UNIVERSITY OF CALIFORNIA SAN DIEGO

Sample Efficient Constrained Reinforcement Learning

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Lijing Kuang

Committee in charge:

Professor Sicun Gao, Chair
Professor Kamalika Chaudhuri
Professor Sanjoy Dasgupta

2020

The Thesis of Lijing Kuang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

TABLE OF CONTENTS

ABSTRACT OF THE THESIS

Sample Efficient Constrained Reinforcement Learning

by

Lijing Kuang

Master of Science in Computer Science

University of California San Diego, 2020

Professor Sicun Gao, Chair

The general assumption in reinforcement learning(RL) that agents are free to explore for searching optimal policies limits its applicability in real-world domains where safe exploration is desired. In this paper, we study the problem of constrained RL in episodic MDPs to investigate efficient exploration in safe RL. We formally describe two different constraint schemes frequently considered in empirical studies — namely, soft constrained RL that focuses on the overall safety satisfaction, and hard constrained RL that aims to provide guarantees throughout learning. While violations may occur in the former scheme, the latter enforces safety by extending the challenging knapsack problem in multi-armed bandits. Accordingly, we propose two novel sample efficient constrained Q-learning algorithms. By balancing exploration and exploitation based on UCB, our methods reduce the notoriously high sample complexity in constrained model-free settings while achieving asymptotically optimal solutions. Our theoretical analyses establish promising regret bounds for both algorithms.

Chapter 1

Introduction

Reinforcement learning (RL) is be widely adopted for addressing the problem of sequential decision making. Since the successful deployment of deep learning (DL) architectures with RL algorithms [Mnih et al., 2013], deep RL methods provide promising solutions to physical control problems with continuous spaces, by resorting to high-capacity function approximators and stochastic optimization techniques. In standard RL frameworks, we are concerned with the task of how an intelligent agent takes optimal actions when interacting with the unknown environment through a trial-and-error learning process, with the goal of maximizing the long-term rewards [Sutton and Barto, 2018]. Under such settings, agents are often assumed to have complete freedom for exploration so as to lead to performance improvement. This assumption however, may not hold in complex problem domains where multi-objectives are involved [Roijers et al., 2013].

In cyber-physical systems where safe exploration is of vital concern, a set of constraints need to be considered for safe-critical tasks such as collision-free navigation for autonomous cars [Lötjens et al., 2019], and prevention of catastrophic behaviors for robotics[Amodei et al., 2016]. In systems where collecting real-world experience data samples is expensive, it is desirable to quickly converge to optimal solutions with least finite samples and time steps. As such, scalar reward itself fails to capture the objectives appropriately, imposing the need for Multi-Objective RL (MORL) [Mossalam et al., 2016, Roijers et al., 2013] to generate safe policies while optimizing the long-term performance. Depending on the level of desired safety, different types of constraint incorporation are considered in empirical studies [Achiam et al., 2017, Roijers et al., 2013, Tessler et al., 2018, Wen and Topcu, 2018, Wu et al., 2018]. Nonetheless, it is an open question how these forms relate theoretically and whether one outperforms others.

To help understand the above issues, in this work, we study constrained RL with stochastic rewards and costs in episodic Markov Decision Processes (MDPs), where a budget constraint is imposed for both exploration and exploitation. Instead of learning actions with the highest expected rewards, our goal is to discover state-action pairs that not only yield favorable rewards but also incur insignificant costs. Specifically, we formally introduce two constrained RL schemes that categorize the general settings both in practice and in empirical works, namely, soft constrained RL and hard constrained RL. Hard constrained RL that enforces safety to provide any-time guarantee is critical for areas such as medical treatment and autonomous driving, whereas soft constrained RL adopts auxiliary costs as regularization to achieve preferable behaviors by minimizing cost violations in applications like sponsored search and recommendation.

Accordingly, to achieve efficient exploration in safe RL, we present two novel constrained RL algorithms based on Q-learning, with provable sample efficiency by adopting *upper confidence bound* (UCB)[Auer, 2002, Auer et al., 2002]. By balancing exploitation of current knowledge to maximize rewards, and exploration of optimism under uncertainty, our proposed methods, UCB-SCQ and UCB-HCQ, are able to guide the safe policies towards rapid convergence. To the best of our knowledge, this is the first work to bring these elements together in model-free settings that guarantees sample efficiency in both constraint schemes with rigorous analyses.

We establish theoretical regret bounds to quantitatively measure the sample complexity of both methods, distinguishing them from existing ones. This is a challenging task for several reasons. Due to the time-varying nature, we need to solve a stochastic constrained optimization problem [Achiam et al., 2017, Chow et al., 2019, Tessler et al., 2018] for asymptotically optimal solutions. Meanwhile, hard constrained RL extends multi-armed bandit (MAB) with knapsacks, which are *NP*-hard [Badanidiyuru et al., 2018, Tran-Thanh et al., 2010]. In addition, the stopping time involved cannot be easily determined because of stochastic costs. Our theoretical analyses conclude promising regret bounds for both algorithms, revealing a trade-off between performance and safety guarantees: UCB-HCQ ensures safety of policies all throughout training at the price of extra regret compared to UCB-SCQ where constraint satisfaction is relaxed and thus lower regret is attainable.

