde{n}^{e-1}(s,a)}{n^{e}(s,a)} \\
&\leq 3 \sum_{e=1}^{E} \mathbb{I}\{\tilde{n}^{e-1}(s,a) \geq H\} \log\left(\frac{\tilde{n}^e(s,a)}{\tilde{n}^{e-1}(s,a)}\right) \\
&\leq 3\mathbb{I}\{\tilde{n}^{E}(s,a) \geq H\} (\log \tilde{n}^E(s,a) - \log(H)) \\
&\leq 3 \log\big(\tilde{n}^E(s,a) \vee 1\big),
\end{aligned}
$$

where the first inequality follows since $\frac{\tilde{n}^e(s,a)}{\tilde{n}^{e-1}(s,a)} \leq \frac{2\tilde{n}^{e-1}(s,a)+H}{\tilde{n}^{e-1}(s,a)} \leq 3$ for $\tilde{n}^{e-1}(s,a) \geq H$, and the second inequality follows by the inequality $\frac{a-b}{a} \leq \log \frac{a}{b}$ for $a \geq b > 0$. Applying Jensen's inequality we conclude the proof:

$$
\begin{aligned}
\sum_{m=1}^{M} \sum_{s,a} \mathbb{I}\{n^{m-1}(s,a) \geq H\} \frac{\sum_{h=1}^{H} \mathbb{I}\{s_h^m = s, a_h^m = a\}}{n^{m-1}(s,a) \vee 1} &\leq 3 \sum_{s,a} \log\big(\tilde{n}^E(s,a) \vee 1\big) \\
&\leq 3SA \log\left(\sum_{s,a} \tilde{n}^E(s,a)\right) \\
&\leq 3SA \log(MH).
\end{aligned}
$$

□

Lemma D.2.13 (Transition Difference to Next State Expectation, [EMSM21], Lemma 28)**.** *Let $Y \in \mathbb{R}^S$ be a vector such that $0 \leq Y(s) \leq 2H$ for all $s \in \mathscr{S}$. Let P_1 and P_2 be two transition models and $n \in \mathbb{R}_+^{SA}$. Let $\Delta P(\cdot \mid s,a) \in \mathbb{R}^S$ and $\Delta P(s'|s,a) = P_1(s'|s,a) - P_2(s'|s,a)$. Assume that*

$$
\forall (s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}, h \in [H] : \ |\Delta P(s'|s,a)| \leq \sqrt{\frac{C_1 L_m P_1(s'|s,a)}{n(s,a) \vee 1}} + \frac{C_2 L_m}{n(s,a) \vee 1},
$$

for some $C_1, C_2 > 0$. Then, for any $\alpha > 0$.

$$|\Delta P(\cdot \mid s,a) \cdot Y| \leq \frac{1}{\alpha}\mathbb{E}_{P_1(\cdot|s,a)}\left[Y(s')\right] + \frac{HL_m(2C_2 + \alpha S C_1/2)}{n(s,a) \vee 1}.$$

Lemma D.2.14 (Law of Total Variance, e.g., [AOM17])**.** *For any π the following holds.*

$$\mathbb{E}\left[\sum_{h=1}^{H} \mathrm{Var}_{P(\cdot|s_h,a_h)}(V_{h+1}^{\pi}) \mid \pi\right] = \mathbb{E}\left[\left(\sum_{h=1}^{H} c(s_h,a_h) + c_f(s_{H+1}) - V_1^{\pi}(s_1)\right)^2 \mid \pi\right].$$

D.3 Extending the reduction to unknown $B_\star$

In this section we assume $B_\star \geq 1$ to simplify presentation, but the results work similarly for $B_\star < 1$. To handle unknown $B_\star$, we leverage techniques from the adversarial SSP literature [RM21b, CL21] for learning the diameter of an SSP problem. Recall that the SSP-diameter D [TGV+20] is defined as $D = \max_{s \in \mathcal{S}} \min_{\pi : s \to A} T^\pi(s)$. So to compute D we can find the optimal policy with respect to the constant cost function $c_1(s,a) = 1$, and compute its cost-to-go function. [RM21b] utilize this observation to estimate the SSP-diameter. They show that one can estimate the expected time from a state s to the goal state g by running the `Bernstein-SSP` algorithm of [RCMK20] with unit costs for $L = \widetilde{O}(D^2 S^2 A)$ episodes and setting the estimator to be the average cost per episode times 10.

Inspired by their approach, we use the `Bernstein-SSP` algorithm on the the actual costs, in order to estimate the expected cost of the optimal policy. Although `Bernstein-SSP` suffers from sub-optimal regret, we run it only for a small number of episodes and therefore we will only suffer from a slightly larger additive factors in our regret bound, but keep minimax optimal regret for large enough K.

By similar proofs to Lemmas 26 and 27 from [RM21b, Appendix J], we can show that the cost-to-go from state s can be estimated up to a constant multiplicative factor by running `Bernstein-SSP` for $L = \widetilde{O}(T_\star^2 S^2 A)$ episodes. This is demonstrated in the following lemma, where the upper bound follows from the regret guarantees of `Bernstein-SSP` and the lower bound follows from concentration arguments (and noticing that the regret is minimized by playing the optimal policy, but even then it is not zero).

Lemma D.3.1. *Let $s \in \mathcal{S}$ and $L \geq 2400 T_\star^2 S^2 A \log^3 \frac{K T_\star S A}{\delta}$. Run* `Bernstein-SSP` *with initial state s for L episodes and denote by $\widetilde{B}_s$ the average cost per episode times* 10. *Then, with probability $1 - \delta$,*

$$V^{\pi^\star}(s) \leq \widetilde{B}_s \leq O(B_\star).$$

Thus, we use the first L visits to each state in order to estimate its cost-to-go. A state which was visited at least L times will be called *$B_\star$-known*, and otherwise *$B_\star$-unknown* (not to be confused with our previous definition of known state-action pair). To that end, we split the total time steps into E epochs. In epoch e, we apply our reduction to a virtual MDP $\mathcal{M}^e$ that is identical to $\mathcal{M}$ in $B_\star$-known states, but turns $B_\star$-unknown states into zero-cost sinks (like the goal state). For every state $s \in \mathcal{S}$ we maintain a `Bernstein-SSP` algorithm $\mathcal{B}_s$. Every time we reach a $B_\star$-unknown state s, we run an episode of $\mathcal{B}_s$ until the goal is reached.

Note that in the virtual MDP $\mathscr{M}^e$ we can compute an upper bound on the optimal cost-to-go using our estimates. Epoch e ends once some $B_\star$-unknown state s is visited L times and thus becomes $B_\star$-known. Therefore the number of epochs E is bounded by S. The important change, introduced by [CL21], is to not completely initialize our finite-horizon algorithm ALG in the beginning of a new epoch as this leads to an extra S factor in the regret. Instead, algorithm ALG inherits the experience (i.e., visit counters and accumulated costs) of the previous epoch in $B_\star$-known states.

The reduction without knowledge of $B_\star$ is presented in Algorithm 15, and next we prove that it maintains the same regret bound up to a slightly larger additive factor.

Theorem D.3.2. *Let ALG be an admissible algorithm for regret minimization in finite-horizon MDPs and denote its regret in M episodes by $\widehat{\mathscr{R}}_{\texttt{ALG}}(M)$. Then, running Algorithm 15 with ALG ensures that, with probability at least $1-2\delta$,*

$$R_K \le \widehat{\mathscr{R}}_{\texttt{ALG}}\left(4K + 4\cdot 10^4 SA\omega_{\texttt{ALG}} \log\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta} + 4\cdot 10^4 T_\star^2 S^3 A \log^3\frac{KT_\star SA}{\delta}\right)$$
$$+ O\left(B_\star\sqrt{K\log\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta}} + T_\star\omega_{\texttt{ALG}} SA\log^2\frac{KT_\star SA\omega_{\texttt{ALG}}}{\delta} + T_\star^3 S^3 A\log^4\frac{KT_\star SA}{\delta}\right),$$

where $\omega_{\texttt{ALG}}$ is a quantity that depends on the algorithm ALG and on S, A, H.

Using the reduction with the ULCVI algorithm, we can again obtain optimal regret for SSP.

Theorem D.3.3. *Running the reduction in Algorithm 15 with the finite-horizon regret minimization algorithm ULCVI ensures, with probability at least $1-2\delta$,*

$$R_K = O\left(B_\star\sqrt{SAK}\log\frac{KT_\star SA}{\delta} + T_\star^5 S^2 A\log^6\frac{KT_\star SA}{\delta} + T_\star^3 S^3 A\log^4\frac{KT_\star SA}{\delta}\right).$$

Algorithm 15 REDUCTION FROM SSP TO FINITE-HORIZON MDP WITH UNKNOWN $B_\star$

```
input: state space $\mathcal{S}$, action space $\mathcal{A}$, initial state $s_{\text{init}}$, goal state $g$, confidence parameter $\delta$, number of episodes $K$, bound on the expected time of the optimal policy $T_\star$ and algorithm ALG for regret minimization in finite-horizon MDPs.
initialize a Bernstein-SSP algorithm $\mathcal{B}_s$ with initial state $s$ and confidence parameter $\delta/S$ for every $s \in \mathcal{S}$.
set $L = 10^4 T_\star^2 S^2 A \log^3 \frac{K T_\star S A}{\delta}$, $\mathcal{S}^1_{\text{known}} = \{s_{\text{init}}\}$ and $N_f(s) = L\mathbb{1}\{s = s_{\text{init}}\}$ for every $s \in \mathcal{S}$.
run $\mathcal{B}_{s_{\text{init}}}$ for $L$ episodes and set $\widetilde{B}_{s_{\text{init}}}$ to be the average cost per episode times 10.
initialize ALG with state space $\widehat{\mathcal{S}} = \mathcal{S} \cup \{g\}$, action space $\mathcal{A}$, horizon $H = 8T_\star \log(8K)$, confidence parameter $\frac{\delta}{4S}$, terminal costs $\hat{c}_f(s) = 8\mathbb{1}\{s = s_{\text{init}}\}\widetilde{B}_{s_{\text{init}}}$ and bound on the expected cost of the optimal policy $9\widetilde{B}_{s_{\text{init}}}$.
initialize intervals counter $m \leftarrow 0$, time steps counter $t \leftarrow 1$ and epochs counter $e \leftarrow 1$.
for $k = L+1, \dots, K$ do
  set $s_t \leftarrow s_{\text{init}}$.
  while $s_t \neq g$ do
    set $m \leftarrow m+1$, feed initial state $s_t$ to ALG and obtain policy $\pi^m = \{\pi^m_h : \widehat{\mathcal{S}} \to \mathcal{A}\}_{h=1}^H$.
    for $h = 1, \dots, H$ do
      play action $a_t = \pi^m_h(s_t)$, suffer cost $C_t \sim c(s_t, a_t)$, and set $s^m_h = s_t, a^m_h = a_t, C^m_h = C_t$.
      observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ and set $t \leftarrow t+1$.
      if $s_t = g$ or $s_t \notin \mathcal{S}^{\rceil}_{\text{known}}$ then
        pad trajectory to be of length $H$ and BREAK.
      end if
    end for
    set $s^m_{H+1} = s_t$.
    feed trajectory $U^m = (s^m_1, a^m_1, \dots, s^m_H, a^m_H, s^m_{H+1})$ and costs $\{C^m_h\}_{h=1}^H$ to ALG.
    if $s_t \notin \mathcal{S}^e_{\text{known}}$ then
      set $N_f(s_t) \leftarrow N_f(s_t) + 1$ and run an episode of $\mathcal{B}_{s_t}$.
      if $N_f(s_t) = L$ then
        set $e \leftarrow e+1$ and $\mathcal{S}^e_{\text{known}} \leftarrow \mathcal{S}^{e-1}_{\text{known}} \cup \{s_t\}$.
        set $\widetilde{B}_{s_t}$ to be the average cost per episode of $\mathcal{B}_{s_t}$ times 10.
        reinitialize ALG by updating the terminal costs as $\hat{c}_f(s) = 8\mathbb{1}\{s \in \mathcal{S}^e_{\text{known}}\} \max_{\tilde{s} \in \mathcal{S}^e_{\text{known}}} \widetilde{B}_{\tilde{s}}$, updating the bound on the expected cost of the optimal policy $9 \max_{\tilde{s} \in \mathcal{S}^e_{\text{known}}} \widetilde{B}_{\tilde{s}}$ and deleting the history of ALG only in state $s_t$.
      end if
    end if
  end while
end for
```

D.3.1 Proof of Theorem D.3.2

We follow the analysis of the known $B_\star$ case under the event that Theorem D.3.1 holds for all states (which happens with probability at least $1-\delta$), i.e., $V^{\pi^\star}(s) \le \widetilde{B}_s \le O(B_\star)$ for every $s \in \mathscr{S}$. We start by decomposing the regret similarly to Theorem 6.3.1. Note that now there is an additional term that comes from the regret of the S `Bernstein-SSP` algorithms that are used to estimate $B_\star$.

Lemma D.3.4. *For $H = 8T_\star \log(8K)$, we have the following bound on the regret of Algorithm 15:*

$$R_K \le \widehat{\mathscr{R}}_{\texttt{ALG}}(M) + \sum_{m=1}^{M}\left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}_1^{\pi^m}(s_1^m)\right) + O\left(T_\star^2 B_\star S^3 A \log^3 \frac{KT_\star SA}{\delta}\right), \tag{D.12}$$

where M is the total number of intervals.

Remark 6. Note that now each interval is considered in the context of the current epoch, i.e., the current $B_\star$-known states. The finite-horizon cost-to-go $\widehat{V}^{\pi^m}$ is with respect to the MDP of $B_\star$-known states. Moreover, for interval m that ends in a $B_\star$-unknown state, the last state in the trajectory s_{H+1}^m will be a $B_\star$-unknown state and the length of the interval may be shorter than H (just like intervals that end in the goal state).

Proof. Every interval ends either in the goal state, in a $B_\star$-known state or in a $B_\star$-unknown state. The first two cases are similar to the proof of Theorem 6.3.1 because our estimates $\widetilde{B}_s$ in all $B_\star$-known states s are upper bounds on $V^{\pi^\star}(s)$. Importantly, we do not initialize `ALG` in the end of an epoch and this allows us to get its regret bound without an extra S factor. The reason is that `ALG` is an admissible (and thus optimistic) algorithm, so it operates based on the observations it collected. Another important note is that the cost in the virtual MDP $\mathscr{M}^e$ is always bounded by the cost in the actual MDP $\mathscr{M}$.

We now focus on the last case. Recall that if interval m ends in a $B_\star$-unknown state s, then the terminal cost is 0 and we run an episode of the `Bernstein-SSP` algorithm $\mathscr{B}_s$. Thus, the excess cost of running `Bernstein-SSP` algorithms is bounded by S times the `Bernstein-SSP` regret plus $SB_\star L$, i.e., we can bound it as follows

$$SB_\star L + O\left(B_\star^{3/2} S^2 \sqrt{AL} \log \frac{KT_\star SA}{\delta} + T_\star^{3/2} S^3 A \log^2 \frac{KT_\star SA}{\delta}\right).$$

To finish the proof we plug in the definition of L. □

Next, we bound the number of intervals. Again, we get a similar bound to Theorem 6.3.3 but with an additional term for all the intervals that ended in a $B_\star$-unknown state (there are at most SL such intervals).

Lemma D.3.5. *Assume that the reduction is performed using an admissible algorithm* $\mathtt{ALG}$*. Then, with probability at least* $1 - {}^{3\delta}/_{8}$,

$$M \leq 4\left(K + 10^4 SA\omega_{\mathtt{ALG}} \log\frac{KT_\star SA\omega_{\mathtt{ALG}}}{\delta} + 10^4 T_\star^2 S^3 A \log^3\frac{KT_\star SA}{\delta}\right).$$

Proof. The proof is based on the claim that in every interval there is a probability of at least $1/2$ that the agent reaches either the goal state, an unknown state-action pair or a $B_\star$-unknown state. This is proved similarly to Theorem D.1.3 since we can look at the MDP of $B_\star$-known states, and then the claim of Theorem D.1.3 is equivalent to reaching either the goal state, an unknown state-action pair or a $B_\star$-unknown state.

With this claim the proof follows easily by following the proof of Theorem 6.3.3. We simply define X^m to be 1 if an unknown state-action pair or the goal or a $B_\star$-unknown state were reached during interval m (and 0 otherwise). Then, we have

$$\sum_{m=1}^{M} X^m \leq K + SA\omega_{\mathtt{ALG}} \log\frac{MHSA}{\delta} + SL,$$

which implies the Lemma following the same argument based on Freedman's inequality. □

Finally, we bound the deviation of the actual cost in each interval from its expected value. The proof is exactly the same as Theorem 6.3.2. The second moment of the accumulated cost until reaching the goal, an unknown state-action pair or a $B_\star$-unknown state is of order $B_\star^2$, and therefore in almost all intervals (except for a finite number) the accumulated cost will be of order $B_\star$ with high probability (in other intervals the cost is trivially bounded by $H + O(B_\star)$).

Lemma D.3.6. *Assume that the reduction is performed using an admissible algorithm* $\mathtt{ALG}$*. Then, the following holds with probability at least* $1 - {}^{3\delta}/_{8}$,

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H} C_h^m + \hat{c}_f(s_{H+1}^m) - \widehat{V}_1^{\pi^m}(s_1^m)\right) = O\left(B_\star\sqrt{M\log\frac{M}{\delta}} + (H+B_\star)\omega_{\mathtt{ALG}} SA\log\frac{MKT_\star SA}{\delta}\right) + O\left((H+B_\star)T_\star^2 S^3 A\log^3\frac{KT_\star SA}{\delta}\right).$$

The proof of the theorem is finished by combining Theorems D.3.4 to D.3.6 together with the guarantees of the admissible algorithm `ALG` and Theorem D.3.1, similarly to Theorem 6.2.1.

D.4 Lower bound

In this section we prove Theorem 6.1.3 which lower bounds the expected regret of any learning algorithm for the case $B_\star < 1$. It complements the lower bound found in [RCMK20] for the case $B_\star \geq 1$.

By Yao's minimax principle, in order to derive a lower bound on the learner's regret, it suffices to show a distribution over MDP instances that forces any deterministic learner to suffer a regret of $\Omega(\sqrt{B_\star SAK})$ in expectation.

To construct this distribution, we follow [RCMK20] with a few modifications. We initially consider the simpler setting with two states: an initial state and the goal state. We now embed a hard MAB instance into our problem where the optimal action has an expected cost of $B_\star$. To that end, consider a distribution over MDPs where a special action $a^\star$ is chosen a-priori uniformly at random. Then, all actions lead to the goal state g with probability 1. The cost $C_k(s_{\text{init}}, a^\star)$ chosen at episode k is 1 w.p. $B_\star$ and 0 otherwise. The cost of any other action $a \neq a^\star$ is 1 w.p. $B_\star + \varepsilon$ and 0 otherwise, where $\varepsilon \in (0, 1/8)$ is a constant to be determined. Thus the optimal policy will always play $a^\star$ and we have $V^\star(s_{\text{init}}) = B_\star$.

Fix any deterministic learning algorithm, we shall now quantify the regret of the learner in terms of the number of times that it plays $a^\star$. Indeed, we have that the optimal cost is $B_\star$, and the learner loses ε in the regret each time she plays an action other than $a^\star$. Therefore,

$$\mathbb{E}[R_K] \geq \varepsilon \cdot (K - \mathbb{E}[N]),$$

where N is the number of times $a^\star$ was chosen in s_{init}.

We now introduce an additional distribution of the costs which denote by $\Pr_{\text{unif}}$. $\Pr_{\text{unif}}$ is identical to the distribution over the costs defined above, and denoted by Pr, except that $\Pr[C_k(s_{\text{init}}, a) = 1] = B_\star + \varepsilon$ for all actions $a \in A$ regardless of the choice of $a^\star$. We denote expectations over $\Pr_{\text{unif}}$ by $\mathbb{E}_{\text{unif}}$, and expectations over Pr by $\mathbb{E}$. The following lemma uses standard lower bound techniques used for multi-armed bandits (see, e.g., [JOA10, Theorem 13]) to bound the difference in the expectation of N when the learner plays in Pr compared to when it plays in $\Pr_{\text{unif}}$.

Lemma D.4.1. *Suppose that $B_\star \leq \frac{1}{2}$. Denote by $\Pr_{unif,a}$, $\mathbb{E}_{unif,a}$, $\Pr_a$, $\mathbb{E}_a$ the distributions and expectations defined above conditioned on $a^\star = a$. For any deterministic learner we have that $\mathbb{E}_a[N] \leq \mathbb{E}_{unif,a}[N] + \varepsilon K \sqrt{\mathbb{E}_{unif,a}[N]/B_\star}$.*

Proof. Fix any deterministic learner. Let us denote by $C^{(k)}$ the sequence of costs observed by the learner up to episode k and including. Now, as $N \le K$ and the fact that N is a deterministic function of $C^{(K)}$, $\mathbb{E}_a[N] \le \mathbb{E}_{\text{unif},a}[N] + K \cdot \text{TV}(\Pr_{\text{unif},a}[C^{(K)}], \Pr[C^{(K)}])$, and Pinsker's inequality yields

$$\text{TV}(\Pr_{\text{unif},a}[C^{(K)}], \Pr[C^{(K)}]) \le \sqrt{\frac{1}{2}\text{KL}(\Pr_{\text{unif},a}[C^{(K)}] \parallel \Pr_a[C^{(K)}])}. \tag{D.13}$$

Next, the chain rule of the KL divergence obtains

$$\begin{aligned}&\text{KL}(\Pr_{\text{unif},a}[C^{(K)}] \parallel \Pr_a[C^{(K)}]) \\ &\quad = \sum_{k=1}^{K}\sum_{C^{(k)}} \Pr_{\text{unif},a}[C^{(k)}] \cdot \text{KL}(\Pr_{\text{unif},a}[C_k(s_{\text{init}}, a_k) \mid C^{(k)}] \parallel \Pr_a[C_k(s_{\text{init}}, a_k) \mid C^{(k)}]),\end{aligned}$$

where a_k is the action chosen by the learner at episode k. (Recall that after which the model transition to the goal state and the episode ends.)

Observe that at any episode, since the learning algorithm is deterministic, the learner chooses an action given $C^{(k)}$ regardless of whether $C^{(k)}$ was generated under Pr or under $\Pr_{\text{unif},a}$. Thus, the $\text{KL}(\Pr_{\text{unif},a}[C_k(s_{\text{init}}, a_k) \mid C^{(k)}] \parallel \Pr_a[C_k(s_{\text{init}}, a_k) \mid C^{(k)}])$ is zero if $a_k \neq a_\star$, and otherwise

$$\begin{aligned}&\text{KL}(\Pr_{\text{unif},a}[C_k(s_{\text{init}}, a_k) \mid C^{(k)}] \parallel \Pr_a[C_k(s_{\text{init}}, a_k) \mid C^{(k)}]) \\ &\quad = (B_\star + \varepsilon)\log\Big(1 + \frac{\varepsilon}{B_\star}\Big) + (1 - B_\star - \varepsilon)\log\Big(1 - \frac{\varepsilon}{1 - B_\star}\Big) \\ &\quad \le \frac{\varepsilon^2}{B_\star(1 - B_\star)},\end{aligned}$$

where we used that $\log(1+x) \le x$ for all $x > -1$, and since we assume $B_\star \le \frac{1}{2}$ and $\varepsilon < \frac{1}{8}$ that imply $-\varepsilon/(1 - B_\star) \ge -\frac{1}{4} > -1$. Plugging the above back into Equation (D.13) and using $B_\star \le \frac{1}{2}$ gives the lemma. □

In the following result, we combine the lemma above with standard techniques from lower bounds of multi-armed bandits (see [ACBFS02] for example).

Theorem D.4.2. *Suppose that $B_\star \le \frac{1}{2}$, $\varepsilon \in (0, \frac{1}{8})$ and $A \ge 2$. For the problem described above we have that*

$$\mathbb{E}[R_K] \ge \varepsilon K\Big(\frac{1}{2} - \varepsilon\sqrt{\frac{K}{AB_\star}}\Big).$$

Proof of Theorem D.4.2. Note that as under $\Pr_{\text{unif}}$ the cost distributions of all actions are identical. Denote by N_a the number of times that the learner chooses action a in s_{init}. Therefore,

$$\sum_{a\in A}\mathbb{E}_{\text{unif},a}[N]=\sum_{a\in A}\mathbb{E}_{\text{unif}}[N_a]=\mathbb{E}_{\text{unif}}\left[\sum_{a\in A}N_a\right]=K. \tag{D.14}$$

Recall that $a^\star$ is sampled uniformly at random before the game starts. Then,

$$\begin{aligned}
\mathbb{E}[R_K]&=\frac{1}{A}\sum_{a\in A}\mathbb{E}_a[R_K]\\
&\geq K-\frac{1}{A}\sum_{a\in A}\mathbb{E}_a[N]\\
&\geq K-\frac{1}{A}\sum_{a\in A}\left(\mathbb{E}_{\text{unif},a}[N]+\varepsilon K\sqrt{\mathbb{E}_{\text{unif},a}[N]/B_\star}\right) && \text{(Theorem D.4.1)}\\
&\geq K-\frac{1}{A}\sum_{a\in A}\mathbb{E}_{\text{unif},a}[N]+\varepsilon K\sqrt{\frac{1}{AB_\star}\sum_{a\in A}\mathbb{E}_{\text{unif},a}[N]} && \text{(Jensen's inequality)}\\
&=K-\frac{K}{A}+\varepsilon K\sqrt{\frac{K}{AB_\star}}, && \text{(Equation (D.14))}
\end{aligned}$$

The theorem follows from $A\geq 2$ and by rearranging. □

Proof of Theorem 6.1.3. Consider the following MDP. Let $\mathscr{S}$ be the set of states disregarding g. The initial state is sampled uniformly at random from $\mathscr{S}$. Each $s\in\mathscr{S}$ has its own special action $a^\star_s$. All actions transition to the goal state with probability 1. The cost $C_k(s,a)$ of action $a\neq a^\star_s$ in episode k and state s is 1 with probability $B_\star+\varepsilon$ and 0 otherwise. The cost of $C_k(s,a^\star_s)$ is 1 with probability $B_\star$ and 0 otherwise.

Note that for each $s\in\mathscr{S}$, the learner is faced with a simple problem as the one described above from which it cannot learn about from other states $s'\neq s$. Therefore, we can apply Theorem D.4.2 for each $s\in\mathscr{S}$ separately and lower bound the learner's expected regret the sum of the regrets suffered at each $s\in\mathscr{S}$, which would depend on the number of times $s\in\mathscr{S}$ is drawn as the initial state. Since the states are chosen uniformly at random there are many states (constant fraction) that are chosen $\Theta(K/S)$ times. Summing the regret bounds of Theorem D.4.2 over only these states and choosing ε appropriately gives the sought-after bound.

Denote by K_s the number of episodes that start in each state $s\in\mathscr{S}$.

$$\mathbb{E}[R_K]\geq\sum_{s\in\mathscr{S}}\mathbb{E}\left[\varepsilon K_s\left(\frac{1}{2}-\varepsilon\sqrt{\frac{K_s}{AB_\star}}\right)\right]=\frac{\varepsilon K}{2}-\varepsilon^2\sqrt{\frac{1}{AB_\star}}\sum_{s\in\mathscr{S}}\mathbb{E}[K_s^{3/2}]. \tag{D.15}$$

Applying Cauchy-Schwartz inequality gives

$$\sum_{s\in\mathscr{S}} \mathbb{E}[K_s^{3/2}] \leq \sum_{s\in\mathscr{S}} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s^2]} = \sum_{s\in\mathscr{S}} \sqrt{\mathbb{E}[K_s]}\sqrt{\mathbb{E}[K_s]^2 + \mathrm{Var}[K_s]}$$
$$= \sum_{s\in\mathscr{S}} \sqrt{\frac{K}{S}}\sqrt{\frac{K^2}{S^2} + \frac{K}{S}\Big(1 - \frac{1}{S}\Big)} \leq K\sqrt{\frac{2K}{S}},$$

where we have used the expectation and variance formulas of the Binomial distribution. The lower bound is now given by applying the inequality above in Equation (D.15) and choosing $\varepsilon = \frac{1}{8}\sqrt{B_\star AS/K}$. □

D.5 General useful results

Lemma D.5.1 (Freedman's Inequality)**.** *Let $\{X_t\}_{t\geq 1}$ be a real valued martingale difference sequence adapted to a filtration $\{F_t\}_{t\geq 0}$. If $|X_t| \leq R$ a.s. then for any $\eta \in (0, 1/R), T \in \mathbb{N}$ it holds with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \leq \eta \sum_{t=1}^{T} \mathbb{E}[X_t^2 | F_{t-1}] + \frac{\log(1/\delta)}{\eta}.$$

Lemma D.5.2 (Consequences of Freedman's Inequality for Bounded and Positive Sequence of Random Variables, e.g., [EMSM21], Lemma 27)**.** *Let $\{Y_t\}_{t\geq 1}$ be a real valued sequence of random variables adapted to a filtration $\{F_t\}_{t\geq 0}$. Assume that for all $t \geq 1$ it holds that $0 \leq Y_t \leq C$ a.s., and $T \in \mathbb{N}$. Then, each of the following inequalities hold with probability at least $1-\delta$.*

$$\sum_{t=1}^{T} \mathbb{E}[Y_t | F_{t-1}] \leq \left(1 + \frac{1}{2C}\right) \sum_{t=1}^{T} Y_t + 2(2C+1)^2 \log \frac{1}{\delta}$$
$$\sum_{t=1}^{T} Y_t \leq 2 \sum_{t=1}^{T} \mathbb{E}[Y_t | F_{t-1}] + 4C \log \frac{1}{\delta}.$$

Lemma D.5.3 (Standard Deviation Difference, e.g., [ZB19])**.** *Let $V_1, V_2 : S \to \mathbb{R}$ be fixed mappings. Let $P(s)$ be a probability measure over the state space. Then, $\sqrt{\mathrm{Var}(V_1)} - \sqrt{\mathrm{Var}(V_2)} \leq \sqrt{\mathrm{Var}(V_1 - V_2)}$.*

E Supplementary Material for Chapter 7

E.1 Examples that illustrate some challenges in adversarial SSPs

E.1.1 Naive application of OMD fails in SSP

In general, the first policy that OMD picks is the one that maximizes the entropy, which is the uniform policy, i.e., $\pi^u(a \mid s) = 1/A$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$. Next we show that, in SSP, this might result in exponential cost of A^S already in the first episode. In the finite-horizon setting, this is not a concern because the cost in a single episode is always bounded by H, while in SSP it can be infinite.

Consider the following MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, s_{\text{init}}, g)$ with the state space $\mathcal{S} = \{1, \dots, S\}$. In every state i there is one action $a(i)$ (picked uniformly at random in advance) such that $P(i+1 \mid i, a(i)) = 1$, while the other actions return the agent to the initial state $s_{\text{init}} = 1$, i.e., $P(1 \mid i, a) = 1$ for every $a \neq a(i)$. Finally, the cost function (for the first episode in which OMD picks π^u) is simply $c(s,a) = 1$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Clearly the best policy in this case is to pick $a(i)$ in state i and then the total cost is S (the SSP-diameter in this example is also S). However, the uniform policy picks this action only with probability $1/A$ which yields exponential expected time to reach the goal (and therefore exponential cost). To see that consider the Bellman equations for π^u:

$$V^{\pi^u}(i) = 1 + \frac{1}{A} \cdot V^{\pi^u}(i+1) + (1 - \frac{1}{A}) \cdot V^{\pi^u}(1) \qquad \forall i = 1, \dots, S-1$$
$$V^{\pi^u}(S) = 1 + \frac{1}{A} \cdot 0 + (1 - \frac{1}{A}) \cdot V^{\pi^u}(1).$$

Solving these equations gives $V^{\pi^u}(s_{\text{init}}) = V^{\pi^u}(1) = \frac{A(A^S-1)}{A-1} \geq A^S$.

E.1.2 The expected time of the best policy in hindsight might be $\Omega(D/c_{min})$

The following example shows that the expected time of the best policy in hindsight might be $\Omega(D/c_{\min})$, and therefore there is no better apriori choice for τ.

Consider the MDP $\mathcal{M} = (\{s_{\text{init}}\}, \{a_1, a_2\}, P, s_{\text{init}}, g)$ that has only one state (other than the goal) and two actions.

Playing action a_1 transitions to the goal with probability $1/D$ and back to s_{init} with probability $1 - 1/D$, i.e., $P(s_{\text{init}} \mid s_{\text{init}}, a_1) = 1 - 1/D$ and $P(g \mid s_{\text{init}}, a_1) = 1/D$. Therefore, the expected time of the policy that plays a_1 is D and so the SSP-diameter is also bounded by D.

Playing action a_2 transitions to the goal with probability $2c_{\min}/D$ and back to s_{init} with probability $1 - 2c_{\min}/D$, i.e., $P(s_{\text{init}} \mid s_{\text{init}}, a_2) = 1 - 2c_{\min}/D$ and $P(g \mid s_{\text{init}}, a_2) = 2c_{\min}/D$. Therefore, the expected time of the policy that plays a_1 is ${}^{D}/_{2c_{\min}}$.

Apriori there is no way to tell if a_1 or a_2 will be the best policy in hindsight. For example, if $c(s_{\text{init}}, a_1) = 1$ and $c(s_{\text{init}}, a_2) = c_{\min}$ then a_2 is better, and if $c(s_{\text{init}}, a_1) = 1$ and $c(s_{\text{init}}, a_2) = 3c_{\min}$ then a_1 is better. Thus, the smallest possible choice for τ in this case is ${}^{D}/_{2c_{\min}} = \Omega(D/c_{\min})$.

E.1.3 A bound on the expected regret does not guarantee a high probability regret bound in SSP

In most online learning problems, algorithms that guarantee bounded regret in expectation also guarantee bounded regret with high probability. The way to show this (in most problems) is by Azuma inequality for bounded martingales. However, the SSP problem is unique in the sense that guaranteeing bounded regret in expectation is significantly easier than guaranteeing bounded regret with high probability. This is illustrated by the following simple example in which there exists a policy with 0 expected regret, but linear regret with constant probability of at least ${}^{1}/_{30}$.

Consider the MDP $\mathcal{M} = (\{s_{\text{init}}, s_1\}, \{a_1, a_2\}, P, s_{\text{init}}, g)$ that has only two states (other than the goal) and two actions. In state s_{init} playing action a_1 simply transitions to the goal, i.e., $P(g \mid s_{\text{init}}, a_1) = 1$. In this state playing action a_2 transitions to the goal with probability $p = 1 - \frac{1-c_{\min}}{10K}$ and transitions to state s_1 with probability $1-p$, i.e., $P(g \mid s_{\text{init}}, a_2) = p$ and $P(s_1 \mid s_{\text{init}}, a_2) = 1-p$. Moreover, in state s_1 both actions have the same effect. They transition to the goal with probability ${}^{1}/_{10K}$ and remain in state s_1 with probability $1 - {}^{1}/_{10K}$, i.e., $P(g \mid s_1, a_i) = {}^{1}/_{10K}$ and $P(s_1 \mid s_1, a_i) = 1 - {}^{1}/_{10K}$ for $i = 1, 2$.

Now consider the simple case where the cost function is the same for all episodes. Playing action a_1 always suffers a cost of 1, i.e., $c(s_{\text{init}}, a_1) = c(s_1, a_1) = 1$. Playing action a_2 suffers cost of $c_{\min}$ in s_{init} but cost of 1 in s_1, i.e., $c(s_{\text{init}}, a_2) = c_{\min}$ and $c(s_1, a_2) = 1$. There are only two policies: π_1 plays action a_1 in state s_{init}, and π_2 plays a_2. Notice that both policies have the same expected cost since clearly $V^{\pi_1}(s_{\text{init}}) = 1$ and

$$V^{\pi_2}(s_{\text{init}}) = c_{\min} + p \cdot 0 + (1-p) \cdot 10K = c_{\min} + \frac{1 - c_{\min}}{10K} \cdot 10K = 1.$$

Moreover, both have similar expected time since clearly $T^{\pi_1}(s_{\text{init}}) = 1$ and

$$T^{\pi_2}(s_{\text{init}}) = 1 + p \cdot 0 + (1-p) \cdot 10K = 1 + \frac{1 - c_{\min}}{10K} \cdot 10K = 2 - c_{\min} \le 2.$$

Thus, playing policy π_2 in all episodes has optimal expected regret of 0 since

$$\mathbb{E}[R_K] = \mathbb{E}\Big[\sum_{k=1}^{K} V^{\pi_2}(s_{\text{init}}) - 1\Big] = \mathbb{E}\Big[\sum_{k=1}^{K} 1 - 1\Big] = 0.$$

However, we now show that with probability at least $1/2$ the actual regret is linear. Define the event E_k – in episode k the agent's cost was at most $2K$. Now define $E = \bigcap_{k=1}^{K} E_k$ as the event that E_k occurs for all episodes. Notice that if E does not occur than the regret is linear in K since in some episode k the cost was at least $2K$ while the overall cost of π_1 in all episodes is just K. The following lemma proves that event E_k occurs with probability at most $1 - 1/26K$ and therefore event E indeed occurs with probability at most $(1 - 1/26K)^K \le e^{-1/26} \le 29/30$.

Lemma E.1.1. *For every $k = 1, \dots, K$ it holds that* $\Pr[E_k] \le 1 - 1/26K$.