1.1 Related Work

The most widely-adopted formulation of RL with a set of constraints is constrained Markov Decision Processes (CMDPs) [Altman, 1999, Yu et al., 2019]. To encode the concept of safety, it augments standard MDP framework with constraints over expectations of auxiliary costs. When models are known in discrete tabular settings, a CMDP is solvable using linear programming (LP) [Altman, 1999]. However, results are limited for model-free scenarios where model dynamics are unknown, and for large-scale or even continuous state action spaces [Achiam et al., 2017, Chow et al., 2018, Yu et al., 2019]. More importantly, both objective and constraint in high-dimentional CMDP settings, where high-capacity function approximators are adopted, are non-convex. Recent methods in solving CMDPs in continuous spaces can be divided into two categories, in terms of ways to incorporate constraints. In soft constrained RL, it is a common practice to apply Lagrangian method with learnable Lagrangian multipliers and solve the converted unconstrained saddle-point optimization problem using policy-based methods [Bohez et al., 2019, Chow et al., 2017, Tessler et al., 2018]. Such Lagrangian methods achieve overall safety when policies converge asymptotically, nevertheless allowing possible violations during training. On the contrary, hard-constrained RL aims to learn safe policies throughout training. Representative works include Constrained Policy Optimization (CPO) based on trust region [Achiam et al., 2017], surrogate algorithms with stepwise [Dalal et al., 2018] and super-martingale [Mossalam et al., 2016] surrogate constraints, as well as Lyapunov-based approaches [Chow et al., 2018, Chow et al., 2019]. Still, there is no studies that provide unified formulations of both problems nor theoretical results are available to quantify their relationship in performance.

Meanwhile, with the booming demand of adopting RL in practical applications, provably efficient RL algorithms gain increasing attention in the community. Compared to model-based RL (MBRL)[Feinberg et al., 2018, Levine and Abbeel, 2014], Model-free RL gains greater popularity with simplicity in implementation, better computational and space complexities [Jin et al., 2018, Song and Sun, 2019]. Although asymptotic performance is achieved without requiring system dynamics, they suffer from notoriously high sample complexity [Feinberg et al., 2018, Gu et al., 2016, Levy and Ermon, 2018]. Encouragingly, recent advances have established finite sample-guarantees for model-free RL by managing the trade-off between exploration and exploitation [Abbasi-Yadkori et al., 2019, Azar et al., 2017, Jin et al., 2018, Jin et al., 2019, Lykouris et al., 2019, Russo, 2019, Song and Sun, 2019, Wainwright, 2019], which can be further improved with additional assumptions on system dynamics [Du et al., 2019], or using low-dimensional representation by parameterizing with features [Yang and Wang, 2019a, Yang and Wang, 2019b]. However, current results are limited in the context of standard RL formulations. Building upon these achievements, we derive regret bounds for constrained RL to develop provable sample efficient RL algorithms in model-free settings with budget constraint.

Chapter 2

Preliminaries

In this section, we describe the settings of standard and extended Markov Decision Process (MDP) in need to formulate the problem setup.

2.1 Unconstrained Markov Decision Process

An episodic MDP can be represented as a five-element tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, such that there are H steps in each of the K episodes during a complete learning process. The state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite discrete spaces with cardinality of S and A respectively. At each step $h \in [H]$, given the observation of a state-action pair (s, a), the transition kernel $\mathbb{P}_h(s_{h+1} | s_h = s, a_h = a)$ provides the distribution of the next state s_{h+1}, and the reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ emits a bounded reward signal between 0 and 1.

2.2 Constrained Markov Decision Process

A constrained MDP (CMDP) [Yu et al., 2019] is an extended MDP framework with budget constraint that reduces the size of policy space by eliminating infeasible policies. In CMDP framework, akin to reward $r_h(s, a)$, each state-action pair (s, a) is associated with a scalar cost value $c_h(s,a)$ that measures how expensive it is to take that specific action at a particular state.

Here, we introduce an episodic CMDP framework, which is a six-element tuple $(\mathcal{S},\mathcal{A},H,\mathbb{P},r,c)$, bounded by a predefined total cost budget B, where K episodes are rolled out for learning and H steps are involved in each of the $k \in [K]$ episodes. Unlike the standard episodic MDP problem, at each step $h \in [H]$, apart from a reward signal r_h realized from an underlying reward distribution with mean $0 \leq \mu_{r,h}(s_h,a_h) \leq 1$, a stochastic cost signal $c_h : \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$ is also taken into account, which can be seen as an immediate cost sampling from its underlying unknown cost distribution with mean $0 \leq c_{min} \leq \mu_{c,h}(s_h,a_h) \leq 1$. Assume the empirical realizations of costs c_h and rewards r_h are independently and identically distributed, and bounded between 0 and 1.

Throughout the learning process, the expected cumulative costs must not exceed an allocated budget. It is expected that high costs are received as penalty for risky actions which may lead to dangerous situations, whereas conservative actions are mapped to lower costs. By considering expected cumulative costs and rewards, the ultimate goal becomes forming policies with actions that satisfy the long-term safety while optimizing the long-term performance.