Proof. Recall that E_k is the event that the actual cost of the learner in episode k is bounded by $2K$. The probability of that is the probability to transition to the goal from s_{init} or to transition to s_1 and stay there for at most $2K$ steps. Thus,

$$\begin{aligned}
\Pr[E_k] &\leq p + (1-p)\sum_{i=1}^{2K}(1-\frac{1}{10K})^i \cdot \frac{1}{10K} \\
&= 1 - \frac{1-c_{\min}}{10K} + \frac{1-c_{\min}}{100K^2}\sum_{i=1}^{2K}(1-\frac{1}{10K})^i \\
&\leq 1 - \frac{1-c_{\min}}{10K} + \frac{1-c_{\min}}{100K^2} \cdot \frac{1-(1-\frac{1}{10K})^{2K+1}}{1/10K} \\
&= 1 - \frac{1-c_{\min}}{10K} + \frac{1-c_{\min}}{10K} \cdot \big(1-(1-\frac{1}{10K})^{2K+1}\big) \\
&\leq 1 - \frac{1-c_{\min}}{10K} + \frac{1-c_{\min}}{10K} \cdot \frac{1}{5} \leq 1 - \frac{1-c_{\min}}{13K} \leq 1 - \frac{1}{26K},
\end{aligned}$$

where the third inequality holds for large enough K since $(1-\frac{1}{10K})^{2K+1} \to e^{-1/5}$, and the last inequality holds for $c_{\min} \leq 1/2$. □

E.2 Implementation details for SSP-O-REPS

E.2.1 Computing q_k

Before describing the algorithm, some more definitions are in order. First, define $\mathrm{KL}(q \parallel q')$ as the unnormalized Kullback-Leibler divergence between two occupancy measures q and q':

$$\mathrm{KL}(q \parallel q') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} q(s,a)\log\frac{q(s,a)}{q'(s,a)} + q'(s,a) - q(s,a).$$

Furthermore, let $R(q)$ define the unnormalized negative entropy of the occupancy measure q:

$$R(q) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} q(s,a)\log q(s,a) - q(s,a).$$

SSP-O-REPS chooses its occupancy measures as follows:

$$\begin{aligned} q_1 = q^{\pi_1} &= \arg\min_{q\in\Delta(\mathcal{M})(D/c_{\min})} R(q) \\ q_{k+1} = q^{\pi_{k+1}} &= \arg\min_{q\in\Delta(\mathcal{M})(D/c_{\min})} \eta\langle q, c^k\rangle + \mathrm{KL}(q \parallel q_k). \end{aligned}$$

As shown by [ZN13], each of these steps can be split into an unconstrained minimization step, and a projection step. Thus, q_1 can be computed as follows:

$$\begin{aligned} q'_1 &= \arg\min_q R(q) \\ q_1 &= \arg\min_{q\in\Delta(\mathcal{M})(D/c_{\min})} \mathrm{KL}(q \parallel q'_1), \end{aligned}$$

where q'_1 has a closed-from solution $q'_1(s,a) = 1$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Similarly, q_{k+1} is computed as follows for every $k = 1,\dots,K-1$:

$$\begin{aligned} q'_{k+1} &= \arg\min_q \eta\langle q, c^k\rangle + \mathrm{KL}(q \parallel q_k) \\ q_{k+1} &= \arg\min_{q\in\Delta(\mathcal{M})(D/c_{\min})} \mathrm{KL}(q \parallel q'_{k+1}), \end{aligned}$$

where again q'_{k+1} has a closed-from solution $q'_{k+1}(s,a) = q_k(s,a)e^{-\eta c^k(s,a)}$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Therefore, we just need to show that the projection step can be computed efficiently (the implementation follows [ZN13]). We start by formulating the projection step as a constrained convex optimization problem:

$$\begin{aligned}
\min_q \quad & \mathrm{KL}(q \,\|\, q'_{k+1}) \\
s.t. \quad & \sum_{a\in\mathscr{A}} q(s,a) - \sum_{s'\in\mathscr{S}} \sum_{a'\in\mathscr{A}} P(s \mid s',a') q(s',a') = \mathbb{I}\{s = s_{\mathrm{init}}\} && \forall s \in \mathscr{S} \\
& \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} q(s,a) \le \frac{D}{c_{\min}} \\
& q(s,a) \ge 0 && \forall (s,a) \in \mathscr{S} \times \mathscr{A}
\end{aligned}$$

To solve the problem, consider the Lagrangian:

$$\begin{aligned}
\mathscr{L}(q,\lambda,v) &= \mathrm{KL}(q \,\|\, q'_{k+1}) + \lambda \left(\sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} q(s,a) - \frac{D}{c_{\min}} \right) \\
&\quad + \sum_{s\in\mathscr{S}} v(s) \left(\sum_{s'\in\mathscr{S}} \sum_{a'\in\mathscr{A}} P(s \mid s',a') q(s',a') + \mathbb{I}\{s = s_{\mathrm{init}}\} - \sum_{a\in\mathscr{A}} q(s,a) \right) \\
&= \mathrm{KL}(q \,\|\, q'_{k+1}) + \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} q(s,a) \left(\lambda + \sum_{s'\in\mathscr{S}} P(s' \mid s,a) v(s') - v(s) \right) \\
&\quad + v(s_{\mathrm{init}}) - \lambda \frac{D}{c_{\min}}
\end{aligned}$$

where λ and $\{v(s)\}_{s\in\mathscr{S}}$ are Lagrange multipliers. Differentiating the Lagrangian with respect to any $q(s,a)$, we get

$$\frac{\partial \mathscr{L}(q,\lambda,v)}{\partial q(s,a)} = \log q(s,a) - \log q'_{k+1}(s,a) + \lambda + \sum_{s'\in\mathscr{S}} P(s' \mid s,a) v(s') - v(s).$$

Hence, setting the gradient to zero, we obtain the formula for $q_{k+1}(s,a)$:

$$\begin{aligned}
q_{k+1}(s,a) &= q'_{k+1}(s,a) e^{-\lambda - \sum_{s'\in\mathscr{S}} P(s'|s,a)v(s') + v(s)} \\
&= q_k(s,a) e^{-\lambda - \eta c^k(s,a) - \sum_{s'\in\mathscr{S}} P(s'|s,a)v(s') + v(s)} \\
&= q_k(s,a) e^{-\lambda + B_k^v(s,a)}, \qquad \text{(E.1)}
\end{aligned}$$

where the second equality follows from the formula of $q'_{k+1}(s,a)$, and setting $c_0(s,a) = 0$ and $q_0(s,a) = 1$ for every $s \in \mathscr{S}$ and $a \in \mathscr{A}$. The last equality follows by defining $B_k^v(s,a) = v(s) - \eta c^k(s,a) - \sum_{s'\in\mathscr{S}} P(s' \mid s,a) v(s')$.

We now need to compute the value of λ and v at the optimum. To that end, we write the dual problem $\mathscr{D}(\lambda,v)=\min_q \mathscr{L}(q,\lambda,v)$ by substituting q_{k+1} back into $\mathscr{L}$:

$$\begin{aligned}\mathscr{D}(\lambda,v) &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q'_{k+1}(s,a) - \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_{k+1}(s,a) + v(s_{\text{init}}) - \lambda\frac{D}{c_{\min}} \\ &= -\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_k(s,a)e^{-\lambda+B_k^v(s,a)} + v(s_{\text{init}}) - \lambda\frac{D}{c_{\min}} + \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q'_{k+1}(s,a).\end{aligned}$$

Now we obtain λ and v by maximizing the dual. Equivalently, we can minimize the negation of the dual (and ignore the term $\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q'_{k+1}(s,a)$), that is:

$$\lambda_{k+1}, v_{k+1} = \arg\min_{\lambda\geq 0, v} \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_k(s,a)e^{-\lambda+B_k^v(s,a)} + \lambda\frac{D}{c_{\min}} - v(s_{\text{init}}).$$

This is a convex optimization problem with only non-negativity constraints (and no constraints about the relations between the variables), which can be solved efficiently using iterative methods like gradient descent.

E.2.2 Computing the SSP-diameter and the fast policy

The fast policy π^f is a deterministic stationary policy that minimizes the time to the goal state from all states simultaneously (its existence is similar to regular MDPs, for a detailed proof see [BT91]). Thus, π^f is the optimal policy w.r.t the constant cost function $c(s,a) = 1$ for every $s\in\mathscr{S}$ and $a\in\mathscr{A}$.

Finding the optimal policy of an SSP instance is known as the planning problem. By [BT91], this problem can be solved efficiently using Linear Programming (LP), Value Iteration (VI) or Policy Iteration (PI).

The SSP-diameter D is an upper bound on the expected time it takes to reach the goal from some state, and therefore $D=\max_{s\in\mathscr{S}} T^{\pi^f}(s)$. Thus, in order to compute π^f and D we need to perform the following steps:

1. Compute the optimal policy π^f w.r.t the constant cost function $c(s,a)=1$.

2. Compute $T^{\pi^f}(s)$ for every $s\in\mathscr{S}$ by solving the linear Bellman equations:

$$T^{\pi^f}(s) = 1 + \sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} \pi^f(a\mid s)P(s'\mid s,a)T^{\pi^f}(s') \quad \forall s\in\mathscr{S}.$$

3. Set $D=\max_{s\in\mathscr{S}} T^{\pi^f}(s)$.

E.3 Pseudo-code for SSP-O-REPS

Algorithm 16 SSP-O-REPS

```
**Input:** state space $\mathscr{S}$, action space $\mathscr{A}$, transition function $P$, minimal cost $c_{\min}$, optimization parameter $\eta$.
**Initialization:**
Compute the SSP-diameter $D$ (see Section E.2.2).
Set $q_0(s,a) = 1$ and $c_0(s,a) = 0$ for every $(s,a) \in \mathscr{S} \times \mathscr{A}$.
**for** $k = 1,2,\ldots$ **do**
    Compute $\lambda_k, v_k$ as follows (using, e.g., gradient descent):
        $$\lambda_k, v_k = \arg\min_{\lambda \geq 0, v} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_{k-1}(s,a) e^{-\lambda + B^v_{k-1}(s,a)} + \lambda \frac{D}{c_{\min}} - v(s_{\text{init}}),$$
    where $B^v_k(s,a) = v(s) - \eta c^k(s,a) - \sum_{s' \in \mathscr{S}} P(s' \mid s,a) v(s')$.
    Compute $q_k$ as follows for every $(s,a) \in \mathscr{S} \times \mathscr{A}$:
        $$q_k(s,a) = q_{k-1}(s,a) e^{-\lambda_k + B^{v_k}_{k-1}(s,a)}.$$
    Compute $\pi_k$ as follows for every $(s,a) \in \mathscr{S} \times \mathscr{A}$:
        $$\pi_k(a \mid s) = \frac{q_k(s,a)}{\sum_{a' \in \mathscr{A}} q_k(s,a')}.$$
    Set $s^k_1 \leftarrow s_{\text{init}}$, $i \leftarrow 1$.
    **while** $s^k_i \neq g$ **do**
        Play action according to $\pi_k$, i.e., $a^k_i \sim \pi_k(\cdot \mid s^k_i)$.
        Observe next state $s^k_{i+1} \sim P(\cdot \mid s^k_i, a^k_i)$, $i \leftarrow i+1$.
    **end while**
    Set $I^k \leftarrow i - 1$.
    Observe cost function $c^k$ and suffer cost $\sum_{j=1}^{I^k} c^k(s^k_j, a^k_j)$.
**end for**
```

E.4 Proofs for Section 7.2.1

Lemma E.4.1. *It holds that* $q^{\pi^\star} \in \Delta(\mathscr{M})(\frac{D}{c_{min}})$.

Proof. Denote by π^f the fast policy, i.e., $\pi^f = \arg\min_{\pi \in \Pi_{\text{proper}}} T^\pi(s_{\text{init}})$. By definition of the SSP-diameter we have that $T^{\pi^f}(s_{\text{init}}) \le D$. Now, recall that $\pi^\star$ is the best policy in hindsight and therefore

$$\frac{1}{K}\sum_{k=1}^K V_k^{\pi^\star}(s_{\text{init}}) \le \frac{1}{K}\sum_{k=1}^K V_k^{\pi^f}(s_{\text{init}}) \le \frac{1}{K}\sum_{k=1}^K T^{\pi^f}(s_{\text{init}}) \le D, \tag{E.2}$$

where the second inequality follows because $c^k(s,a) \le 1$.

However, we also have that $c^k(s,a) \ge c_{\min}$ and therefore $V_k^{\pi^\star}(s_{\text{init}}) \ge c_{\min} T^{\pi^\star}(s_{\text{init}})$. Thus, combining with Equation (E.2), we obtain

$$c_{\min} T^{\pi^\star}(s_{\text{init}}) \le \frac{1}{K}\sum_{k=1}^K V_k^{\pi^\star}(s_{\text{init}}) \le D.$$

This finishes the proof since $T^{\pi^\star}(s_{\text{init}}) \le \frac{D}{c_{\min}}$. □

E.4.1 Proof of Theorem 7.2.1

Lemma E.4.2. *Let* $\tau \ge 1$. *For every* $q \in \Delta(\mathscr{M})(\tau)$ *it holds that* $R(q) \le \tau \log \tau$.

Proof.

$$\begin{aligned}
R(q) &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log q(s,a) - \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a) \\
&\le \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log q(s,a) \\
&= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log \frac{q(s,a)}{\tau} + \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log\tau \\
&\le \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log\tau \le \tau\log\tau,
\end{aligned}$$

where the first two inequalities follow from non-positivity, and the last one from the definition of $\Delta(\mathscr{M})(\tau)$. □

Lemma E.4.3. *Let* $\tau \ge 1$. *For every* $q \in \Delta(\mathscr{M})(\tau)$ *it holds that* $-R(q) \le \tau(1+\log(SA))$.

Proof. Similarly to Theorem E.4.2 we have that

$$\begin{aligned}-R(q) &= -\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log\frac{q(s,a)}{\tau} + \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a) - \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q(s,a)\log\tau \\ &\le -\tau\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} \frac{q(s,a)}{\tau}\log\frac{q(s,a)}{\tau} + \tau \le \tau\log(SA) + \tau,\end{aligned}$$

where the first inequality follows because the last term is non-positive and from the definition of $\Delta(\mathscr{M})(\tau)$, and the last inequality follows from properties of Shannon's entropy. □

Proof of Theorem 7.2.1. We start with a fundamental inequality of OMD (see, e.g., [ZN13]) that holds for every $q \in \Delta(\mathscr{M})(D/c_{\min})$ (by Theorem E.4.1 it also holds for $q^{\pi^\star}$),

$$\sum_{k=1}^{K}\langle q_k - q^{\pi^\star}, c^k\rangle \le \sum_{k=1}^{K}\langle q_k - q'_{k+1}, c^k\rangle + \frac{\mathrm{KL}(q^{\pi^\star}\parallel q_1)}{\eta}. \tag{E.3}$$

For the first term we use the exact form of q'_{k+1} and the inequality $e^x \ge 1 + x$ to obtain

$$q'_{k+1}(s,a) = q_k(s,a)e^{-\eta c^k(s,a)} \ge q_k(s,a) - \eta q_k(s,a)c^k(s,a).$$

We substitute this back and obtain

$$\begin{aligned}\sum_{k=1}^{K}\langle q_k - q'_{k+1}, c^k\rangle &\le \eta\sum_{k=1}^{K}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_k(s,a)c^k(s,a)^2 \le \eta\sum_{k=1}^{K}\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q_k(s,a) \\ &= \eta\sum_{k=1}^{K} T^{\pi_k}(s_{\text{init}}) \le \eta K\frac{D}{c_{\min}},\end{aligned} \tag{E.4}$$

where the last inequality follows from the definition of $\Delta(\mathscr{M})(D/c_{\min})$.

Next we use Theorems E.4.2 and E.4.3 to bound the second term of Equation (E.3). Recall that q_1 minimizes R in $\Delta(\mathscr{M})(D/c_{\min})$, this implies that $\langle\nabla R(q_1), q^{\pi^\star} - q_1\rangle \ge 0$ because otherwise we could decrease R by taking small step in the direction $q^{\pi^\star} - q_1$. Thus we obtain

$$\begin{aligned}\mathrm{KL}(q^{\pi^\star}\parallel q_1) &= R(q^{\pi^\star}) - R(q_1) - \langle\nabla R(q_1), q^{\pi^\star} - q_1\rangle \le R(q^{\pi^\star}) - R(q_1) \\ &\le \frac{D}{c_{\min}}\log\frac{D}{c_{\min}} + \frac{D}{c_{\min}}(1+\log(SA)) \le \frac{3D}{c_{\min}}\log\frac{DSA}{c_{\min}}.\end{aligned} \tag{E.5}$$

By substituting Equations (E.4) and (E.5) into Equation (E.3) and choosing $\eta = \sqrt{\frac{3\log\frac{DSA}{c_{\min}}}{K}}$, we obtain,

$$\sum_{k=1}^{K} \langle q_k - q^{\pi^\star}, c^k \rangle \leq \eta K \frac{D}{c_{\min}} + \frac{3D}{c_{\min}\eta} \log \frac{DSA}{c_{\min}} \leq \frac{2D}{c_{\min}} \sqrt{3K \log \frac{DSA}{c_{\min}}}. \tag{E.6}$$

This finishes the proof since

$$\mathbb{E}[R_K] = \mathbb{E}\left[\sum_{k=1}^{K} \langle q_k - q^{\pi^\star}, c^k \rangle\right].$$

□

E.4.2 SSP-O-REPS picks proper policies

For every policy π_k chosen by SSP-O-REPS it holds that $T^{\pi_k}(s_{\text{init}}) \leq D/c_{\min}$. If there exists some state $s \in \mathscr{S}$ such that $T^{\pi_k}(s) = \infty$, then the probability to reach it must be zero, since otherwise $T^{\pi_k}(s_{\text{init}}) = \infty$. Thus there exists $B > 0$ such that if s is reachable from s_{init} using π_k then $T^{\pi_k}(s) \leq B$. By Theorem E.6.1, this implies that the goal state will be reached in every episode with probability 1. Thus, all policies chosen by SSP-O-REPS are proper.

E.5 Pseudo-code for SSP-O-REPS2

Algorithm 17 SSP-O-REPS2

```
Input: state space $\mathscr{S}$, action space $\mathscr{A}$, transition function $P$, minimal cost $c_{\min}$, optimization parameter $\eta$.
Initialization:
Compute the SSP-diameter $D$ and the fast policy $\pi^f$ (see Section E.2.2).
Set $q_0(s,a) = 1$ and $c_0(s,a) = 0$ for every $(s,a) \in \mathscr{S} \times \mathscr{A}$.
for $k = 1,2,\dots$ do
    Compute $\lambda_k, v_k$ as follows (using, e.g., gradient descent):
        $$\lambda_k, v_k = \arg\min_{\lambda \geq 0, v} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} q_{k-1}(s,a) e^{-\lambda + B^v_{k-1}(s,a)} + \lambda \frac{D}{c_{\min}} - v(s_{\text{init}}),$$
    where $B^v_k(s,a) = v(s) - \eta c^k(s,a) - \sum_{s' \in \mathscr{S}} P(s' \mid s,a) v(s')$.
    Compute $q_k$ as follows for every $(s,a) \in \mathscr{S} \times \mathscr{A}$: $q_k(s,a) = q_{k-1}(s,a) e^{-\lambda_k + B^{v_k}_{k-1}(s,a)}$.
    Compute $\pi_k$ as follows for every $(s,a) \in \mathscr{S} \times \mathscr{A}$: $\pi_k(a \mid s) = \frac{q_k(s,a)}{\sum_{a' \in \mathscr{A}} q_k(s,a')}$.
    Set $T^{\pi_k}(s) \leftarrow \frac{D}{c_{\min}}$ for every $s \in \mathscr{S}$ such that $q^{\pi_k}(s) = \sum_{a \in \mathscr{A}} q^{\pi_k}(s,a) = 0$.
    Compute $T^{\pi_k}$ by solving the following linear equations (the Bellman equations):
        $$T^{\pi_k}(s) = 1 + \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} \pi_k(a \mid s) P(s' \mid s,a) T^{\pi_k}(s') \quad \forall s \in \{s \in \mathscr{S} : \sum_{a \in \mathscr{A}} q^{\pi_k}(s,a) > 0\}.$$
    Set $s^k_1 \leftarrow s_{\text{init}}$, $i \leftarrow 1$.
    while $s^k_i \neq g$ and $T^{\pi_k}(s^k_i) < \frac{D}{c_{\min}}$ do
        Play action according to $\pi_k$, i.e., $a^k_i \sim \pi_k(\cdot \mid s^k_i)$.
        Observe next state $s^k_{i+1} \sim P(\cdot \mid s^k_i, a^k_i)$, $i \leftarrow i+1$.
    end while
    while $s^k_i \neq g$ do
        Play action according to $\pi^f$, i.e., $a^k_i \sim \pi^f(\cdot \mid s^k_i)$.
        Observe next state $s^k_{i+1} \sim P(\cdot \mid s^k_i, a^k_i)$, $i \leftarrow i+1$.
    end while
    Set $I^k \leftarrow i-1$, observe cost function $c^k$ and suffer cost $\sum_{j=1}^{I^k} c^k(s^k_j, a^k_j)$.
end for
```

E.6 Proofs for Section 7.2.2

E.6.1 Proof of Theorem 7.2.2

Lemma E.6.1. *Let σ be a strategy such that the expected time of reaching the goal state when starting at state s is at most τ for every $s \in \mathscr{S}$. Then, the probability that σ takes more than m steps to reach the goal state is at most $2e^{-\frac{m}{4\tau}}$.*

Proof. By Markov inequality, the probability that σ takes more than 2τ steps before reaching the goal state is at most $1/2$. Iterating this argument, we get that the probability that σ takes more than $2k\tau$ steps before reaching the goal state is at most 2^{-k} for every integer $k \geq 0$. In general, for any $m \geq 0$, the probability that σ takes more than m steps before reaching the goal state is at most $2^{-\lfloor \frac{m}{2\tau} \rfloor} \leq 2 \cdot 2^{-\frac{m}{2\tau}} \leq 2e^{-\frac{m}{4\tau}}$. □

Proof of Theorem 7.2.2. Define

$$X_k = \sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) - \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) \mid P, \sigma_k, s_1^k = s_{\text{init}}\Big].$$

This is a martingale difference sequence, and in order to use Theorem E.12.5 we need to show that $\Pr[|X_k| > m] \leq 2e^{-\frac{m}{4\tau}}$ for every $k = 1, 2, \ldots$ and $m \geq 0$. This follows immediately from Theorem E.6.1 since the total cost is bounded by the total time.

By Theorem E.12.5, $\sum_{k=1}^{K} X_k \leq 44\tau\sqrt{K \log^3 \frac{4K}{\delta}}$ with probability $1-\delta$, which gives the Lemma's statement. □

E.6.2 Proof of Theorem 7.2.3

Lemma E.6.2. *For every $k = 1, \ldots, K$ it holds that*

$$\mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) \mid P, \sigma_k, s_1^k = s_{init}\Big] \leq \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) \mid P, \pi_k, s_1^k = s_{init}\Big] = V_k^{\pi_k}(s_{init}).$$

Proof. Until a state $s \in \mathscr{S}$ with $T^{\pi_k}(s) \geq D/c_{\min}$ is reached, the strategy σ_k is the same as the policy π_k. If such a state is reached then $V^{\pi_k}(s) \geq c_{\min} T^{\pi_k}(s) \geq c_{\min} \frac{D}{c_{\min}} = D$, where the first inequality is because all costs are bounded from below by $c_{\min}$. On the other hand, $V^{\pi^f}(s) \leq T^{\pi^f}(s) \leq D$, where the last inequality follows by the definition of the fast policy and the SSP-diameter. Therefore, $V^{\pi^f}(s) \leq V^{\pi_k}(s)$. □

Lemma E.6.3. *For every $k = 1,\ldots,K$, the strategy σ_k of the learner ensures that the expected time to the goal state from any initial state is at most D/c_{min}.*

Proof. Let $s \in \mathscr{S}$. If $T^{\pi_k}(s) \geq D/c_{\min}$, then we play the fast policy π^f when we start in s. Thus, the expected time to the goal when starting in s will be at most D.

If $T^{\pi_k}(s) < D/c_{\min}$, then the expected time to the goal when starting in s will also be at most $D/c_{\min}$ since playing σ_k only decreases the expected time. □

Proof of Theorem 7.2.3. We decompose the regret into two terms as follows,

$$\begin{aligned} R_K &= \sum_{k=1}^{K}\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) - \sum_{k=1}^{K} \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) \mid P,\sigma_k,s_1^k = s_{\text{init}}\Big] \\ &\quad + \sum_{k=1}^{K} \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) \mid P,\sigma_k,s_1^k = s_{\text{init}}\Big] - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}). \end{aligned}$$

The first term accounts for the deviations in the performance of the learner's strategies from their expected value, and is bounded with high probability using Theorem 7.2.2.

The second term is the difference between the expected performance of the learner's strategies and the best policy in hindsight. Using Theorem E.6.2, we can bound it as follows,

$$\begin{aligned} \sum_{k=1}^{K} \mathbb{E}\Big[\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) \mid P,\sigma_k,s_1^k = s_{\text{init}}\Big] - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}) &\leq \sum_{k=1}^{K} V_k^{\pi_k}(s_{\text{init}}) - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}) \\ &= \sum_{k=1}^{K} \langle q^{\pi_k} - q^{\pi^\star}, c^k\rangle \\ &\leq \frac{2D}{c_{\min}}\sqrt{3K\log\frac{DSA}{c_{\min}}}, \end{aligned}$$

where the last inequality follows from Equation (E.6), and the equality follows because

$$V_k^{\pi}(s_{\text{init}}) = \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}} q^{\pi}(s,a)c^k(s,a) = \langle q^{\pi}, c^k\rangle.$$

□

E.7 Implementation details for SSP-O-REPS3

E.7.1 Computing q_k

After extending the occupancy measures, we must extend our additional definitions. Define $\mathrm{KL}(q \parallel q')$ as the unnormalized Kullback-Leibler divergence between two occupancy measures q and q':

$$\mathrm{KL}(q \parallel q') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \frac{q(s,a,s')}{q'(s,a,s')} + q'(s,a,s') - q(s,a,s'),$$

where $\mathcal{S}^+ = \mathcal{S} \cup \{g\}$. Furthermore, let $R(q)$ define the unnormalized negative entropy of the occupancy measure q:

$$R(q) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log q(s,a,s') - q(s,a,s').$$

SSP-O-REPS3 chooses its occupancy measures as follows:

$$\begin{aligned} q_1 = q^{P_1,\pi_1} &= \arg \min_{q\in\widetilde{\Delta(\mathcal{M})}_{e(1)}(D/c_{\min})} R(q) \\ q_{k+1} = q^{P_{k+1},\pi_{k+1}} &= \arg \min_{q\in\widetilde{\Delta(\mathcal{M})}_{e(k+1)}(D/c_{\min})} \eta \langle q, c^k \rangle + \mathrm{KL}(q \parallel q_k). \end{aligned}$$

As shown in [RM19a], each of these steps can be split into an unconstrained minimization step, and a projection step. Thus, q_1 can be computed as follows:

$$\begin{aligned} q_1' &= \arg\min_q R(q) \\ q_1 &= \arg \min_{q\in\widetilde{\Delta(\mathcal{M})}_{e(1)}(D/c_{\min})} \mathrm{KL}(q \parallel q_1'), \end{aligned}$$

where q_1' has a closed-from solution $q_1'(s,a,s') = 1$ for every $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}^+$. Similarly, q_{k+1} is computed as follows for every $k = 1,\dots,K-1$:

$$\begin{aligned} q_{k+1}' &= \arg\min_q \eta \langle q, c^k \rangle + \mathrm{KL}(q \parallel q_k) \\ q_{k+1} &= \arg \min_{q\in\widetilde{\Delta(\mathcal{M})}_{e(k+1)}(D/c_{\min})} \mathrm{KL}(q \parallel q_{k+1}'), \end{aligned}$$

where again q_{k+1}' has a closed-from solution $q_{k+1}'(s,a,s') = q_k(s,a,s')e^{-\eta c^k(s,a)}$ for every

$(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$.

Therefore, we just need to show that the projection step can be computed efficiently (the implementation follows [RM19a, JJL$^+$20]). We start by formulating the projection step as a constrained convex optimization problem (where $e = e(k+1)$):

$$\begin{aligned}
\min_{q} \quad & \mathrm{KL}(q \parallel q'_{k+1}) \\
s.t. \quad & \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} q(s,a,s') - \sum_{s'\in\mathscr{S}} \sum_{a'\in\mathscr{A}} q(s',a',s) = \mathbb{I}\{s = s_{\text{init}}\} && \forall s \in \mathscr{S} \\
& q(s,a,s') \le \left(\bar{P}_e(s' \mid s,a) + \varepsilon_e(s' \mid s,a)\right) \sum_{s''\in\mathscr{S}^+} q(s,a,s'') && \forall (s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+ \\
& q(s,a,s') \ge \left(\bar{P}_e(s' \mid s,a) - \varepsilon_e(s' \mid s,a)\right) \sum_{s''\in\mathscr{S}^+} q(s,a,s'') && \forall (s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+ \\
& \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} q(s,a,s') \le \frac{D}{c_{\min}} \\
& q(s,a,s') \ge 0 && \forall (s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+
\end{aligned}$$

To solve the problem, consider the Lagrangian:

$$\begin{aligned}
\mathscr{L}(q,\lambda,v,\mu) &= \mathrm{KL}(q \parallel q'_{k+1}) + \lambda \left(\sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} q(s,a,s') - \frac{D}{c_{\min}} \right) \\
&\quad + \sum_{s\in\mathscr{S}} v(s) \left(\sum_{s'\in\mathscr{S}} \sum_{a'\in\mathscr{A}} q(s',a',s) + \mathbb{I}\{s = s_{\text{init}}\} - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} q(s,a,s') \right) \\
&\quad + \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} \mu^+(s,a,s') \left(q(s,a,s') - \left(\bar{P}_e(s' \mid s,a) + \varepsilon_e(s' \mid s,a)\right) \sum_{s''\in\mathscr{S}^+} q(s,a,s'') \right) \\
&\quad + \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} \mu^-(s,a,s') \left(\left(\bar{P}_e(s' \mid s,a) - \varepsilon_e(s' \mid s,a)\right) \sum_{s''\in\mathscr{S}^+} q(s,a,s'') - q(s,a,s') \right) \\
&= \mathrm{KL}(q \parallel q'_{k+1}) + v(s_{\text{init}}) - \lambda \frac{D}{c_{\min}} \\
&\quad + \sum_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} q(s,a,s') \Bigg(\lambda + v(s') - v(s) + \mu^+(s,a,s') - \mu^-(s,a,s') \\
&\qquad\qquad - \sum_{s''\in\mathscr{S}^+} \bar{P}_e(s'' \mid s,a)(\mu^+(s,a,s'') - \mu^-(s,a,s'')) \\
&\qquad\qquad - \sum_{s''\in\mathscr{S}^+} \varepsilon_e(s'' \mid s,a)(\mu^+(s,a,s'') + \mu^-(s,a,s'')) \Bigg)
\end{aligned}$$

where λ, $\{v(s)\}_{s\in\mathscr{S}}$, $\{\mu^+(s,a,s')\}_{(s,a,s')\in\mathscr{S}\times\mathscr{A}\times\mathscr{S}^+}$ and $\{\mu^-(s,a,s')\}_{(s,a,s')\in\mathscr{S}\times\mathscr{A}\times\mathscr{S}^+}$ are

Lagrange multipliers, and we set $v(g) = 0$ for convenience. Differentiating the Lagrangian with respect to any $q(s,a,s')$, we get

$$\begin{aligned}\frac{\partial \mathscr{L}(q,\lambda,v,\mu)}{\partial q(s,a,s')} = \log \frac{q(s,a,s')}{q'_{k+1}(s,a,s')} + \lambda + v(s') - v(s) + \mu^+(s,a,s') - \mu^-(s,a,s') \\ - \sum_{s''\in\mathscr{S}^+} \bar{P}_e(s'' \mid s,a)(\mu^+(s,a,s'') - \mu^-(s,a,s'')) \\ - \sum_{s''\in\mathscr{S}^+} \varepsilon_e(s'' \mid s,a)(\mu^+(s,a,s'') + \mu^-(s,a,s'')).\end{aligned}$$

Next we define

$$\begin{aligned}B_k^{v,\mu}(s,a,s') = v(s) - v(s') + \mu^-(s,a,s') - \mu^+(s,a,s') - \eta c^k(s,a) \\ + \sum_{s''\in\mathscr{S}^+} \bar{P}_{e(k+1)}(s'' \mid s,a)(\mu^+(s,a,s'') - \mu^-(s,a,s'')) \\ + \sum_{s''\in\mathscr{S}^+} \varepsilon_{e(k+1)}(s'' \mid s,a)(\mu^+(s,a,s'') + \mu^-(s,a,s'')).\end{aligned} \tag{E.7}$$

Hence, setting the gradient to zero, we obtain the formula for $q_{k+1}(s,a)$:

$$\begin{aligned}q_{k+1}(s,a,s') &= q'_{k+1}(s,a,s')e^{-\lambda+\eta c^k(s,a)+B_k^{v,\mu}(s,a,s')} \\ &= q_k(s,a,s')e^{-\lambda+B_k^{v,\mu}(s,a,s')},\end{aligned} \tag{E.8}$$

where the last equality follows from the formula of $q'_{k+1}(s,a,s')$, and setting $c_0(s,a) = 0$ and $q_0(s,a,s') = 1$ for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$.

We now need to compute the value of λ, v, μ at the optimum. To that end, we write the dual problem $\mathscr{D}(\lambda,v,\mu) = \min_q \mathscr{L}(q,\lambda,v,\mu)$ by substituting q_{k+1} back into $\mathscr{L}$:

$$\begin{aligned}\mathscr{D}(\lambda,v,\mu) &= \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q'_{k+1}(s,a,s') - \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q_{k+1}(s,a,s') + v(s_{\text{init}}) - \lambda\frac{D}{c_{\min}} \\ &= -\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q_k(s,a,s')e^{-\lambda+B_k^{v,\mu}(s,a,s')} + v(s_{\text{init}}) - \lambda\frac{D}{c_{\min}} + \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q'_{k+1}(s,a,s').\end{aligned}$$

Now we obtain λ, v, μ by maximizing the dual. Equivalently, we can minimize the negation of the dual (and ignore the term $\sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q'_{k+1}(s,a,s')$), that is:

$$\lambda_{k+1}, v_{k+1}, \mu_{k+1} = \arg\min_{\lambda\geq 0, v, \mu\geq 0} \sum_{s\in\mathscr{S}}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+} q_k(s,a,s')e^{-\lambda+B_k^{v,\mu}(s,a,s')} + \lambda\frac{D}{c_{\min}} - v(s_{\text{init}}).$$

This is a convex optimization problem with only non-negativity constraints (and no constraints about the relations between the variables), which can be solved efficiently using iterative methods like gradient descent.

E.7.2 Computing the optimistic fast policy

The optimistic fast policy $\widetilde{\pi}_e^f$ is a deterministic stationary policy that together with the optimistic fast transition function from the confidence set of epoch e, minimizes the time to the goal state from all states simultaneously out of all pairs of policies and transition functions from the confidence set. Essentially, this is the optimal pair of policy and transition function from the confidence set w.r.t the constant cost function $c(s,a) = 1$ for every $s \in \mathscr{S}$ and $a \in \mathscr{A}$.

The existence of the optimistic fast policy is proven in [TGV$^+$20], and there they also show that it can be computed efficiently with Extended Value Iteration. In [RCMK20], the authors compute the following optimistic fast transition function for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}$:

$$\widetilde{P}_e^f(s' \mid s,a) = \max\left\{0, \bar{P}_e(s' \mid s,a) - 28A^e(s,a) - 4\sqrt{\bar{P}_e(s' \mid s,a)A^e(s,a)}\right\},$$

where the remaining probability mass goes to $\widetilde{P}_e^f(g \mid s,a)$. Then, $\widetilde{\pi}_e^f$ is computed by finding the fast policy w.r.t $\widetilde{P}_e^f$ (see Section E.2.2).

While this method is simpler and more efficient than Extended Value Iteration, the authors do not prove that this is indeed the optimistic fast policy. However, this policy is sufficient for their analysis and for our analysis as well. For simplicity, throughout the analysis we assume that $\widetilde{\pi}_e^f$ is the optimistic fast policy, but every step of the proof works with this computation as well.

E.8 Pseudo-code for SSP-O-REPS3

Algorithm 18 SSP-O-REPS3

```
Input: state space $\mathscr{S}$, action space $\mathscr{A}$, minimal cost $c_{\min}$, optimization parameter $\eta$ and confidence parameter $\delta$.
Initialization:
Obtain SSP-diameter $D$ from user or estimate it (see Section E.10).
Set $q_0(s,a,s') = 1$ and $c_0(s,a) = 0$ for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$.
Set $e \leftarrow 0$ and $\forall (s,a,s')$: $N^0(s,a) \leftarrow 0, N^0(s,a,s') \leftarrow 0, n^0(s,a) \leftarrow 0, n^0(s,a,s') \leftarrow 0$.
for $k = 1,2,\ldots$ do
    $e \leftarrow e+1$, start new epoch (Algorithm 19), set $s_1^k \leftarrow s_{\text{init}}$, $i \leftarrow 1$.
    while $s_i^k \neq g$ and $\widetilde{T}_k^{\pi_k}(s_i^k) < \frac{D}{c_{\min}}$ and $\forall a. n^e(s_i^k,a) + N^e(s_i^k,a) > \alpha \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$ do
        Play according to $\pi_k$, i.e., $a_i^k \sim \pi_k(\cdot \mid s_i^k)$, and observe next state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$.
        Update counters: $n^e(s_i^k,a_i^k) \leftarrow n^e(s_i^k,a_i^k) + 1, n^e(s_i^k,a_i^k,s_{i+1}^k) \leftarrow n^e(s_i^k,a_i^k,s_{i+1}^k) + 1$.
        $i \leftarrow i+1$.
        if $n^e(s_{i-1}^k,a_{i-1}^k) \geq N^e(s_{i-1}^k,a_{i-1}^k)$ then
            $e \leftarrow e+1$, start new epoch (Algorithm 19), and BREAK.
        end if
    end while
    while $s_i^k \neq g$ do
        if $\exists a \in \mathscr{A}. n^e(s_i^k,a) + N^e(s_i^k,a) \leq \alpha \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$ then
            Play the least played action $a_i^k = \arg\min_{a \in \mathscr{A}} n^e(s_i^k,a) + N^e(s_i^k,a)$.
        else
            Play according to $\widetilde{\pi}_e^f$, i.e., $a_i^k \sim \widetilde{\pi}_e^f(\cdot \mid s_i^k)$.
        end if
        Observe next state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$.
        Update counters: $n^e(s_i^k,a_i^k) \leftarrow n^e(s_i^k,a_i^k) + 1, n^e(s_i^k,a_i^k,s_{i+1}^k) \leftarrow n^e(s_i^k,a_i^k,s_{i+1}^k) + 1$.
        $i \leftarrow i+1$.
        if $n^e(s_{i-1}^k,a_{i-1}^k) \geq N^e(s_{i-1}^k,a_{i-1}^k)$ then
            $e \leftarrow e+1$, start new epoch (Algorithm 19).
        end if
    end while
    Set $I^k \leftarrow i-1$, observe cost function $c^k$ and suffer cost $\sum_{j=1}^{I^k} c^k(s_j^k,a_j^k)$.
end for
```

Algorithm 19 START NEW EPOCH

```
Update counters for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$:
    $N^e(s,a) \leftarrow N^{e-1}(s,a) + n^{e-1}(s,a) \quad ; \quad n^e(s,a) \leftarrow 0$
    $N^e(s,a,s') \leftarrow N^{e-1}(s,a,s') + n^{e-1}(s,a,s') \quad ; \quad n^e(s,a,s') \leftarrow 0$
Update confidence set for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$:
    $\bar{P}_e(s' \mid s,a) = \frac{N^e(s,a,s')}{N^e_+(s,a)}$
    $\varepsilon_e(s' \mid s,a) = 4\sqrt{\bar{P}_e(s' \mid s,a) A^e(s,a)} + 28A^e(s,a),$
where $A^e(s,a) = \frac{\log(SAN^e_+(s,a)/\delta)}{N^e_+(s,a)}$.
if $e$ is the first epoch of episode $k$ then
    Compute $\lambda_k, \nu_k, \mu_k$ as follows (using, e.g., gradient descent):
        $\lambda_k, \nu_k, \mu_k = \arg\min_{\lambda \geq 0, \nu, \mu \geq 0} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_{k-1}(s,a,s') e^{-\lambda + B^{\nu,\mu}_{k-1}(s,a,s')} + \lambda \frac{D}{c_{\min}} - \nu(s_{\text{init}}),$
    where $B^{\nu,\mu}_k(s,a,s')$ is defined in Equation (E.7).
    Compute $q_k$ as follows for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$:
        $q_k(s,a,s') = q_{k-1}(s,a,s') e^{-\lambda_k + B^{\nu_k,\mu_k}_{k-1}(s,a,s')}.$
    Compute $\pi_k$ and $P_k$ as follows for every $(s,a,s') \in \mathscr{S} \times \mathscr{A} \times \mathscr{S}^+$:
        $\pi_k(a \mid s) = \frac{\sum_{s' \in \mathscr{S}^+} q_k(s,a,s')}{\sum_{a' \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_k(s,a',s')} \quad ; \quad P_k(s' \mid s,a) = \frac{q_k(s,a,s')}{\sum_{s'' \in \mathscr{S}^+} q_k(s,a,s'')}$
    Set $\widetilde{T}^{\pi_k}_k(s) \leftarrow \frac{D}{c_{\min}}$ for every $s \in \mathscr{S}$ such that $\sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_k(s,a,s') = 0$.
    Compute $\widetilde{T}^{\pi_k}_k$ by solving the following linear equations:
        $\widetilde{T}^{\pi_k}_k(s) = 1 + \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}} \pi_k(a|s) P_k(s'|s,a) \widetilde{T}^{\pi_k}_k(s') \quad \forall s \in \{s \in \mathscr{S} : \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_k(s,a,s') > 0\}.$
else
    Compute the optimistic fast policy $\widetilde{\pi}^f_e$ (see Section E.7.2).
end if
```

E.9 Analysis of SSP-O-REPS3 (proofs for Section 7.3)

E.9.1 Overview

Our analysis follows the framework of [RCMK20] for analyzing optimism in SSPs, but makes the crucial adaptations needed to handle the adversarial environment.