Chapter 3

Problem Setup: Episodic MDP with Constraints

We focus on episodic MDPs with finite spaces, finite time horizon, and an a-priori budget constraint. Instead of studying fixed and deterministic costs in earlier MAB studies [Tran-Thanh et al., 2012], we consider the more realistic and complex RL settings with stochastic rewards and time-varying costs, which are examined in recent empirical works of safe RL [Achiam et al., 2017, Tessler et al., 2018, Wen and Topcu, 2018].

At each episode $k \in [K]$, a policy π is a deterministic function $\{\pi : \mathcal{S} \times [H] \to \mathcal{A}\}$ that maps states into actions for each step. At step $h \in [H]$, the on-policy value function $V_h^\pi(s)$ gives the expected long term return by following policy π, starting from $s_h = s$, and the corresponding action-value function $Q_h^\pi(s,a)$ gives the same measure by also considering taking action $a_h = a$ when starting at $s_h = s$. Similarly, we introduce an episodic cost function $C_h^\pi(s)$ that describes the expected cost and its corresponding action-cost function $P_h^\pi(s,a)$. The above functions are formally denoted as:

$$
\begin{aligned}
V_h^\pi(s) &= \mathbb{E}^\pi\left[\sum_{i=h}^{H} r_i(s_i, \pi(s_i, i)) | s_h = s\right] = Q_h^\pi(s, \pi(s,h))\ , \\
Q_h^\pi(s,a) &= \mathbb{E}^\pi\left[\sum_{i=h}^{H} r_i(s_i, \pi(s_i, i)) | s_h = s, a_h = a\right] = r_h(s,a) + \sum_{s'} \mathbb{P}_h(s'|s,a) V_{h+1}^\pi(s')\ , \\
C_h^\pi(s) &= \mathbb{E}^\pi\left[\sum_{i=h}^{H} c_i(s_i, \pi(s_i, i)) | s_h = s\right] = P_h^\pi(s, \pi(s,h))\ , \\
P_h^\pi(s,a) &= \mathbb{E}^\pi\left[\sum_{i=h}^{H} c_i(s_i, \pi(s_i, i)) | s_h = s, a_h = a\right] = c_h(s,a) + \sum_{s'} \mathbb{P}_h(s'|s,a) C_{h+1}^\pi(s').
\end{aligned}
$$

For any function $Y \in \{Q\} \cup \{P\}$, define the linear operator $\mathbb{P}_h$ as: $[\mathbb{P}_h Y_{h+1}^\pi](s,a) \triangleq \sum_{s'} \mathbb{P}_h(s'|s,a)$ $Y_{h+1}^\pi(s') = \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} Y_{h+1}^\pi(s')$, and its empirical counterpart in episode k as: $[\hat{\mathbb{P}}_h^k Y_{h+1}^\pi](s,a) :=$ $Y_{h+1}^\pi(s_{h+1}^k)$, where s_{h+1}^k is realized from transition kernel $\mathbb{P}_h(\cdot|s_h^k = s, a_h^k = a)$. With Bellman optimality equation, the corresponding optimal functions can be specified as:

$$
\begin{aligned}
V_h^*(s) &= \sup_\pi V_h^\pi(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a), \qquad Q_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a). \\
C_h^*(s) &= \inf_\pi C_h^\pi(s) = \min_{a \in \mathcal{A}} P_h^*(s,a), \qquad P_h^*(s,a) = c_h(s,a) + [\mathbb{P}_h C_{h+1}^*](s,a).
\end{aligned}
$$

Definition 1. *Consider two performance metrics. The reward regret $\mathcal{G}_r(K)$ evaluates the expected sum of deviations from V^* across all episodes when following a chosen policy π_k in each episode $k \in [K]$. The cost regret $\mathcal{G}_c(K)$ measures the expected deviations of empirical cost from the optimal cost C^*. Both metrics are greater than or equal to* 0*:*

$$
\mathcal{G}_r(K) = \mathbb{E}\left[\textstyle\sum_{k=1}^K \left(V_1^{\pi^*}(s_1^k) - V_1^{\pi_k}(s_1^k)\right)\right], \quad \mathcal{G}_c(K) = \mathbb{E}\left[\textstyle\sum_{k=1}^K \left(C_1^{\pi_k}(s_1^k) - C_1^{\pi^*}(s_1^k)\right)\right].
$$

In practice of safe RL, recent studies employ constraints in distinct ways, which lead to diversified solutions and performance. In order to study how efficient exploration can be achieved when different types of constraint integration are being used, and to quantitatively measure the difference of sample complexity, we formally define two types of constrained RL problems.

3.1 Hard Constrained RL

In this setting, the learning process terminates whenever constraint violation occurs. Like MAB with knapsacks, optimal solutions need to be achieved for the reward-related objective function constrained by cost-related metrics. Suppose there exists an optimal policy π^* that gives the optimal value $V_h^*(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ starting from any step $h \in [H]$. Denote the policy space that consists of all stationary policies for the learning task as Π. In a learning process with K episodes, an agent chooses a policy π_k at the beginning of each episode k to complete a single roll-out.

Definition 2. (Hard Constrained RL). *Consider an episodic CMDP with budget constraint B. The goal in hard constrained RL is to find a set of valid policies $\{\pi_k | k \in [K]\}$ in the policy space Π, so as to minimize the reward regret over all episodes while ensuring the sum of expected cost does not exceed budget constraint B. Denote the shrunken policy space that consists of all valid policies as: $\Pi_B = \{\pi \in \Pi : \sum_{k=1}^K C_1^{\pi_k}(s_1^k) \le B\}$. The problem and its solution can be formally described as:*

$$\min_{\pi_k \in \Pi_B, k \in [K]} \mathcal{G}_r(K) \ , \quad s.t. \ \sum_{k=1}^{K} C_1^{\pi_k}(s_1^k) \le B \ .$$

$$\forall s \in \mathcal{S}, \quad \pi_k^* = \arg\min_{\pi \in \Pi_B} \mathcal{G}_r(K).$$

3.2 Soft Constrained RL

In soft constrained RL, constraint satisfaction is relaxed to allow possible violations. We introduce a new constrained reward function z_h based on Lagrangian methods, then formulate the soft constrained RL problems with a new composite objective function that concerns both rewards and costs.

Definition 3. *The constrained reward function w_h in episodic CMDP is defined as a weighted average of the immediate reward and immediate cost with a Lagrange multiplier λ:*

$$w_h(\lambda, s, a) = r_h(s,a) - \lambda c_h(s,a) \text{ , where } \lambda \in [0,1].$$

Let $[\mathbb{P}_h W^{\pi}_{h+1}](\lambda, s, a) \triangleq \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} W^{\pi}_{h+1}(s')$. The corresponding episodic on-policy constrained value function and constrained action-value function at step h can be denoted as:

$$W^{\pi}_h(\lambda, s) = \mathbb{E}^{\pi}\big[\sum_{i=h}^{H} r_i(s_i, \pi(s_i, i)) - \lambda c_i(s_i, \pi(s_i, i))|s_h = s\big] = V^{\pi}_h(s) - \lambda C^{\pi}_h(s),$$

$$F^{\pi}_h(\lambda, s, a) = w_h(\lambda, s, a) + [\mathbb{P}_h W^{\pi}_{h+1}](\lambda, s, a) = Q^{\pi}_h(s,a) - \lambda P^{\pi}_h(s,a).$$

Definition 4. (Soft Constrained RL). *The goal in soft constrained RL is to find a set of valid policies $\{\pi_k | k \in [K]\}$ in the policy space Π, so as to minimize a weighted average of reward regret $\mathcal{G}_r(K)$ and cost regret $\mathcal{G}_r(K)$ over all episodes. Intuitively, the optimal policies here are the ones that yield optimal combinations of high long-term return with small constraint violations:*

$$\min_{\pi_k \in \Pi, k \in [K]} \mathcal{G}_r(K) + \lambda \mathcal{G}_c(K) \text{ , where } \lambda \in [0,1],\ \mathcal{G}_r(K) \geq 0,\ \mathcal{G}_c(K) \geq 0.$$

$$\forall s \in \mathcal{S},\ \pi^*_k = \operatorname*{arg\,min}_{\pi \in \Pi}\ \mathcal{G}_r(K) + \lambda \mathcal{G}_c(K).$$

Chapter 4

Constrained Q-learning with UCB Exploration

In this chapter, we address the problem of efficient exploration in safe RL. To reduce sample complexity, one crucial factor is to strike the balance between exploration and exploitation. It is known that the lower regret bound achievable in any learning algorithm in terms of the total time steps T is logarithmic in T [Auer et al., 2002, Lai and Robbins, 1985]. While greedy and ε-greedy algorithms are commonly used for exploration due to simplicity, they lead to sub-optimal regrets that grow linearly in time. Recent studies have revealed and established that *minimax* regret bounds in tabular MDPs with finite-horizon for model-free RL can be bounded at the scale of $\sqrt{T}$ [Jin et al., 2018, Simchowitz and Jamieson, 2019].

Algorithm 1: Constrained Q-learning for Soft Constrained RL (UCB-SCQ)

```
   Input: number of episodes $K$, number of steps in each episode $H$, state space $\mathcal{S}$,
          action space $\mathcal{A}$, weighted coeffcient $\lambda$.
1  Initialization: $t = 0, \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], F_h(\lambda, s, a) = H, N_h(s,a) = 0, \hat{C} = 0$
2  for episode $k = 1, \dots, K$ do
3      Generate initial state $s_1$
4      for step $h = 1, \dots, H$ do
5          Choose action $a_h = \arg\max_{a \in A} F_h(\lambda, s_h, a)$
6          Observe reward $r_h(s_h, a_h)$ and cost $c_h(s_h, a_h)$
7          Update observed total cost $\hat{C} = \hat{C} + c_h(s_h, a_h)$
8          Calculate constrained-reward $w_h(\lambda, s_h, a_h) = r_h(s_h, a_h) - \lambda c_h(s_h, a_h)$
9          Generate the next state $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h)$
10         Update visiting counter $N_h(s_h, a_h) = N_h(s_h, a_h) + 1$, $t = N_h(s_h, a_h)$
11         Set learning rate as $\alpha_t = 2H/(2H - 1 + t)$
12         Set UCB exploration bonus as $U_{1t} = 4\sqrt{H^3 \ln(2SAT/\delta)/t}$
13         $F_h(\lambda, s_h, a_h) \leftarrow$
             $(1 - \alpha_t) F_h(\lambda, s_h, a_h) + \alpha_t [w_h(\lambda, s_h, a_h) + W_{h+1}(\lambda, s_{h+1}) + U_{1t}]$
14         $W_h(\lambda, s_h) = \min\{H,\ \max_{a \in A} F_h(\lambda, s_h, a)\}$
15     end
16 end
```