We have two objectives: bounding the number of steps T taken by the algorithm (to show that we reach the goal in every episode) and bounding the regret. To bound the total time we split the time steps into *intervals*. The first interval begins at the first time step, and an interval ends once (1) an episode ends, (2) an epoch ends, (3) an unknown state is reached, or (4) a state s such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$ is reached when playing π_k in episode k, i.e., there is a switch.

Intuitively, we will bound the length of every interval by $\widetilde{O}(D/c_{\min})$ with high probability, and then use the number of intervals $\widetilde{O}(K + DS^2A/c_{\min}^2)$ to bound the total time T. Finally, we will show that the regret scales with the square root of the total variance (which is the number of intervals times the variance in each interval) to finish the proof. While intuitive, this approach is technically difficult and therefore we apply these principles in a different way.

We start by showing that the confidence sets contain P with high probability, which is a common result (see, e.g., [ZB19, EMGM19]). Define Ω^m the event that P is in the confidence set of the epoch that interval m belongs to.

Lemma E.9.1 ([RCMK20], Lemma 4.2)**.** *With probability at least* $1 - \delta/2$*, the event* Ω^m *holds for all intervals m simultaneously.*

There are two dependant probabilistic events that are important for the analysis. The first are the events Ω^m, and the second is that the deviation in the cost of a given policy from its expected value is not large. To disentangle these events we define an alternative regret for every $M = 1, 2, \dots,$

$$\widetilde{R}_M = \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{a \in \mathcal{A}} \tilde{\pi}_m(a \mid s_h^m) c^m(s_h^m, a) \mathbb{I}\{\Omega^m\} - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}),$$

where $c^m = c^k$ for the episode k that interval m belongs to, $\tilde{\pi}_m$ is the policy followed by the learner in interval m, H^m is the length of interval m, and the trajectory visited in interval m is $U^m = (s_1^m, a_1^m, \dots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$.

We focus on bounding $\widetilde{R}_M$ because we can use it to obtain a bound on R_K. This is done using Theorem E.9.1 and an application of Azuma inequality, when M is the number of intervals in which the first K episodes elapse (we show that the learner indeed completes these K episodes).

As mentioned, bounding the length of each interval complicates the analysis, and therefore we introduce artificial intervals. That is, an interval m also ends at the first time step H such that $\sum_{h=1}^{H}\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a \mid s_h^m)c^m(s_h^m,a) \geq D/c_{\min}$. The artificial intervals are only introduced for the analysis and do not affect the algorithm. Now, the length of each interval is bounded by $2D/c_{\min}^2$ and we can bound the number of intervals as follows.

Lemma E.9.2. *Let* $\widetilde{C}^M = \sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a \mid s_h^m)c^m(s_h^m,a)$. *The total time satisfies* $T \leq \widetilde{C}^M/c_{min}$ *and the total number of intervals satisfies*

$$M \leq \frac{c_{min}\widetilde{C}^M}{D} + 2SA\log T + 2K + 2\alpha\frac{DS^2A}{c_{min}^2}\log\frac{DSA}{\delta c_{min}}.$$

Note that a confidence set update occurs only in the end of an epoch and thus $\Omega^m = \Omega^{m-1}$ for most intervals. Also, for artificial intervals the policy does not change. Next we bound $\widetilde{C}^M$ as a function of the number of intervals M. Through summation of our confidence bounds, and by showing that the variance in each interval is bounded by $D^2/c_{\min}^2$ we are able to obtain the following, when Theorem E.9.1 holds,

$$\widetilde{C}^M \leq \sum_{k=1}^{K}\langle q_k, c^k\rangle + \widetilde{O}\left(\frac{DS}{c_{\min}}\sqrt{MA} + \frac{D^2S^2A}{c_{\min}^2}\right).$$

Substituting in Theorem E.9.2 and solving for $\widetilde{C}^M$ we get

$$\begin{aligned}\widetilde{R}_M = \widetilde{C}^M - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{\text{init}}) &\leq \sum_{k=1}^{K}\langle q_k - q^{P,\pi^\star}, c^k\rangle \\ &+ \widetilde{O}\left(\frac{DS}{c_{\min}}\sqrt{AK} + \frac{D^2S^2A}{c_{\min}^2}\right),\end{aligned}$$

Notice that the first term on the RHS of the inequality is exactly the regret of OMD, and therefore analyzing it similarly to Theorem 7.2.1 gives the final bound (see Section E.9.7).

E.9.2 Notations

Denote the trajectory visited in interval m by $U^m = (s_1^m, a_1^m, \ldots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$, where a_h^m is the action taken in s_h^m, and H^m is the length of the interval. In addition, the concatenation of trajectories in the intervals up to and including interval m is denoted by $\bar{U}^m$, that is $\bar{U}^m = \cup_{m'=1}^{m} U^{m'}$.

The policy that the learner follows in interval m is denoted by $\tilde{\pi}_m$, and the transition function that was involved in the choice of $\tilde{\pi}_m$ is denoted by $\widetilde{P}_m$. For the first interval of every episode these are chosen by OMD, i.e., π_k and P_k, and for other intervals these are the optimistic fast policy $\widetilde{\pi}_e^f$ and the transition function chosen from the confidence set together with it $\widetilde{P}_e^f$, for the epoch e that interval m belongs to. Notice that intervals with unknown states are of length 1. Thus, there is no policy since only one action is performed – we ignore visits to unknown states and we suffer their cost directly in Theorem E.9.5.

The expected cost of $\tilde{\pi}_m$ w.r.t $\widetilde{P}_m$ is denoted by $\widetilde{V}^m$, and the expected time to the goal is denoted by $\widetilde{T}^m$. For intervals in which we follow the optimistic fast policy, we will show that $\widetilde{T}^m(s) \le D$ for every $s \in \mathscr{S}$ when Ω^m holds. We would like to have a similar property for intervals in which we follow the OMD policy, i.e., the first interval of every episode.

Note that for the first interval m of episode k, we have that $\widetilde{T}_k^{\pi_k} = \widetilde{T}^m$, and recall that reaching a state $s \in \mathscr{S}$ such that $\widetilde{T}_k^{\pi_k}(s) \ge D/c_{\min}$ ends the current interval. We would like to take advantage of this fact in order to make sure that $\widetilde{T}^m$ is always bounded by $D/c_{\min}$. Similarly to Section 7.2.2, we compute $\widetilde{T}_k^{\pi_k}(s)$ only for states s that are reachable from s_{init} w.r.t P_k. Since reaching a state s with $\widetilde{T}_k^{\pi_k}(s) \ge D/c_{\min}$ yields the start of a new interval for which we use the optimistic fast policy, we can set $\widetilde{T}_k^{\pi_k}(s) = D/c_{\min}$ for states that are not reachable from s_{init} without affecting the algorithm's choices.

We make another change to $\widetilde{P}_m$ for interval m that is the first interval of episode k. Since reaching a state $s \in \mathscr{S}$ such that $\widetilde{T}_k^{\pi_k}(s) \ge D/c_{\min}$ ends the interval, we tweak $\widetilde{P}_m$ such that from such a state it goes directly to the goal with expected time of $D/c_{\min}$ and expected cost of D (can be done with a self-loop that has $c_{\min}/D$ probability to go to g). Thus, when we consider the expected cost of $\tilde{\pi}_m$ w.r.t $\widetilde{P}_m$, we have that $\widetilde{V}^m(s_{\text{init}}) \le \widetilde{V}_k^{\pi_k}(s_{\text{init}})$ because we only decreased the cost from some states. However, notice that now $\widetilde{P}_m$ is in the confidence set only for states that we did not tweak. We show that this does not affect the analysis, since reaching those states ends the interval.

We would like to emphasize that tweaking $\widetilde{P}_m$ is only done in hindsight as a part of the analysis, and does not change the algorithm.

E.9.3 Properties of the learner's policies

Lemma E.9.3. *Let m be an interval. If m is the first interval of episode k then $\widetilde{T}^m(s) \le D/c_{min}$ for every $s \in \mathscr{S}$. Otherwise, if Ω^m holds then $\widetilde{T}^m(s) \le D$ for every $s \in \mathscr{S}$.*

Proof. The first case holds by definition of $\widetilde{P}_m$ for intervals that are in the beginning of some episode (see discussion in Section E.9.2). The second case follows by optimism and the fact that P is in the confidence set (see [RCMK20], Lemma B.2). □

Lemma E.9.4. *Let m be an interval and let $1 \le h \le H^m$. If Ω^m holds then the following Bellman equations hold:*

$$\widetilde{V}^m(s_h^m) = \sum_{a\in\mathscr{A}} \tilde{\pi}_m(a \mid s_h^m) c^m(s_h^m, a) + \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{V}^m(s')$$
$$\widetilde{T}^m(s_h^m) = 1 + \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{T}^m(s').$$

Proof. For the optimistic fast policy $\tilde{\pi}_e^f$ the Bellman equations hold for every $s \in \mathscr{S}$ since it is proper w.r.t $\widetilde{P}_e^f$ (see [RCMK20], Lemma B.11). When $\tilde{\pi}_m$ is the policy chosen by OMD π_k, reaching a state s such that $q^{P_k,\pi_k}(s) = 0$ will end the interval (since we set $\widetilde{T}_k^{\pi_k}(s) = D/c_{\min}$ for these states). Thus, it suffices to show that the Bellman equations hold for all states in $\{s \in \mathscr{S} : q^{P_k,\pi_k}(s) > 0\}$.

For these states we have that $\widetilde{T}^m$ is bounded by $D/c_{\min}$ and therefore $\tilde{\pi}_m$ is proper w.r.t $\widetilde{P}_m$ and the Bellman equations hold. Note that we did not make changes to $\widetilde{P}_m$ or c^m in states that can be visited during the interval. □

E.9.4 Regret decomposition

Lemma E.9.5. *It holds that*

$$\widetilde{R}_M \le \sum_{m=1}^{M} \widetilde{R}_m^1 + \sum_{m=1}^{M} \widetilde{R}_m^2 - \sum_{k=1}^{K} V_k^{\pi^\star}(s_{init}) + \alpha \frac{DS^2A}{c_{min}^2} \log \frac{DSA}{\delta c_{min}},$$

where

$$\widetilde{R}_m^1 = \big(\widetilde{V}^m(s_1^m) - \widetilde{V}^m(s_{H^m+1}^m)\big)\mathbb{I}\{\Omega^m\}$$
$$\widetilde{R}_m^2 = \sum_{h=1}^{H^m} \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}.$$

Proof. First we have a cost of at most 1 every time we visit an unknown state. Each state becomes known after $\alpha A \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$ visits, and therefore the total cost from these visits is at most $\alpha SA \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$. From now on we will ignore visits to unknown states throughout the analysis because we calculated their contribution to the total cost.

We can use the Bellman equations w.r.t $\widetilde{P}_m$ (Theorem E.9.4) to have the following interpretation of the costs for every interval m and time h:

$$
\begin{aligned}
&\sum_{a\in\mathscr{A}} \tilde{\pi}_m(a \mid s_h^m)c^m(s_h^m,a)\mathbb{I}\{\Omega^m\} = \\
&\qquad= \left(\widetilde{V}^m(s_h^m) - \sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m)\widetilde{P}_m(s' \mid s_h^m,a)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\} \\
&\qquad= \left(\widetilde{V}^m(s_h^m) - \widetilde{V}^m(s_{h+1}^m)\right)\mathbb{I}\{\Omega^m\} \\
&\qquad\quad+ \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m)\widetilde{P}_m(s' \mid s_h^m,a)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\}.
\end{aligned}
\tag{E.9}
$$

We now write $\widetilde{R}_M = \sum_{m=1}^M \sum_{h=1}^{H^m} \sum_{a\in\mathscr{A}} \tilde{\pi}_m(a \mid s_h^m)c^m(s_h^m,a)\mathbb{I}\{\Omega^m\} - \sum_{k=1}^K V_k^{\pi^\star}(s_{\text{init}})$, and substitute for each cost using Equation (E.9) to get the lemma, noting that the first term telescopes within the interval. □

Lemma E.9.6. *It holds that*

$$
\sum_{m=1}^M \widetilde{R}_m^1 \leq 2DSA\log T + \alpha\frac{D^2S^2A}{c_{min}^2}\log\frac{DSA}{\delta c_{min}} + \sum_{k=1}^K \widetilde{V}_k^{\pi_k}(s_{init})\mathbb{I}\{\Omega^{m(k)}\},
$$

where $m(k)$ is the first interval of episode k.

Proof. For every two consecutive intervals $m, m+1$ we have one of the following:

1. If interval m ended in the goal state then $\widetilde{V}^m(s_{H^m+1}^m) = \widetilde{V}^m(g) = 0$ and $\widetilde{V}^{m+1}(s_1^{m+1}) = \widetilde{V}^{m(k)}(s_{\text{init}}) \leq \widetilde{V}_k^{\pi_k}(s_{\text{init}})$, where $m+1$ is the first interval of episode k. Therefore,

   $$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} \leq \widetilde{V}_k^{\pi_k}(s_{\text{init}})\mathbb{I}\{\Omega^{m(k)}\}.$$

   This happens at most K times, once for every value k.

2. If interval m ended since the sum of expected costs in the interval passed $D/c_{\min}$, then we did not change policy. Thus, $\widetilde{V}^m = \widetilde{V}^{m+1}$, $\Omega^m = \Omega^{m+1}$ and $s_1^{m+1} = s_{H^m+1}^m$. We get

   $$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} = 0.$$

3. If interval m ended by reaching an unknown state, then we switch policy. Thus,

$$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} \le \widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} \le D,$$

   where the last inequality follows because we switched to the optimistic fast policy and thus its expected time will be bounded by D if P is in the confidence set (see Theorem E.9.3). This happens at most $SA\alpha\frac{DS}{c_{\min}^2}\log\frac{DSA}{\delta c_{\min}}$ times.

   Here we ignored the unknown state (since we accounted for its cost in Theorem E.9.5) and jumped right to the next interval, which is controlled by the optimistic fast policy.

4. If interval m ended with doubling the visits to some state-action pair, then similarly to the previous article,

$$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} \le \widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} \le D.$$

   This happens at most $2SA\log T$.

5. If m is the first interval of an episode k and it ended because we reached a "bad" state then $\widetilde{V}^m(s_{H^m+1}^m) = D$ and $\widetilde{V}^{m+1}(s_1^{m+1}) \le D$ since this is the optimistic fast policy. Thus,

$$\widetilde{V}^{m+1}(s_1^{m+1})\mathbb{I}\{\Omega^{m+1}\} - \widetilde{V}^m(s_{H^m+1}^m)\mathbb{I}\{\Omega^m\} \le 0.$$

□

Lemma E.9.7. *With probability at least* $1-\delta/6$*, the following holds for all* $M = 1,2,\dots$ *simultaneously.*

$$\sum_{m=1}^{M} \widetilde{R}_m^2 \le \sum_{m=1}^{M} \mathbb{E}\big[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\big] + \frac{6D}{c_{min}}\sqrt{M\log\frac{4M}{\delta}},$$

where $\mathbb{E}[\cdot \mid \bar{U}^{m-1}]$ *is the expectation conditioned on the trajectories up to interval m.*

Proof. Consider the martingale difference sequence $(Y^m)_{m=1}^{\infty}$ defined by $Y^m = X^m - \mathbb{E}[X^m \mid \bar{U}^{m-1}]$ and

$$X^m = \sum_{h=1}^{H^m}\left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \tilde{\pi}_m(a \mid s_h^m)\widetilde{P}_m(s' \mid s_h^m, a)\widetilde{V}^m(s')\right)\mathbb{I}\{\Omega^m\}.$$

The Bellman equations of $\tilde{\pi}_m$ w.r.t $\widetilde{P}_m$ (Theorem E.9.4) obtain

$$|X^m| = \Bigg| \Bigg(\underbrace{\widetilde{V}^m(s^m_{H^m+1}) - \widetilde{V}^m(s^m_1)}_{\le D/c_{\min}} + \\ + \underbrace{\sum_{h=1}^{H^m} \widetilde{V}^m(s^m_h) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s^m_h)\widetilde{P}_m(s' \mid s^m_h, a)\widetilde{V}^m(s')}_{=\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a|s^m_h)c^m(s^m_h,a)} \Bigg) \mathbb{I}\{\Omega^m\} \Bigg| \\ \le \frac{D}{c_{\min}} + \sum_{h=1}^{H^m} \sum_{a\in\mathscr{A}} \tilde{\pi}_m(a \mid s^m_h) c^m(s^m_h, a) \le \frac{3D}{c_{\min}}$$

where for the first inequality we used Theorems E.9.3 and E.9.4, and the last inequality follows because the cost in every interval is at most $2D/c_{\min}$.

Therefore, we use anytime Azuma inequality (Theorem E.12.1) to obtain that with probability at least $1-\delta/6$:

$$\sum_{m=1}^{M} X^m \le \sum_{m=1}^{M} \mathbb{E}\big[X^m \mid \bar{U}^{m-1}\big] + \frac{6D}{c_{\min}}\sqrt{M \log\frac{4M}{\delta}}.$$

□

E.9.5 Bounding the variance within an interval

Lemma E.9.8 ([RCMK20], Lemma B.13)**.** *Denote $A^m(s,a) = \frac{\log(SAN_+^{e(m)}(s,a)/\delta)}{N_+^{e(m)}(s,a)}$, where $e(m)$ is the epoch that interval m belongs to. When Ω^m holds we have for any $(s,a,s') \in \mathscr{S}\times\mathscr{A}\times\mathscr{S}^+$:*

$$\big|P(s' \mid s,a) - \widetilde{P}_m(s' \mid s,a)\big| \le 8\sqrt{P(s' \mid s,a)A^m(s,a)} + 136A^m(s,a).$$

Lemma E.9.9. *Denote $A^m_h = A^m(s^m_h, a^m_h)$. For every interval m it holds that,*

$$\mathbb{E}[\widetilde{R}^2_m \mid \bar{U}^{m-1}] \le 16\mathbb{E}\left[\sum_{h=1}^{H^m}\sqrt{S\mathbb{V}^m_h A^m_h}\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] + 272\mathbb{E}\left[\sum_{h=1}^{H^m}\frac{D}{c_{min}}SA^m_h\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right],$$

where $\mathbb{V}^m_h$ is the empirical variance defined as

$$\mathbb{V}^m_h = \sum_{s'\in\mathscr{S}^+} P(s' \mid s^m_h, a^m_h)\left(\widetilde{V}^m(s') - \mu^m_h\right)^2,$$

and $\mu_h^m = \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} \tilde{\pi}_m(a \mid s_h^m) P(s' \mid s_h^m, a) \widetilde{V}^m(s')$.

Proof. Denote

$$X^m = \sum_{h=1}^{H^m} \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}$$

$$Z_h^m = \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) P(s' \mid s_h^m, a) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\}.$$

Think of the interval as an infinite stochastic process, and note that, conditioned on $\bar{U}^{m-1}$, $\left(Z_h^m\right)_{h=1}^{\infty}$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^{\infty}$, where U^h is the trajectory of the learner from the beginning of the interval and up to and including time h. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval.

Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^{\infty}$ which is bounded by $2D/c_{\min}^2$. Hence by the optional stopping theorem $\mathbb{E}[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1}] = 0$, which gets us

$$\begin{aligned}
&\mathbb{E}[X^m \mid \bar{U}^{m-1}] = \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m} \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{V}^m(s') \right) \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1} \right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m} Z_h^m \mid \bar{U}^{m-1} \right] + \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \big(P(s' \mid s_h^m, a) - \widetilde{P}_m(s' \mid s_h^m, a) \big) \tilde{\pi}_m(a \mid s_h^m) \widetilde{V}^m(s') \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1} \right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H^m} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \big(P(s' \mid s_h^m, a) - \widetilde{P}_m(s' \mid s_h^m, a) \big) \tilde{\pi}_m(a \mid s_h^m) \widetilde{V}^m(s') \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1} \right].
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}}\big(P(s'\mid s_h^m,a)-\widetilde{P}_m(s'\mid s_h^m,a)\big)\tilde{\pi}_m(a\mid s_h^m)\widetilde{V}^m(s')\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] = \\
&\quad = \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+}\big(P(s'\mid s_h^m,a)-\widetilde{P}_m(s'\mid s_h^m,a)\big)\tilde{\pi}_m(a\mid s_h^m)\widetilde{V}^m(s')\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] \\
&\quad = \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s'\in\mathscr{S}^+}\big(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\big)\widetilde{V}^m(s')\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] \\
&\quad = \mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{s'\in\mathscr{S}^+}\big(P(s'\mid s_h^m,a_h^m)-\widetilde{P}_m(s'\mid s_h^m,a_h^m)\big)\left(\widetilde{V}^m(s')-\mu_h^m\right)\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] \\
&\quad \leq \mathbb{E}\left[8\sum_{h=1}^{H^m}\sum_{s'\in\mathscr{S}^+}\sqrt{A_h^m P(s'\mid s_h^m,a_h^m)\left(\widetilde{V}^m(s')-\mu_h^m\right)^2}\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] \\
&\qquad + \mathbb{E}\left[136\sum_{h=1}^{H^m}\sum_{s'\in\mathscr{S}^+}A_h^m\left|\widetilde{V}^m(s')-\mu_h^m\right|\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right] \\
&\quad \leq \mathbb{E}\left[16\sum_{h=1}^{H^m}\sqrt{S\mathbb{V}_h^m A_h^m}\mathbb{1}\{\Omega^m\}+272S\frac{D}{c_{\min}}A_h^m\mathbb{1}\{\Omega^m\}\mid \bar{U}^{m-1}\right],
\end{aligned}$$

where the first equality follows because $\widetilde{V}^m(g)=0$ and the second by the definition of a_h^m. The third equality follows since $P(\cdot\mid s_h^m,a_h^m)$ and $\widetilde{P}_m(\cdot\mid s_h^m,a_h^m)$ are probability distributions over S^+ whence μ_h^m does not depend on s'. The first inequality follows from Theorem E.9.8, and the second inequality from Jensen's inequality, Theorem E.9.3, $|S^+|\leq 2S$, and the definition of $\mathbb{V}_h^m$. □

The following lemma will help us bound the variance within an interval, and it follows by the fact that known states were visited many times so our estimation of the transition function in these states is relatively accurate.

Lemma E.9.10 ([RCMK20], Lemma B.14)**.** *Let m be an interval and s be a known state. If Ω^m holds then for every $a\in\mathscr{A}$ and $s'\in\mathscr{S}^+$,*

$$\big|\widetilde{P}_m\big(s'\mid s,a\big)-P\big(s'\mid s,a\big)\big|\leq\frac{1}{8}\sqrt{\frac{c_{min}^2\cdot P\big(s'\mid s,a\big)}{SD}}+\frac{c_{min}^2}{4SD}.$$

Define $\mu^m(s)=\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+}\tilde{\pi}_m(a\mid s)P(s'\mid s,a)\widetilde{V}^m(s')$ and therefore $\mu_h^m=\mu^m(s_h^m)$. Similarly, define $\mathbb{V}^m(s,a)=\sum_{s'\in\mathscr{S}^+}P(s'\mid s,a)\left(\widetilde{V}^m(s')-\mu^m(s)\right)^2$ and therefore $\mathbb{V}_h^m=$

$\mathbb{V}^m(s_h^m, a_h^m)$. The next lemma bounds the variance within a single interval.

Lemma E.9.11. *For any interval m it holds that $\mathbb{E}\big[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \leq 64D^2/c_{min}^2$.*

Proof. Denote

$$Z_h^m = \left(\widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) P(s' \mid s_h^m, a) \widetilde{V}^m(s')\right) \mathbb{I}\{\Omega^m\},$$

and think of the interval as an infinite stochastic process. Note that, conditioned on $\bar{U}^{m-1}$, $\big(Z_h^m\big)_{h=1}^{\infty}$ is a martingale difference sequence w.r.t $(U^h)_{h=1}^{\infty}$, where U^h is the trajectory of the learner from the beginning of the interval and up to time h and including. This holds since, by conditioning on $\bar{U}^{m-1}$, Ω^m is determined and is independent of the randomness generated during the interval. Note that H^m is a stopping time with respect to $(Z_h^m)_{h=1}^{\infty}$ which is bounded by $2D/c_{\min}^2$. Therefore, applying Theorem E.12.2 obtains

$$\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] = \mathbb{E}\left[\left(\sum_{h=1}^{H^m} Z_h^m \mathbb{I}\{\Omega^m\}\right)^2 \mid \bar{U}^{m-1}\right]. \tag{E.10}$$

We now proceed by bounding $|\sum_{h=1}^{H^m} Z_h^m|$ when Ω^m occurs. Therefore,

$$\begin{aligned}
\left|\sum_{h=1}^{H^m} Z_h^m\right| &= \left|\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) P(s' \mid s_h^m, a) \widetilde{V}^m(s')\right| \\
&\leq \left|\sum_{h=1}^{H^m} \widetilde{V}^m(s_{h+1}^m) - \widetilde{V}^m(s_h^m)\right| \qquad \text{(E.11)} \\
&\quad + \left|\sum_{h=1}^{H^m} \widetilde{V}^m(s_h^m) - \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}} \tilde{\pi}_m(a \mid s_h^m) \widetilde{P}_m(s' \mid s_h^m, a) \widetilde{V}^m(s')\right| \qquad \text{(E.12)} \\
&\quad + \left|\sum_{h=1}^{H^m} \sum_{a\in\mathscr{A}} \sum_{s'\in\mathscr{S}^+} \tilde{\pi}_m(a \mid s_h^m) \Big(\widetilde{P}_m(s' \mid s_h^m, a) - P(s' \mid s_h^m, a)\Big) \Big(\widetilde{V}^m(s') - \mu_h^m\Big)\right|, \qquad \text{(E.13)}
\end{aligned}$$

where Equation (E.13) is given as $P(\cdot \mid s_h^m, a)$ and $\widetilde{P}_m(\cdot \mid s_h^m, a)$ are probability distributions over S^+, μ_h^m is constant w.r.t s', and $\widetilde{V}^m(g) = 0$.

We now bound each of the three terms above individually. Equation (E.11) is a telescopic sum that is at most $D/c_{\min}$ on Ω^m (Theorem E.9.3). For Equation (E.12), we use the Bellman equations for $\tilde{\pi}_m$ w.r.t $\widetilde{P}_m$ (Theorem E.9.4) thus it is at most $2D/c_{\min}$ (see proof of Theorem E.9.7). For Equation (E.13), recall that all states at times $h = 1, \ldots, H^m$ are known by definition of H^m. Hence by Theorem E.9.10,

$$\begin{aligned}\left|\sum_{s'\in\mathscr{S}^+}\left(P(s'\mid s_h^m,a)-\widetilde{P}_m(s'\mid s_h^m,a)\right)\left(\widetilde{V}^m(s')-\mu_h^m\right)\right| &\leq \frac{1}{8}\sum_{s'\in\mathscr{S}^+}\sqrt{\frac{c_{\min}^2 P(s'\mid s_h^m,a)\left(\widetilde{V}^m(s')-\mu_h^m\right)^2}{SD}}\\ &\quad+\sum_{s'\in\mathscr{S}^+}\frac{c_{\min}^2}{4SD}\underbrace{\left|\widetilde{V}^m(s')-\mu_h^m\right|}_{\leq D/c_{\min}}\\ &\leq \frac{1}{4}\sqrt{\frac{c_{\min}^2\mathbb{V}^m(s_h^m,a)}{D}}+\frac{c_{\min}}{2},\end{aligned}$$

where the last inequality follows from Jensen's inequality and because $|S^+|\leq 2S$. Therefore,

$$\begin{aligned}&\left|\sum_{a\in\mathscr{A}}\sum_{s'\in\mathscr{S}^+}\tilde{\pi}_m(a\mid s_h^m)\left(P(s'\mid s_h^m,a)-\widetilde{P}_m(s'\mid s_h^m,a)\right)\left(\widetilde{V}^m(s')-\mu_h^m\right)\right|\leq\\ &\quad\leq\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a\mid s_h^m)\left(\frac{1}{4}\sqrt{\frac{c_{\min}^2\mathbb{V}^m(s_h^m,a)}{D}}+\frac{c_{\min}}{2}\right)\\ &\quad\leq\frac{1}{4}\sqrt{\frac{c_{\min}^2\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a\mid s_h^m)\mathbb{V}^m(s_h^m,a)}{D}}+\frac{c_{\min}}{2},\end{aligned}$$

where the last inequality follows again from Jensen's inequality. We use Jensen's inequality one last time to obtain

$$\begin{aligned}&\sum_{h=1}^{H^m}\frac{1}{4}\sqrt{\frac{c_{\min}^2\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a\mid s_h^m)\mathbb{V}^m(s_h^m,a)}{D}}+\sum_{h=1}^{H^m}\frac{c_{\min}}{2}\leq\\ &\quad\leq\frac{1}{4}\sqrt{H^m\sum_{h=1}^{H^m}\frac{c_{\min}^2\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a\mid s_h^m)\mathbb{V}^m(s_h^m,a)}{D}}+\frac{c_{\min}H^m}{2}\\ &\quad\leq\frac{1}{2}\sqrt{\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}}\tilde{\pi}_m(a\mid s_h^m)\mathbb{V}^m(s_h^m,a)}+\frac{D}{c_{\min}},\end{aligned}$$

where we used the fact that $H^m\leq 2D/c_{\min}^2$.

Plugging these bounds back into Equation (E.10) gets us

$$\begin{aligned}\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] &\le \mathbb{E}\left[\left(\frac{4D}{c_{\min}} + \frac{1}{2}\sqrt{\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}} \tilde{\pi}_m(a|s_h^m)\mathbb{V}^m(s_h^m,a)\mathbb{I}\{\Omega^m\}}\right)^2 \,\middle|\, \bar{U}^{m-1}\right] \\ &\le \frac{32D^2}{c_{\min}^2} + \frac{1}{2}\mathbb{E}\left[\sum_{h=1}^{H^m}\sum_{a\in\mathscr{A}} \tilde{\pi}_m(a|s_h^m)\mathbb{V}^m(s_h^m,a)\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] \\ &= \frac{32D^2}{c_{\min}^2} + \frac{1}{2}\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right],\end{aligned}$$

where the second inequality is by the elementary inequality $(a+b)^2 \le 2(a^2+b^2)$, and the last equality is by definition of a_h^m and $\mathbb{V}_h^m$. Rearranging gets us $\mathbb{E}\big[\sum_{h=2}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\big] \le 64D^2/c_{\min}^2$, and the lemma follows. □

Lemma E.9.12. *With probability at least* $1-\delta/6$, *the following holds for all* $M = 1, 2, \ldots$ *simultaneously.*

$$\sum_{m=1}^{M} \mathbb{E}[\widetilde{R}_m^2 \mid \bar{U}^{m-1}] \le 573\frac{DS}{c_{min}}\sqrt{MA\log^2\frac{TSA}{\delta}} + 5440\frac{D}{c_{min}}S^2A\log^2\frac{TSA}{\delta}.$$

Proof. From Theorem E.9.9 we have that

$$\mathbb{E}[\widetilde{R}_m^2 \mid \bar{U}^{m-1}] \le 16\mathbb{E}\left[\sum_{h=1}^{H^m}\sqrt{S\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] + 272\mathbb{E}\left[\sum_{h=1}^{H^m}\frac{D}{c_{\min}}SA_h^m\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right],$$

Moreover, by applying the Cauchy-Schwartz inequality twice, we get that

$$\begin{aligned}\mathbb{E}\left[\sum_{h=1}^{H^m}\sqrt{\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right] &\le \mathbb{E}\left[\sqrt{\sum_{h=1}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\}}\cdot\sqrt{\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}} \,\middle|\, \bar{U}^{m-1}\right] \\ &\le \sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]}\cdot\sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m}\mathbb{V}_h^m\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]} \\ &\le \frac{8D}{c_{\min}}\sqrt{\mathbb{E}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\} \,\middle|\, \bar{U}^{m-1}\right]},\end{aligned}$$

where the last inequality is by Theorem E.9.11. We sum over all intervals to obtain

$$\begin{aligned}\sum_{m=1}^{M}\mathbb{E}[\widetilde{R}_m^2 \mid \bar{U}^{m-1}] &\le \frac{128D}{c_{\min}}\sum_{m=1}^{M}\sqrt{S\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right]}+\frac{272DS}{c_{\min}}\sum_{m=1}^{M}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right]\\ &\le \frac{128D}{c_{\min}}\sqrt{MS\sum_{m=1}^{M}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right]}+\frac{272DS}{c_{\min}}\sum_{m=1}^{M}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}\;\middle|\;\bar{U}^{m-1}\right],\end{aligned}$$

where the last inequality follows from Jensen's inequality. We finish the proof using Theorem E.9.13 below. □

Lemma E.9.13. *With probability at least* $1-\delta/6$*, the following holds for* $M=1,2,\dots$ *simultaneously.*

$$\sum_{m=1}^{M}\mathbb{E}\left[\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}\mid\bar{U}^{m-1}\right]\le 20SA\log^2\frac{TSA}{\delta}.$$

Proof. Define the infinite sequence of random variables: $X^m=\sum_{h=1}^{H^m}A_h^m\mathbb{I}\{\Omega^m\}$ for which $|X^m|\le 2$ due to Theorem E.9.14 below. We apply Equation (E.22) of Theorem E.12.3 to obtain with probability at least $1-\delta/6$, for all $M=1,2,\dots$ simultaneously

$$\sum_{m=1}^{M}\mathbb{E}\big[X^m\mid\bar{U}^{m-1}\big]\le 2\sum_{m=1}^{M}X^m+8\log\frac{12M}{\delta}.$$

Now, we bound the sum over X^m by rewriting it as a sum over epochs (since the confidence sets update only in the beginning of a new epoch):

$$\sum_{m=1}^{M}X^m\le\sum_{m=1}^{M}\sum_{h=1}^{H^m}\frac{\log(SAN_+^{e(m)}(s_h^m,a_h^m)/\delta)}{N_+^{e(m)}(s_h^m,a_h^m)}\le\log\frac{SAT}{\delta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{e=1}^{E}\frac{n^e(s,a)}{N_+^e(s,a)},$$

where $n^e(s,a)$ is the number of visits to (s,a) during epoch e. From Theorem E.9.15 below we have that for every $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\sum_{e=1}^{E}\frac{n^e(s,a)}{N_+^e(s,a)}\le 2\log N_{E+1}(s,a)\le 2\log T.$$

We now plugin the resulting bound for $\sum_{m=1}^{M}X^m$ and simplify the acquired expression by using $M\le T$. □

Lemma E.9.14. *For any interval* m*,* $|\sum_{h=1}^{H^m}A_h^m|\le 2$.

Proof. Note that all states during the interval are known. Hence, $N_+^{e(m)}(s_h^m, a_h^m) \geq \alpha \cdot \frac{DS}{c_{\min}^2} \log \frac{DSA}{\delta c_{\min}}$. Therefore, since $\log(x)/x$ is decreasing and since $A \geq 2$ (otherwise the learner has no choices),

$$\sum_{h=1}^{H^m} A_h^m = \sum_{h=1}^{H^m} \frac{\log(SAN_+^{e(m)}(s_h^m, a_h^m)/\delta)}{N_+^{e(m)}(s_h^m, a_h^m)} \leq \frac{c_{\min}^2 H^m}{D} \leq 2.$$

□

Lemma E.9.15 ([RCMK20], Lemma B.18)**.** *For any sequence of integers $z_1, \dots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$ and $Z_0 = 1$, it holds that*

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq 2 \log Z_n.$$

E.9.6 Proof of Theorem 7.3.1

Theorem E.9.16 (Restatement of Theorem 7.3.1)**.** *Under Theorem 5.2.1, running SSP-O-REPS3 with known SSP-diameter D and $\eta = \sqrt{\frac{6\log(DSA/c_{min})}{K}}$ ensures that, with probability at least $1 - \delta$,*

$$R_K \leq O\left(\frac{DS}{c_{min}} \sqrt{AK} \log \frac{KDSA}{\delta c_{min}} + \frac{D^2 S^2 A}{c_{min}^2} \log^2 \frac{KDSA}{\delta c_{min}}\right) = \widetilde{O}\left(\frac{DS}{c_{min}} \sqrt{AK}\right),$$

where the last equality holds for $K \geq D^2 S^2 A / c_{min}^2$.

Proof of Theorem 7.3.1. With probability at least $1 - \delta$, via a union bound, we have that Theorems E.9.1, E.9.7 and E.9.12 hold and the following holds by Azuma inequality for every $T = 1, 2, \dots$ simultaneously,

$$\sum_{m=1}^{M} \sum_{h=1}^{H^m} c^m(s_h^m, a_h^m) \leq \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{a \in \mathscr{A}} \tilde{\pi}_m(a \mid s_h^m) c^m(s_h^m, a) + 4\sqrt{T \log \frac{T}{\delta}}. \tag{E.14}$$

We start by bounding $\widetilde{R}_M$ and in the end we explain how this yields a bound on R_K.