Based on these findings, we present Algorithm 1(UCB-SCQ) and Algorithm 2(UCB-HCQ) to solve the above two constrained RL schemes respectively. By adopting UCB for proactive exploration, our methods will be shown to have nice learning properties with provable sample efficiency. Both algorithms maintain their own action-value function, F_h or Q_h, action-cost function P_h, as well as visiting counter of state-action pairs N_h for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

To encourage exploration for the less visited state-action pairs for potential performance improvement, UCB bonuses are incorporated into the update rules in such a way that bonus terms U_{1t} (line 13), U_{2t} (line 15) are added into the estimates of action-value functions to increase selection probability. Here, $t := N_h(s,a)$ so that the bonus decreases as we become more confident. It should point out that conversely, for action-cost function P (line 16), the bonus term U_{3t} is subtracted from the new estimate. It is so, as lower cost is preferred and actions with greater uncertainty should be kept in feasible action set (line 5) for further consideration.

Our algorithms upper bound action-value functions, F (line 14) and Q (line 17), by H, lower bound action-cost function P (line 18) by 0, to restrict the optimism effect of UCB bonus under uncertainty not to perturb future updates and thus, reasonably control error propagation. In addition, the bounded reward signal $r_h \in [0,1]$ and cost $c_h \in [0,1]$ ensures that Q values are lower bounded by 0, so do V and C. However, it may be possible that F and W are negative in unfavorable circumstances, where costs are much more expensive than rewards received.

Furthermore, choosing appropriate learning rates for update rules is critical to reducing sample complexity. It is expected to learn from the most recent data for accurate estimates, whilst gradually forgetting earlier experience. They shall enable proactive exploration at the early stage of the learning process, then gradually reducing exploration to encourage more exploitation from accumulated knowledge along with time. Hence, desired learning rates should decay monotonically and emphasize on the most recent value estimates. Details are discussed in chapter 4.2.

4.1 Non-recursive Update Rules

To design learning rates and to upper bound regrets in a systematic way, we need to understand how errors propagate through episodes for reward-related functions Q_h, V_h, cost-related functions P_h C_h, and composite constraint functions W_h and F_h. Here, we derive the non-recursive update rules from their iterative expressions in UCB-SCQ (line 13) and UCB-HCQ (line 15, 16).

Algorithm 2: Constrained Q-learning for Hard Constrained RL (UCB-HCQ)

```
Input: number of episodes $K$, number of steps in each episode $H$, state space $\mathcal{S}$,
       action space $\mathcal{A}$, episodic budget proportion $\gamma$.
Initialization:
$t = 0, \forall(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H], Q_h(s,a) = H, P_h(s,a) = 0, N_h(s,a) = 0, \hat{C} = 0$
for episode $k = 1, \ldots, K$ do
    Generate initial state $s_1$
    for step $h = 1, \ldots, H$ do
        Generate feasible action set $A_h = \{a | P_h(s_h, a) \leq \gamma B\} \subseteq \mathcal{A}$
        Choose action $a_h = \arg\max_{a \in A_h} Q_h(s_h, a)$
        Observe reward $r_h(s_h, a_h)$ and cost $c_h(s_h, a_h)$
        Update observed total cost $\hat{C} = \hat{C} + c_h(s_h, a_h)$
        if $\hat{C} \leq B$ then
            Generate the next state $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h)$
            Update visiting counter $N_h(s_h, a_h) = N_h(s_h, a_h) + 1, t = N_h(s_h, a_h)$
            Set learning rate for function Q and P as $\alpha_t = \nu_t = 2H/(2H - 1 + t)$
            Set UCB exploration bonus for Q function as
              $U_{2t} = 2\sqrt{H^3 \ln(2SAT/\delta)/t}$
            Set UCB exploration bonus for P function as
              $U_{3t} = 2\sqrt{\gamma^2 B^2 H \ln(2SAT/\delta)/t}$
            $Q_h(s_h, a_h) \leftarrow (1 - \alpha_t) Q_h(s_h, a_h) + \alpha_t [r_h(s_h, a_h) + V_{h+1}(s_{h+1}) + U_{2t}]$
            $P_h(s_h, a_h) \leftarrow (1 - \nu_t) P_h(s_h, a_h) + \nu_t [c_h(s_h, a_h) + C_{h+1}(s_{h+1}) - U_{3t}]$
            $V_h(s_h) = \min\{H, \max_{a \in A_h} Q_h(s_h, a)\}$        // upper bounded by H
            $C_h(s_h) = \max\{0, \min_{a \in A_h} P_h(s_h, a)\}$        // lower bounded by 0
        else
            Terminate the learning process
        end
    end
end
```

Let $(s_h^k, a_h^k) = (s,a)$ be the observed state-action pair at step h in episode k, and Q_h^k, V_h^k, P_h^k, C_h^k, F_h^k, W_h^k be the corresponding $Q_h, V_h, P_h, C_h, F_h, W_h$ functions at the beginning of episode k.

Lemma 4.1. (Non-recursive Update Rules). *Consider $t = N_h^k(s,a)$ at the beginning of episode k and $(s_h^k, a_h^k) = (s,a)$, where (s,a) has been traversed at step h in some of the previous episodes $k_1, \ldots, k_t < k$. The non-recursive update rules for action-value functions $F_h(\lambda, s, a)$,*

$Q_h(s,a)$ and action-cost function $P_h(s,a)$ in UCB-SCQ and UCB-HCQ can be expressed as:

$$F_h^k(\lambda,s,a) = \alpha_t^0 H + \sum_{i=1}^{t} \left[\alpha_t^i\big(w_h(\lambda,s,a) + W_{h+1}^{k_i}(\lambda, s_{h+1}^{k_i}) + U_{1i}\big)\right], \tag{4.1}$$

$$Q_h^k(s,a) = \alpha_t^0 H + \sum_{i=1}^{t} \left[\alpha_t^i\big(r_h(s,a) + V_{h+1}^{k_i}(s_{h+1}^{k_i}) + U_{2i}\big)\right], \tag{4.2}$$

$$P_h^k(s,a) = \sum_{i=1}^{t} \left[\nu_t^i\big(c_h(s,a) + C_{h+1}^{k}(s_{h+1}^{k_i}) - U_{3i}\big)\right], \tag{4.3}$$

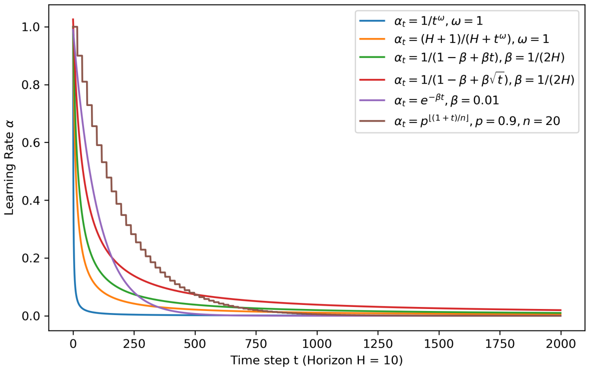
$$\textit{where}\ \ \alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j),\ \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j),\ \nu_t^i = \nu_i \prod_{j=i+1}^{t}(1-\nu_j),\ \ i \geq 1. \tag{4.4}$$

Proof of Lemma 4.1 is provided in Appendix A.

4.2 Selection of Learning Rate Schedules

To degenerate exploration in the later phase, learning rates α_t, ν_t are expected to decay, ideally exponentially, within domain $[0,1]$, throughout the learning process. Oppositely, with Lemma 4.1, the update weights, α_t^i and ν_t^i where $i = 0,1,...,t$, defined in (4.4) need to increase, ideally exponentially, along with time to stress on recent experience. Besides, for initialization purpose, α_t^i and ν_t^i should be zero when $t = 0$. To find out appropriate ones that satisfy all these requirements, we compare several different learning rate schedules in Fig. 1, and illustrate changes of the corresponding update weights in Fig. 2. Though all learning rate families being examined decay monotonically, some of them weight samples uniformly, others concentrate on data collected in different time frames. Based on our findings, we consider two suitable learning rate schedules for α_t and ν_t:

$$\alpha_t(\nu_t) = \begin{cases} \frac{H+1}{H+t^{\omega}}, & \text{if } t \geq 1 \\ 0, & \text{if } t = 0 \end{cases} \quad (4.5) \quad \text{or} \quad \begin{cases} \frac{1}{1-\beta+\beta t^{\omega}}, & \text{if } t \geq 1 \\ 0, & \text{if } t = 0 \end{cases} \quad (4.6),\ \omega \in [\frac{1}{2}, 1],\ \beta \in (0, \frac{1}{2}).$$



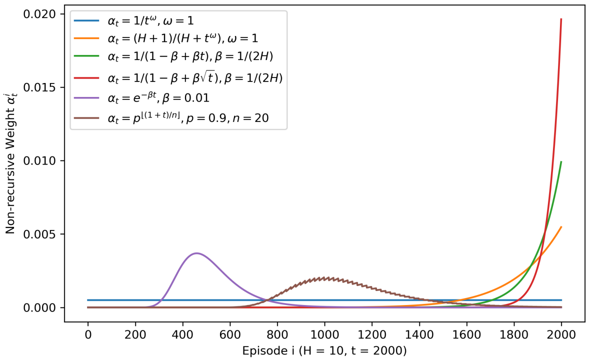


Fig. 1. (Left) Learning Rate Schedule Comparison ($H = 10$). We illustrate several learning rate schedules that give decaying learning rates. Hyperparameters are chosen for illustartion purpose.

Fig. 2. (Right) Update Weights Comparison ($H = 10$). The changes of α_t^i for the corresponding learning rates are depicted.

Learning rates in the above forms have been shown to converge in polynomial time [Even-Dar and Mansour, 2003, Ge et al., 2019]. From Fig. 1 and Fig. 2, it is also notable that the power of time step t affects the decaying speed of the resulting learning rates, which in turn decides the proportion of data that are with non-zero update weights. Depending on practical scenarios, it may be desirable to tune ω delicately. Essentially, if (4.5) and (4.6) share the same power of t, the resulting learning rate in (4.5) can be recovered from (4.6) by calibrating hyperparameter β. In particular, the specially-designed learning rate $\alpha_t = \frac{H+1}{H+t}$ adopted by [Jin et al., 2018] in form (4.5) can be recovered by setting $\omega = 1$, $\beta = 1/(H+1)$ in (4.6). Thus, we deem the learning rate schedule defined in (4.6) as a general form of interest.