Plugging in the bounds of Theorems E.9.6, E.9.7 and E.9.12 into Theorem E.9.5, we have that for any number of intervals M:

$$\widetilde{C}^M \leq \sum_{k=1}^{K} \widetilde{V}_k^{\pi_k}(s_{\text{init}}) \mathbb{I}\{\Omega^{m(k)}\} + O\left(\frac{DS}{c_{\min}} \sqrt{MA} \log \frac{TSA}{\delta} + \frac{D^2 S^2 A}{c_{\min}^2} \log^2 \frac{TSA}{\delta}\right).$$

We now plug in the bound on M from Theorem E.9.2 into the bound above. After simplifying this gets us

$$\begin{aligned}\widetilde{C}^M \leq \sum_{k=1}^{K} \widetilde{V}_k^{\pi_k}(s_{\text{init}})\mathbb{I}\{\Omega^{m(k)}\} + O\bigg(&\sqrt{\frac{D^2S^2A}{c_{\min}^2}K\log^2\frac{TDSA}{\delta c_{\min}}} \\ &+\sqrt{\frac{D^4S^4A^2}{c_{\min}^4}\log^4\frac{TDSA}{\delta c_{\min}}}+\sqrt{\frac{DS^2A}{c_{\min}}\widetilde{C}^M\log^2\frac{TDSA}{\delta c_{\min}}}\bigg).\end{aligned}$$

From which, by solving for $\widetilde{C}^M$ (using that $x \leq a\sqrt{x}+b$ implies $x \leq (a+\sqrt{b})^2$ for $a \geq 0$ and $b \geq 0$), and simplifying the resulting expression by applying $\widetilde{V}_k^{\pi_k}(s_{\text{init}}) \leq D/c_{\min}$ and our assumptions that $K \geq S^2A$, $A \geq 2$, we get that

$$\widetilde{C}^M \leq \sum_{k=1}^{K} \widetilde{V}_k^{\pi_k}(s_{\text{init}})\mathbb{I}\{\Omega^{m(k)}\} + O\bigg(\frac{DS}{c_{\min}}\sqrt{AK}\log\frac{TDSA}{\delta c_{\min}}+\frac{D^2S^2A}{c_{\min}^2}\log^2\frac{TDSA}{\delta c_{\min}}\bigg). \quad \text{(E.15)}$$

Note that in particular, by simplifying the bound above, we obtain a polynomial bound on the total cost: $\widetilde{C}^M = O\Big(\sqrt{D^4S^4A^2KT/c_{\min}^4\delta}\Big)$. Next we combine this with the fact, stated in Theorem E.9.2 that $T \leq \widetilde{C}^M/c_{\min}$. Isolating T gets $T = O\Big(\frac{D^4S^4A^2K}{c_{\min}^4\delta}\Big)$, and plugging this bound back into Equation (E.15) and simplifying gets us

$$\widetilde{C}^M \leq \sum_{k=1}^{K} \widetilde{V}_k^{\pi_k}(s_{\text{init}})\mathbb{I}\{\Omega^{m(k)}\} + O\bigg(\frac{DS}{c_{\min}}\sqrt{AK}\log\frac{KDSA}{\delta c_{\min}}+\frac{D^2S^2A}{c_{\min}^2}\log^2\frac{KDSA}{\delta c_{\min}}\bigg). \quad \text{(E.16)}$$

Recall that

$$\sum_{k=1}^{K} \widetilde{V}_k^{\pi_k}(s_{\text{init}}) - V_k^{\pi^\star}(s_{\text{init}}) = \sum_{k=1}^{K}\langle q_k - q^{P,\pi^\star}, c^k\rangle,$$

and thus applying OMD analysis (see Section E.9.7) we obtain

$$\widetilde{R}_M \leq O\bigg(\frac{DS}{c_{\min}}\sqrt{AK}\log\frac{KDSA}{\delta c_{\min}}+\frac{D^2S^2A}{c_{\min}^2}\log^2\frac{KDSA}{\delta c_{\min}}\bigg).$$

Now, as Ω^m hold for all intervals, we use Equation (E.14) to bound the actual regret (together with $T \leq \widetilde{C}^M/c_{\min}$) for any number of intervals M, with the bound we have for $\widetilde{R}_M$.

We note that the bound above holds for any number of intervals M as long as K episodes do not elapse. As the instantaneous costs in the model are positive, this means that the learner must eventually finish the K episodes from which we derive the bound for R_K claimed by the theorem. □

E.9.7 OMD analysis

This analysis follows the lines of Section E.4, but it is adjusted to extended occupancy measures.

Lemma E.9.17. *Let $\tau \geq 1$. For every $q \in \widetilde{\Delta(\mathcal{M})}_m(\tau)$ it holds that $R(q) \leq \tau \log \tau$.*

Proof.

$$\begin{aligned}
R(q) &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log q(s,a,s') - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \\
&\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log q(s,a,s') \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \frac{q(s,a,s')}{\tau} + \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \tau \\
&\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \tau \leq \tau \log \tau,
\end{aligned}$$

where the first two inequalities follow from non-positivity, and the last one from the definition of $\widetilde{\Delta(\mathcal{M})}_m(\tau)$. □

Lemma E.9.18. *Let $\tau \geq 1$. For every $q \in \widetilde{\Delta(\mathcal{M})}_m(\tau)$ it holds that $-R(q) \leq \tau(1 + \log(S^2 A))$.*

Proof. Similarly to Theorem E.4.2 we have that

$$\begin{aligned}
-R(q) &= -\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \frac{q(s,a,s')}{\tau} + \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \\
&\quad - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} q(s,a,s') \log \tau \\
&\leq -\tau \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}^+} \frac{q(s,a,s')}{\tau} \log \frac{q(s,a,s')}{\tau} + \tau \leq \tau \log(S^2 A) + \tau,
\end{aligned}$$

where the first inequality follows because the last term is non-positive and from the definition of $\widetilde{\Delta(\mathcal{M})}_m(\tau)$, and the last inequality follows from properties of Shannon's entropy. □

Lemma E.9.19. *If Ω^m holds for all intervals m, then*

$$\sum_{k=1}^{K} \langle q_k - q^{P,\pi^\star}, c^k \rangle \leq \frac{2D}{c_{min}} \sqrt{6K \log \frac{DSA}{c_{min}}}.$$

Proof. We start with a fundamental inequality of OMD (see, e.g., [RM19a]) that holds for every $q \in \widetilde{\Delta(\mathscr{M})}_m(D/c_{\min})$ for every m (since Ω^m holds it also holds for $q^{P,\pi^\star}$),

$$\sum_{k=1}^{K} \langle q_k - q^{P,\pi^\star}, c^k \rangle \leq \sum_{k=1}^{K} \langle q_k - q'_{k+1}, c^k \rangle + \frac{\mathrm{KL}(q^{P,\pi^\star} \parallel q_1)}{\eta}. \tag{E.17}$$

For the first term we use the exact form of q'_{k+1} and the inequality $e^x \geq 1 + x$ to obtain

$$q'_{k+1}(s,a,s') = q_k(s,a,s') e^{-\eta c^k(s,a)} \geq q_k(s,a,s') - \eta q_k(s,a,s') c^k(s,a).$$

We substitute this back and obtain

$$\begin{aligned}\sum_{k=1}^{K} \langle q_k - q'_{k+1}, c^k \rangle &\leq \eta \sum_{k=1}^{K} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_k(s,a,s') c^k(s,a)^2 \leq \eta \sum_{k=1}^{K} \sum_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}} \sum_{s' \in \mathscr{S}^+} q_k(s,a,s') \\ &= \eta \sum_{k=1}^{K} \widetilde{T}_k^{\pi_k}(s_{\mathrm{init}}) \leq \eta K \frac{D}{c_{\min}}, \end{aligned} \tag{E.18}$$

where the last inequality follows from the definition of $\widetilde{\Delta(\mathscr{M})}_{m(k)}(D/c_{\min})$.

Next we use Theorems E.9.17 and E.9.18 to bound the second term of Equation (E.17). Recall that q_1 minimizes R in $\widetilde{\Delta(\mathscr{M})}_1(D/c_{\min})$, this implies that $\langle \nabla R(q_1), q^{P,\pi^\star} - q_1 \rangle \geq 0$ because otherwise we could decrease R by taking small step in the direction $q^{P,\pi^\star} - q_1$. Thus we obtain

$$\begin{aligned}\mathrm{KL}(q^{P,\pi^\star} \parallel q_1) &= R(q^{P,\pi^\star}) - R(q_1) - \langle \nabla R(q_1), q^{P,\pi^\star} - q_1 \rangle \leq R(q^{P,\pi^\star}) - R(q_1) \\ &\leq \frac{D}{c_{\min}} \log \frac{D}{c_{\min}} + \frac{D}{c_{\min}}(1 + \log(S^2 A)) \leq \frac{6D}{c_{\min}} \log \frac{DSA}{c_{\min}}. \end{aligned} \tag{E.19}$$

By substituting Equations (E.18) and (E.19) into Equation (E.17) and choosing $\eta = \sqrt{\frac{6 \log \frac{DSA}{c_{\min}}}{K}}$, we obtain,

$$\sum_{k=1}^{K} \langle q_k - q^{P,\pi^\star}, c^k \rangle \leq \eta K \frac{D}{c_{\min}} + \frac{6D}{c_{\min}\eta} \log \frac{DSA}{c_{\min}} \leq \frac{2D}{c_{\min}} \sqrt{6K \log \frac{DSA}{c_{\min}}}.$$

□

E.10 Estimating the SSP-diameter

When D is given, we use it to get the upper bound $D/c_{\min}$ on the expected time of the best policy in hindsight $T^{\pi^\star}(s_{\text{init}})$. The reason that $T^{\pi^\star}(s_{\text{init}}) \leq D/c_{\min}$ is that D is an upper bound on the expected time of the fast policy, i.e., $T^{\pi^f}(s_{\text{init}}) \leq D$ (see Theorem E.4.1). Thus, we want to compute $\widetilde{D}(s_{\text{init}})$ to be an upper bound on $T^{\pi^f}(s_{\text{init}})$.

We would like to use the first L episodes in order to estimate an upper bound $\widetilde{D}(s_{\text{init}})$ on the expected time of the fast policy, and then we can run SSP-O-REPS3 and obtain the same regret bound as in Theorem 7.3.1 but with $\widetilde{D}(s_{\text{init}})$ replacing D.

Notice that π^f is the optimal policy w.r.t the constant cost function $c(s,a) = 1$, and its expected cost is $T^{\pi^f}(s_{\text{init}})$. Thus, we run the SSP regret minimization algorithm of [RCMK20] with the cost function $c(s,a) = 1$ for L episodes. Then, we set $\widetilde{D}(s_{\text{init}})$ to be the average cost per episode times 10, that is,

$$\widetilde{D}(s_{\text{init}}) = \frac{10}{L} \sum_{k=1}^{L} \sum_{i=1}^{I^k} c(s_i^k, a_i^k) = \frac{10}{L} \sum_{k=1}^{L} I^k.$$

We start by showing that $\widetilde{D}(s_{\text{init}})$ is indeed an upper bound on $T^{\pi^f}(s_{\text{init}})$, given L is large enough.

Lemma E.10.1. *If* $L \geq \frac{2400 D^2}{T^{\pi^f}(s_{init})^2} \log^3 \frac{4K}{\delta}$ *then, with probability at least* $1-\delta$, $T^{\pi^f}(s_{init}) \leq \widetilde{D}(s_{init})$.

Proof. Notice that playing π^f during the first L episodes will result in smaller total cost then running the regret minimization algorithm. Thus, it suffices to prove the Lemma as if we are playing the fast policy. Define

$$X_k = \sum_{i=1}^{I^k} c(s_i^k, a_i^k) - \mathbb{E}\Big[\sum_{i=1}^{I^k} c(s_i^k, a_i^k) \mid P, \pi^f, s_1^k = s_{\text{init}}\Big] = \sum_{i=1}^{I^k} c(s_i^k, a_i^k) - T^{\pi^f}(s_{\text{init}}).$$

This is a martingale difference sequence, and in order to use Theorem E.12.5 we need to show that $\Pr[|X_k| > m] \leq 2e^{-\frac{m}{4D}}$ for every $k = 1, 2, \ldots$ and $m \geq 0$. This follows immediately from Theorem E.6.1 since the total cost is equal to the total time for the cost function $c(s,a) = 1$.

By Theorem E.12.5, $\left|\sum_{k=1}^{L} X_k\right| \leq 44D\sqrt{L \log^3 \frac{4L}{\delta}}$ with probability $1-\delta$. Therefore we have,

$$\sum_{k=1}^{L}\sum_{i=1}^{I^k} c(s_i^k,a_i^k) \ge LT^{\pi^f}(s_{\text{init}}) - 44D\sqrt{L\log^3\frac{4L}{\delta}},$$

and thus,

$$\frac{\widetilde{D}(s_{\text{init}})}{10} = \frac{1}{L}\sum_{k=1}^{L}\sum_{i=1}^{I^k} c(s_i^k,a_i^k) \ge T^{\pi^f}(s_{\text{init}}) - 44D\sqrt{\frac{\log^3\frac{4L}{\delta}}{L}}. \tag{E.20}$$

Since $L \ge \frac{2400D^2}{T^{\pi^f}(s_{\text{init}})^2}\log^3\frac{4K}{\delta}$, we have that $44D\sqrt{\frac{\log^3\frac{4L}{\delta}}{L}} \le \frac{9}{10}T^{\pi^f}(s_{\text{init}})$ and therefore we obtain from Equation (E.20) that $T^{\pi^f}(s_{\text{init}}) \le \widetilde{D}(s_{\text{init}})$. □

Next, we show that $\widetilde{D}(s_{\text{init}})$ is a good estimation of $T^{\pi^f}(s_{\text{init}})$, given L is large enough.

Lemma E.10.2. *If $L \ge S^2A\sqrt{D}\log^2\frac{KDSA}{\delta}$ then, with probability at least $1-\delta$, $\widetilde{D}(s_{init}) \le O(D)$.*

Proof. By the regret bound of the SSP regret minimization algorithm we have, with probability at least $1-\delta$,

$$\frac{1}{L}\sum_{k=1}^{L}\sum_{i=1}^{I^k} c(s_i^k,a_i^k) - T^{\pi^f}(s_{\text{init}}) \le O\left(\frac{DS\sqrt{A}\log\frac{LDSA}{\delta}}{\sqrt{L}} + \frac{D^{3/2}S^2A\log^2\frac{LDSA}{\delta}}{L}\right).$$

Since $T^{\pi^f}(s_{\text{init}}) \le D$ we obtain

$$\widetilde{D}(s_{\text{init}}) \le O\left(D + \frac{DS\sqrt{A}\log\frac{LDSA}{\delta}}{\sqrt{L}} + \frac{D^{3/2}S^2A\log^2\frac{LDSA}{\delta}}{L}\right) \le O(D).$$

where the last inequality follows because $L \ge S^2A\sqrt{D}\log^2\frac{KDSA}{\delta}$. □

The second place in which the SSP-O-REPS3 algorithm uses D is to determine when to switch to the optimistic fast policy. The switch happens when we reach a state with expected time larger than $D/c_{\min}$. A careful look at the analysis (especially Theorem E.9.6) shows that we actually need to switch in state s if the expected time is larger than $T^{\pi^f}(s)/c_{\min}$. Thus, we need to estimate an upper bound $\widetilde{D}(s)$ on $T^{\pi^f}(s)$ which is done exactly as we estimated $T^{\pi^f}(s_{\text{init}})$, i.e., in the first L visits to s we switch to the optimistic fast policy and then we can estimate $T^{\pi^f}(s)$ (by taking the average time to the goal times 10). This just means that now the threshold for a state to become known is L instead of $\frac{DS}{c_{\min}^2}\log\frac{DSA}{\delta c_{\min}}$.

Assuming L is large enough we get a good enough estimate, similarly to what we just proved for $T^{\pi^f}(s_{\text{init}})$.

To summarize, the algorithm proceeds as follows. We start by running the regret minimization algorithm of [RCMK20] with constant cost of 1 for L episodes and use it to estimate an upper bound on $T^{\pi^f}(s_{\text{init}})$. Then, we run SSP-O-REPS3 (setting η as a function of our estimate instead of D) with a known state threshold of L. When a state s becomes known we compute an upper bound on $T^{\pi^f}(s)$ and in the next episodes we make the switch in this state using this estimate and not when the expected time is larger than $D/c_{\min}$. The following theorem shows that we can set $L \approx \sqrt{K}$, and this leads to the same regret bound (as if we knew D in advance) assuming K is large enough. Otherwise, the regret is just bounded by some constant that does not depend on K.

Theorem E.10.3. *Under Theorem 5.2.1, running SSP-O-REPS3 with* $\eta = \sqrt{\frac{3\log(\widetilde{D}(s_{init})SA/c_{min})}{K}}$ *and*
$L = 2400 \max\{S^2 A \log^2 \frac{KSA}{\delta c_{min}}, \frac{\sqrt{K}}{c_{min}\sqrt{A}} \log \frac{KSA}{\delta c_{min}}\}$ *ensures that, with probability at least* $1-\delta$*,*

$$R_K \le O\left(\frac{DS}{c_{min}}\sqrt{AK}\log\frac{KDSA}{\delta c_{min}} + \frac{D^2S^2A}{c_{min}^2}\log^3\frac{KDSA}{\delta c_{min}}\right),$$

for $K \ge \max\left\{c_{min}^2 DS^4A^3\log^2\frac{DSA}{\delta c_{min}}, \frac{c_{min}^2 D^4 A}{\min_{s\in\mathscr{S}} T^{\pi^f}(s)^4}\log^4\frac{DSA}{\delta c_{min}}\right\}$*. For smaller K, we have*

$$R_K \le \widetilde{O}\left(\frac{D^3S^2A}{c_{min}^2} + c_{min}^2 D^5 A + D^2S^3A^2\right) \le \widetilde{O}\left(\frac{D^5S^3A^2}{c_{min}^2}\right).$$

Proof. First assume that K is large enough. By union bounds, Theorems E.10.1 and E.10.2 and the regret bound of SSP-O-REPS3 all hold with probability at least $1-3SA\delta$ (because of the $O(\cdot)$ notation it is the same as $1-\delta$). Therefore, $T^{\pi^f}(s) \le \widetilde{D}(s) \le O(D)$ for all $s \in \mathscr{S}$. During the first L episodes our cost is bounded as follows,

$$\begin{aligned}\sum_{k=1}^{L}\sum_{i=1}^{I^k} c^k(s_i^k,a_i^k) &\le LD + O\left(DS\sqrt{AL}\log\frac{LDSA}{\delta} + D^{3/2}S^2A\log^2\frac{LDSA}{\delta}\right)\\ &\le O\left(\frac{DS}{c_{\min}}\sqrt{AK}\log\frac{KDSA}{\delta} + \frac{D^{3/2}S^2A}{c_{\min}^2}\log^3\frac{KDSA}{\delta}\right),\end{aligned}$$

and then we bound the regret as in Theorem 7.3.1 to get the final result (the extra regret that comes from enlarging the known state threshold is at most $LDSA$ which is of the same order).

When K is too small we might encounter an underestimate or an overestimate of some $T^{\pi^f}(s)$. In the rest of the proof assume that $K > 2400 S^2 A \log^2 \frac{DSA}{\delta c_{\min}}$ because otherwise we never go past the diameter estimation phase and the regret is bounded by

$$R_K \leq \widetilde{O}\Big(D^{3/2} S^2 A\Big).$$

By following the proof of Theorem E.10.2, for $K > 2400 S^2 A c_{\min}^2 \log^2 \frac{DSA}{\delta c_{\min}}$ we have that $\widetilde{D} \leq O(D^{3/2})$.

Underestimate. The problem with an underestimate is that now our regret bound does not hold against $\pi^\star$, but only against the best policy in hindsight $\pi^\star(\widetilde{D}(s_{\text{init}}))$ with expected time of at most $\widetilde{D}(s_{\text{init}})$. In addition, we may loose D every time we switch to the fast policy (and the reason was reaching a "bad" state) by the proof of Theorem E.9.6. Thus, the regret bound of SSP-O-REPS3 (without diameter estimation) gives

$$\sum_{k=1}^{K} \sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) - \sum_{k=1}^{K} V_k^{\pi^\star(\widetilde{D}(s_{\text{init}}))}(s_{\text{init}}) \leq O\bigg(\frac{DS}{c_{\min}} \sqrt{AK} \log \frac{KDSA}{\delta c_{\min}} + \frac{D^2 S^2 A}{c_{\min}^2} \log^2 \frac{KDSA}{\delta c_{\min}} + KD \bigg). \tag{E.21}$$

We can use this to bound the total cost, and therfore the regret of the learner, as follows

$$\begin{aligned}
R_K &\leq \sum_{k=1}^{K} \sum_{i=1}^{I^k} c^k(s_i^k, a_i^k) \\
&\leq \sum_{k=1}^{K} V_k^{\pi^\star(\widetilde{D}(s_{\text{init}}))}(s_{\text{init}}) + O\bigg(\frac{DS}{c_{\min}} \sqrt{AK} \log \frac{KDSA}{\delta c_{\min}} + \frac{D^2 S^2 A}{c_{\min}^2} \log^2 \frac{KDSA}{\delta c_{\min}} + KD \bigg) \\
&\leq O\bigg(K\widetilde{D}(s_{\text{init}}) + \frac{DS}{c_{\min}} \sqrt{AK} \log \frac{KDSA}{\delta c_{\min}} + \frac{D^2 S^2 A}{c_{\min}^2} \log^2 \frac{KDSA}{\delta c_{\min}} + KD \bigg) \\
&\leq O\bigg(KD + \frac{DS}{c_{\min}} \sqrt{AK} \log \frac{KDSA}{\delta c_{\min}} + \frac{D^2 S^2 A}{c_{\min}^2} \log^2 \frac{KDSA}{\delta c_{\min}} \bigg) \\
&\leq \widetilde{O}\bigg(D^5 A c_{\min}^2 + D^3 SA + \frac{D^2 S^2 A}{c_{\min}^2} \bigg),
\end{aligned}$$

where the second inequality follows from Equation (E.21), the third because the expected time of $\pi^\star(\widetilde{D}(s_{\text{init}}))$ is at most $\widetilde{D}(s_{\text{init}})$, the forth because $\widetilde{D}(s_{\text{init}}) \leq D$ as an underestimate, and the last one is because underestimation may occur when $K < \frac{c_{\min}^2 D^4 A}{\min_{s \in \mathcal{S}} T^{\pi^f}(s)^4} \log^4 \frac{DSA}{\delta c_{\min}}$ according to Theorem E.10.1.

Overestimate. At this situation our regret bound holds, but its dependence in $\widetilde{D}(s)$ is problematic because $\widetilde{D}(s)$ overestimates D for some $s \in \mathscr{S}$. However, according to Theorem E.10.2, this may occur only when $K < c_{\min}^2 DS^4A^3 \log^2 \frac{DSA}{\delta c_{\min}}$. In addition, as mentioned before, $\widetilde{D}(s) \leq O(D^{3/2})$. Thus, we have

$$\begin{aligned}
R_K &\leq O\bigg(\frac{\widetilde{D}(s)S}{c_{\min}}\sqrt{AK}\log\frac{KDSA}{\delta c_{\min}} + \frac{\widetilde{D}(s)^2S^2A}{c_{\min}^2}\log^2\frac{KDSA}{\delta c_{\min}}\bigg) \\
&\leq O\bigg(\frac{D^{3/2}S}{c_{\min}}\sqrt{AK}\log\frac{KDSA}{\delta c_{\min}} + \frac{D^3S^2A}{c_{\min}^2}\log^2\frac{KDSA}{\delta c_{\min}}\bigg) \\
&\leq \widetilde{O}\bigg(D^2S^3A^2 + \frac{D^3S^2A}{c_{\min}^2}\bigg).
\end{aligned}$$

□

E.11 Zero costs

We can artificially fulfil Theorem 5.2.1 by adding a small $\varepsilon > 0$ perturbation to the costs. That is, when c^k is revealed, we pass to the learner the perturbed cost function $\tilde{c}^k(s,a) = \max\{c^k(s,a), \varepsilon\}$ for every $s \in \mathscr{S}$ and $a \in \mathscr{A}$.

Notice that changing the cost function does not change the transition function or the SSP-diameter. However, the bias introduced by our perturbation adds an additional $\varepsilon D^\star K$ term to the regret, where $D^\star$ is the expected time it takes the best policy in hindsight to reach the goal state.

Choosing ε to balance the algorithms' regret with the new term yields the following regret bounds for the general case. Theorem E.11.1 matches Theorem 7.2.1, Theorem E.11.2 matches Theorem 7.2.3, Theorem E.11.3 matches Theorem 7.3.1, and Theorem E.11.4 matches Theorem E.10.3.

Theorem E.11.1. *Running SSP-O-REPS with known transition function, $\eta = \sqrt{\frac{3\log(DSA/\varepsilon)}{K}}$ and $\varepsilon = K^{-1/4}$ ensures that*

$$\mathbb{E}[R_K] \le O\left(D^\star K^{3/4}\sqrt{\log(KDSA)}\right).$$

Theorem E.11.2. *Running SSP-O-REPS2 with known transition function, $\eta = \sqrt{\frac{3\log(DSA/\varepsilon)}{K}}$ and $\varepsilon = K^{-1/4}\sqrt{\log\frac{KDSA}{\delta}}$ ensures that, with probability $1-\delta$,*

$$R_K \le O\left(D^\star K^{3/4}\log\frac{KDSA}{\delta}\right).$$

Theorem E.11.3. *Running SSP-O-REPS3 with known SSP-diameter D, $\eta = \sqrt{\frac{3\log(DSA/\varepsilon)}{K}}$ and*
$\varepsilon = K^{-1/4}S\sqrt{A\log\frac{KDSA}{\delta}}$ ensures that, with probability $1-\delta$,

$$R_K \le O\left(D^\star S\sqrt{A}K^{3/4}\log\frac{KDSA}{\delta} + D^2\sqrt{K}\log\frac{KDSA}{\delta}\right).$$

Theorem E.11.4. *Running SSP-O-REPS3 with $\varepsilon = K^{-1/4}S\sqrt{A\log\frac{K\widetilde{D}(s_{init})SA}{\delta}}$, $\eta = \sqrt{\frac{3\log(\widetilde{D}(s_{init})SA/\varepsilon)}{K}}$ and*

$L = 2400\max\{S^2 A \log^2 \frac{KSA}{\delta\varepsilon}, \frac{\sqrt{K}}{\varepsilon\sqrt{A}} \log \frac{KSA}{\delta\varepsilon}\}$ ensures that, with probability at least $1-\delta$,

$$R_K \le O\left(D^\star S\sqrt{A}K^{3/4} \log \frac{KDSA}{\delta} + D^2\sqrt{K}\log^2 \frac{KDSA}{\delta}\right),$$

for $K \ge \max\left\{D^{2/3}S^4A^{8/3}\log^2\frac{DSA}{\delta}, \frac{D^{8/3}S^{4/3}A^{4/3}\log^{10/3}\frac{DSA}{\delta}}{\min_{s\in\mathscr{S}} T^{\pi^f}(s)^{8/3}}\right\}$. For smaller K, we have

$$\begin{aligned}
R_K &\le \widetilde{O}\left(D^\star S\sqrt{A}K^{3/4} + D^3\sqrt{K} + \frac{D^5 S^2 A^2}{\sqrt{K}} + D^2 S^3 A^2\right) \\
&\le \widetilde{O}\left(D^\star S\sqrt{A}K^{3/4} + D^3\sqrt{K} + D^5 A + D S^3 A^2\right) \\
&\le \widetilde{O}\left(D^\star S\sqrt{A}K^{3/4} + D^3\sqrt{K} + D^5 S^3 A^2\right).
\end{aligned}$$

Note that for $\varepsilon \le 1$ in Theorems E.11.3 and E.11.4, we need $K \ge S^4A^2$. However, if $K < S^4A^2$ (this is something the algorithm can check) we can just stay in the diameter estimation phase (i.e., assume all costs are $c(s,a) = 1$) and get a regret of $\widetilde{O}\big(DS^4A^2 + D^{3/2}S^2A\big)$ (or tune ε especially for this case for better results).

E.12 Concentration inequalities

Theorem E.12.1 (Anytime Azuma)**.** *Let $(X_n)_{n=1}^{\infty}$ be a martingale difference sequence such that $|X_n| \leq B_n$ almost surely. Then with probability at least $1-\delta$,*

$$\left|\sum_{n=1}^{N} X_n\right| \leq 4\sqrt{\sum_{n=1}^{N} B_n^2 \log\frac{N}{\delta}} \quad \forall N \geq 1.$$

Lemma E.12.2 ([RCMK20], Lemma B.15)**.** *Let $(X_t)_{t=1}^{\infty}$ be a martingale difference sequence adapted to the filtration $(\mathscr{F}_t)_{t=0}^{\infty}$. Let $Y_n = (\sum_{t=1}^{n} X_t)^2 - \sum_{t=1}^{n} \mathbb{E}[X_t^2 \mid \mathscr{F}_{t-1}]$. Then $(Y_n)_{n=0}^{\infty}$ is a martingale, and in particular if τ is a stopping time such that $\tau \leq c$ almost surely, then $\mathbb{E}[Y_\tau] = 0$.*

Lemma E.12.3 ([RCMK20], Lemma D.4)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables with expectation adapted to the filtration $(\mathscr{F}_n)_{n=0}^{\infty}$. Suppose that $0 \leq X_n \leq B$ almost surely. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathscr{F}_{i-1}] \leq 2\sum_{i=1}^{n} X_i + 4B\log\frac{2n}{\delta}. \tag{E.22}$$

Lemma E.12.4. *Let X be a non-negative random variable such that $\Pr[|X| > m] \leq ae^{-m/b}$ ($a \geq 1$) for all $m \geq 0$. Then, $\mathbb{E}[X\mathbb{I}\{X > r\}] \leq a(r+b)e^{-r/b}$.*

Proof. We have that,

$$\mathbb{E}[X\mathbb{I}\{X > r\}] = r\Pr[X > r] + \mathbb{E}[(X-r)\mathbb{I}\{X-r > 0\}],$$

and

$$\begin{aligned}
\mathbb{E}[(X-r)\mathbb{I}\{X-r > 0\}] &= \int_{m=0}^{\infty} \Pr[X-r > m]dm \\
&= \int_{m=r}^{\infty} \Pr[X > m]dm \\
&\leq \int_{m=r}^{\infty} ae^{-m/b}dm \\
&= abe^{-r/b}.
\end{aligned}$$

Hence $\mathbb{E}[X\mathbb{I}\{X > r\}] \leq a(r+b)e^{-r/b}$ as required. □

Theorem E.12.5 (Anytime Azuma for Unbounded Martingales)**.** *Let $(X_n)_{n=1}^{\infty}$ be a non-negative martingale difference sequence adapted to the filtration $(\mathscr{F}_n)_{n=1}^{\infty}$ such that $\Pr[|X_n| >$*

$m] \leq ae^{-m/b}$ $(a \geq 1)$ for all $n \geq 1$ and $m \geq 0$. Then, with probability at least $1-\delta$,

$$\Big|\sum_{n=1}^{N} X_n\Big| \leq 11b\sqrt{N\log^3\frac{2aN}{\delta}} \quad \forall N \geq 1.$$

Proof. Define $r_n = 2b\log\frac{2an}{\delta}$, and note that $\Pr[|X_n| > r_n] \leq \frac{\delta}{4n^2}$.

Additionally define $Y_n = X_n\mathbb{1}\{|X_n| \leq r_n\} - \mathbb{E}\left[X_n\mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right]$. $(Y_n)_{n=1}^{\infty}$ is a bounded martingale difference sequence, and by Theorem E.12.1 we have that with probability at least $1-\frac{\delta}{2}$,

$$\Big|\sum_{n=1}^{N} Y_n\Big| \leq 4\sqrt{\sum_{n=1}^{N} r_n^2\log\frac{N}{\delta}} \quad \forall N \geq 1.$$

Therefore, by a union bound, both the above holds and $|X_n| \leq r_n$ for all $n \geq 1$ with probability at least $1-\delta$. We get that

$$\Big|\sum_{n=1}^{N} X_n\mathbb{1}\{|X_n| \leq r_n\} - \mathbb{E}\left[X_n\mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right]\Big| \leq 4\sqrt{\sum_{n=1}^{N} r_n^2\log\frac{N}{\delta}},$$

and simplifying using the definition of r_n gets

$$\Big|\sum_{n=1}^{N} X_n\mathbb{1}\{|X_n| \leq r_n\}\Big| \leq \Big|\sum_{n=1}^{N} \mathbb{E}\left[X_n\mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right]\Big| + 8b\sqrt{N\log^3\frac{2aN}{\delta}}.$$

It thus remains to upper bound $\Big|\sum_{n=1}^{N} \mathbb{E}\left[X_n\mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right]\Big|$. First note that (since X_n is a martingale difference sequence)

$$\begin{aligned}\mathbb{E}\left[X_n\mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right] &= \mathbb{E}[X_n \mid \mathscr{F}_{n-1}] - \mathbb{E}\left[X_n\mathbb{1}\{|X_n| > r_n\} \mid \mathscr{F}_{n-1}\right] \\ &= -\mathbb{E}\left[X_n\mathbb{1}\{|X_n| > r_n\} \mid \mathscr{F}_{n-1}\right],\end{aligned}$$

from which

$$\begin{aligned}
\left|\sum_{n=1}^{N} \mathbb{E}\left[X_n \mathbb{1}\{|X_n| \leq r_n\} \mid \mathscr{F}_{n-1}\right]\right| &= \left|\sum_{n=1}^{N} \mathbb{E}\left[X_n \mathbb{1}\{|X_n| > r_n\} \mid \mathscr{F}_{n-1}\right]\right| \\
&\leq \sum_{n=1}^{N} \mathbb{E}\Big[|X_n| \mathbb{1}\{|X_n| > r_n\} \mid \mathscr{F}_{n-1}\Big] \\
&\leq \sum_{n=1}^{N} a(r_n + b)e^{-r_n/b} \\
&\leq \sum_{n=1}^{N} 3ab\left(\frac{\delta}{2an}\right)^2 \log\frac{2an}{\delta} \\
&\leq \sum_{n=1}^{N} 6ab\left(\frac{\delta}{2an}\right)^2 \left(\frac{2an}{\delta}\right)^{1/2} \\
&= \sum_{n=1}^{N} 6ab\left(\frac{\delta}{2an}\right)^{3/2} \\
&\leq \sum_{n=1}^{N} \frac{3b}{n^{3/2}} \leq 3b\log(N+1) \leq 3b\log(2N),
\end{aligned}$$

where the second inequality follows from Theorem E.12.4 and and the forth inequality follows because $\log x \leq 2\sqrt{x}$. □

תקציר

התחום של למידת מכונה באמצעות חיזוקים (RL) חוקר את השאלה הבסיסית ביותר בבינה מלאכותית (AI) – *איך יכול סוכן ללמוד לבצע החלטות נכונות באמצעות אינטראקציה עם הסביבה?*

למרות שלמידה באמצעות חיזוקים ראתה הצלחות אמפיריות מרשימות במגוון תצורות, הביצועים של אלגוריתמי למידה באמצעות חיזוקים משתנים באופן דרמטי בין תחומים שונים והם אפילו עלולים להיכשל בתהליך הלמידה בסיבות מסוימות. יכולות להיות לכך מגוון סיבות שונות אבל בתזה הזאת אנו נתמקד בשלוש הסיבות העיקריות הבאות:

1. *אקספלורציה.* רבים מהאלגוריתמים הפופולריים בתחום מסתמכים על היוריסטיקות פשוטות לביצוע אקפלורציה, כמו למשל ההיוריסטיקה האפסילון חמדנית. לכן, הם עלולים להיכשל בסביבות שבהן קשה להגיע לאזורים מסוימים של מרחב המצבים.
2. *סביבה משתנה.* המודל הפופולארי ביותר בלמידה באמצעות חיזוקים הוא מודל ההחלטה המרקובי (MDP). מודל סטוכסטי לחלוטין שאינו משתנה לאורך הזמן. למרות זאת, באפליקציות רבות, הסביבה אינה קבועה ומשתנה אפילו תוך כדי תהליך הלמידה. אלגוריתמים רבים לא מצליחים להסתגל לשינויים אלה.
3. *מודל לא מדויק.* הספרות על למידה באמצעות חיזוקים חוקרת בעיקר מודלים מרקובים עם קריטריוני ההצלחה של אופק סופי, תגמול מופחת או רווח ממוצע. למרות זאת, מצבים רבים (כמו ניווט וניתוב) לא מתאימים למסגרות אלה. לפיכך, אלגוריתמים רבים לא מצליחים לתפוס מצבים שכאלה באופן מספק.

תזה זו מציגה אלגוריתמים חדשים ותיאוריות חדשות לטיפול בנושאים שלעיל. האלגוריתמים שלנו מתמודדים עם האתגרים של אקספלורציה במגוון סביבות, ולכן ההצלחה שלהם נמדדת באמצעות *החרטה* – ההפרש בין ההפסד הכולל של הסוכן לאורך תהליך הלמידה ובין תוחלת ההפסד של המדיניות הטובה ביותר בדיעבד.

התזה מורכבת משני קווי מחקר עיקריים: תהליכי החלטה מרקובים אדברסריאלים (adversarial MDP) ובעיות מציאת מסלולים קלים ביותר סטוכסטיות (SSP). לאחר שאנחנו חוקרים את שני המודלים הללו, אנחנו חוקרים בנוסף מודל חדש, בעיות מציאת מסלולים קלים ביותר סטוכסטיות עם מחירים אדברסריאלים (adversarial SSP), שמשלב את שניהם לכדי מודל שהינו רובסטי וכללי יותר.

מטרתם של תהליכי החלטה מרקובים אדברסריאלים היא להתמודד עם סביבות משתנות. בניגוד לתהליכי החלטה מרקובים רגילים שהם סטוכסטיים ולא משתנים עם הזמן, בתהליכי החלטה מרקובים אדברסריאלים פונקצית המחירים יכולה להשתנות באופן ארביטרארי (כאשר פונקציית המעברים נותרת קבועה וסטוכסטית). מודל זה הוא כללי הרבה יותר מתהליכי החלטה מרקובים רגילים מכיוון שהוא מאפשר לפונקציית המחירים להיבחר בידי יריב, במקום להיות מוגרלת מתוך התפלגות לא ידועה כלשהי. בעבודה זו אנו מקדמים את ההבנה של מודל זה באופן משמעותי. אנחנו מציגים את חסמי החרטה הראשונים שהינם בהסתברות גבוהה עבור תהליכי החלטה מרקובים אדברסריאלים עם פונקציית מעברים לא ידועה ופידבק מלא (full-information feedback), כלומר הסוכן רואה את כל פונקציית המחירים לאחר שהיא משתנה. בנוסף, אנחנו מציגים את חסמי החרטה הראשונים עבור המודל המציאותי הרבה יותר של תהליכי החלטה מרקובים עם פונקציית מעברים לא ידועה ופידבק חלקי (bandit feedback), כלומר הסוכן רואה רק את המחירים עבור הפעולות שביצע. האלגוריתמים שלנו בנויים על מתודלוגיות רגולריזציה אנטרופית, שידועות כיעילות מאוד בפועל.

בעיות מציאת מסלולים קלים ביותר סטוכסטיות הן המודל הבסיסי ביותר של למידה באמצעות חיזוקים. הן מכילות את המודלים של החזר מופחת ואופק סופי כמקרים פרטיים. בבעיות אלה המטרה של הסוכן היא להגיע למצב מטרה שמוגדר מראש בתולחת הפסד מינימאלית. מודל זה

תופס מגוון של מצבים מציאותיים, כגון ניווט מכוניות, משחקים והטסת רחפנים ; כלומר משימות שמתבצעות באפיזודות שלבסוף מסתיימות. בעבודה זו אנו מציגים את חסמי החרטה הראשונים עבור בעיות מציאת מסלולים קלים ביותר סטוכסטיות שהינם קרובים לאופטימלים. לאחר מכן, אנו מפתחים אלגוריתם משופר שמבוסס על רדוקציה למודל האופק הסופי, ומוכיחים שאלגוריתם זה משיג חסם חרטה אופטימלי (כאשר מתעלמים מגורמים לוגריתמים).

אוניברסיטת תל-אביב

הפקולטה למדעים מדויקים ע״ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע״ש בלווטניק

אלגוריתמים למזעור חרטה בתהליכי החלטה מרקוביים

חיבור זה הוגש כעבודת מחקר לקראת התואר ״מוסמך דוקטור לפילוסופיה״ במדעי המחשב
על ידי

אביב רוזנברג

העבודה נעשתה בבית הספר למדעי המחשב
בהנחיית פרופ׳ ישי מנצור

אלול תשפ״ב

אוניברסיטת תל-אביב

הפקולטה למדעים מדויקים ע״ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע״ש בלווטניק

אלגוריתמים למזעור חרטה בתהליכי החלטה מרקוביים

חיבור זה הוגש כעבודת מחקר לקראת התואר ״דוקטור לפילוסופיה״ במדעי המחשב
על ידי

אביב רוזנברג

TEL AVIV UNIVERSITY
אוניברסיטת תל אביב

אלול תשפ״ב