In this work, we set hyperparameter $\omega = 1$ to smoothly replace data in use. One may select smaller value of ω for more aggressive evolution. In the later analyses, we wish to establish performance bounds in terms of T and H, therefore, β is defined as a function of H. Later we will show for $\omega = 1$, any $\beta(H) \in (0, \frac{1}{2})$ validates the provided sample efficiency. Without loss of generality, we consider a simple form of β and define the following learning rate for our algorithms:

$$\beta(H) = \frac{1}{2H}, \quad \alpha_t = \frac{2H}{2H-1+t}, \; H \geq 1, \; t \geq 0. \tag{4.7}$$

Chapter 5

Regret Analysis

In this chapter, we demonstrate our theoretical results for UCB-SCQ and UCB-HCQ.

5.1 Bounded Optimism with Confidence

As we will see, estimation errors in standard Q-learning that lead to the upward bias [Hessel et al., 2018, Van Hasselt et al., 2016] also exist in constrained Q-learning. To establish theoretical guarantees for constrained RL, we first measure the estimation error of constrained action value (cost) functions, then derive confidence bounds following principle of optimism in the face of uncertainty. To do so, one key step is to decompose optimal values (costs) in form of empirical estimates obtained by update rules in Lemma 4.1.

Lemma 5.1. (Estimation Error of Constrained Q-learning). *Consider $t = N_h^k(s,a)$ at the beginning of episode k and $(s_h^k, a_h^k) = (s,a)$, where (s,a) is a valid state-action pair that has been traversed at step h in some of the previous episodes $k_1, \ldots, k_t < k$. Here, t is the number of times (s,a) being visited at step h at the beginning of episode k. The estimation error of action-value* *and action-cost functions in UCB-SCQ and UCB-HCQ can be denoted as:*

$$\begin{aligned}
(F_h^k - F_h^*)(\lambda, s, a) &= \alpha_t^0 (H - F_h^*(\lambda, s, a)) \\
&\quad + \sum_{i=1}^t \alpha_t^i \big([(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) W_{h+1}^*](\lambda, s, a) + (W_{h+1}^{k_i} - W_{h+1}^*)(\lambda, s_{h+1}^{k_i}) + U_{1i} \big). \\
(Q_h^k - Q_h^*)(s, a) &= \alpha_t^0 (H - Q_h^*(s, a)) \\
&\quad + \sum_{i=1}^t \alpha_t^i \big([(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^*](s, a) + (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + U_{2i} \big). \\
(P_h^k - P_h^*)(s, a) &= -\nu_t^0 P_h^*(s, a) \\
&\quad + \sum_{i=1}^t \nu_t^i \big([(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) C_{h+1}^*](s, a) + (C_{h+1}^{k_i} - C_{h+1}^*)(s_{h+1}^{k_i}) - U_{3i} \big).
\end{aligned}$$

Lemma 5.1 allows us to naturally apply Azuma-Hoeffding concentration inequality and construct confidence bonus used in UCB-SCQ and UCB-HCQ.

Lemma 5.2. (Concentrated Bonus). *Let bonus be:*

$$U_{1t} = \sqrt{\frac{8H^2 \ln(2SAT/\delta)}{\beta t}}, U_{2t} = \sqrt{\frac{2H^2 \ln(2SAT/\delta)}{\beta t}}, and U_{3t} = \sqrt{\frac{2\gamma^2 B^2 \ln(2SAT/\delta)}{\beta t}},$$

where β is a function of horizon H. For constrained value function W, there exists a martingale difference sequence such that with probability at least $1-\delta$:

$$\left| \sum_{i=1}^t \alpha_t^i \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) W_{h+1}^*](\lambda, s, a) \right| \leq U_{1t},$$

where $t = N_h^k(s,a)$ and $k_1, \ldots, k_t \leq k$ are the episodes where (s,a) was traversed at step h. The same statement holds for Q and P in UCB-HCQ, with upper bound U_{2t} and U_{3t} respectively.

Proof detail is provided in Appendix C. In the following theorem, we show that estimation in constrained Q-learning is always optimistic in the sense that for value (cost) functions, estimates are always greater (smaller) than its optimal. With concentrated bonus, such optimism is reasonably bounded to control error propagation, and thus help establish upper regret bounds.

Theorem 1. (Bounded Optimism under Uncertainty). *Estimation in UCB-SCQ and UCB-HCQ is always optimistic. By setting bonus U_{1t}, U_{2t}, U_{3t} suggested in Lemma 5.2, then for any small $\delta > 0$, with probability at least $1-\delta$, such optimism is bounded almost surely:*

$$0 \le (F_h^k - F_h^*)(\lambda, s, a) \le 2\alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \big[(W_{h+1}^{k_i} - W_{h+1}^*)(\lambda, s_{h+1}^{k_i})\big] + 3U_{1t}.$$

$$0 \le (Q_h^k - Q_h^*)(s, a) \le \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \big[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\big] + 3U_{2t}.$$

$$-\nu_t^0 \gamma B + \sum_{i=1}^{t} \nu_t^i \big[(C_{h+1}^{k_i} - C_{h+1}^*)(s_{h+1}^{k_i})\big] - 3U_{3t} \le (P_h^k - P_h^*)(s, a) \le 0.$$

5.2 Regret Bounds

Definition 5. (Stability of Constrained Q-learning). *Two feasible action sets A_h, A_h' are equivalent if and only if they contain exactly the same actions, i.e., $A_h \equiv A_h'$ iff $\forall a \in \mathcal{A}$, $a \in A_h \iff a \in A_h'$. In constrained Q-learning, a feasible action set A_h^k at step h in episode k is deemed to be stable if for any small $\varepsilon \in [0,1] : A_h^k \equiv A_h^{k-1}$ and $\forall a \in A_h^k, s \in \mathcal{S}$, $\big|P_h^k(s,a) - P_h^{k-1}(s,a)\big| \le \varepsilon$. When a feasible action set becomes stable, constrained Q-learning is said to be stable, and the step is denoted as h_a. Define $\tau_{h_a,\pi}$ as the stable time of constrained Q-learning, which is a random variable depending on the above stability condition and can be formally defined as:*

$$\tau_{h_a,\pi} = \inf\{t \ge 0 : \exists h_a \in H, k_a = \lceil \tfrac{t}{H} \rceil \in [K], \textit{ such that } A_{h_a}^{k_a} \equiv A_{h_a}^{k_a-1} \text{ and }$$
$$\forall a \in A_h^{k_a}, s \in \mathcal{S}, \ \big|P_{h_a}^{k_a}(s,a) - P_{h_a}^{k_a-1}(s,a)\big| \le \varepsilon, \ \textit{under policy } \pi\}.$$

Assumption 1. *There exists an integer $T_a \le T < \infty$ such that $\mathbb{E}\big[\tau_{h_a,\pi}\big] \le T_a$, where the expectation is taken over the randomness of MDP and policy π.*

Theorem 2 (Regret in Constrained RL). *Consider the setting where Assumption 1 holds, $\gamma \le \frac{1}{K}$, $\beta = \frac{1}{2H}$, and $B \le T$ such that budget constraint is effective for learning. With bonus chosen in Lemma 5.2, for any $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 (UCB-SCQ) with bonus term U_{1t} achieves regret at most $O\big(\sqrt{H^4 SAT \ln(SAT/\delta)}\big)$; and regret for Algorithm 2 (UCB-HCQ) with bonus U_{2t}, U_{3t} is at most $O\big(H^2 SA + \sqrt{H^4 SAT_a \ln(SAT/\delta)} + (T-T_a)(H+\varepsilon)\big)$.*

Proof Sketch. Regret for UCB-SCQ can be easily shown to recover the result in [Jin et al., 2018]. For UCB-HCQ, result is established based on the worst case scenario where optimal action is costly.

Step 1: decompose reward regret. Based on stable time, decompose regret into two parts. Before stability, exploration dominates for estimates to converge, which switchs to exploitation afterwards:

$$\mathcal{G}_r(K) \le \sum_{k=1}^{k_a-1} \delta_1^k + \big(K-k_a\big)\big[\sum_{h=1}^{H}\big(\max_{a\in A} Q_h^*(s_1^{k_a},a) - Q_h^{\pi_{k_a}}(s_h^{k_a},a_h^{k_a})\big) + \varepsilon\big],$$

where $\delta_1^k := (V_1^k - V_1^{\pi_k})(s_1^k)$ is the approximate surrogate regret for each episode.

Step 2: bound localized surrogate regret. Let $\phi_h^k := (V_h^k - V_h^*)(s)$ be estimation error, $\xi_{h+1}^k := \big[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\big](s_h^k, a_h^k)$ be a martingale difference sequence. Bound δ_h^k at each step:

$$\begin{aligned}\sum_{k=1}^{k_a-1} \delta_h^k &\le \sum_{k=1}^{k_a-1} \big[\alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \phi_{h+1}^{k_i} + 3U_{2t} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k\big] \\ &\le SAH + (1+\frac{1}{2H-1}) \sum_{k=1}^{k_a-1} \delta_{h+1}^k + \sum_{k=1}^{k_a-1} \big(3U_{2N_h^k} + \xi_{h+1}^k\big)\end{aligned}$$

Step 3: recurse surrogate regret. With the above recursive expression, obtain complexity for δ_1^k:

$$\sum_{k=1}^{k_a-1} \delta_1^k \le O\big(H^2 SA + \sum_{h=1}^{H} \sum_{k=1}^{k_a-1} \big(3U_{2N_h^k} + \xi_{h+1}^k\big)\big).$$

Using UCB and Azuma-Hoeffding concentration bounds, we have:

$$\sum_{k=1}^{k_a-1}\sum_{h=1}^{H} 3U_{2N_h^k} = O(\sqrt{H^4 SAT_a \ln(SAT/\delta)}),$$
$$|\sum_{h=1}^{H}\sum_{k=1}^{k_a-1} \xi_{h+1}^k| \leq H\sqrt{2T_a \ln(2SAT/\delta)}.$$

Step 4: bound regret after stability. When UCB-HCQ becomes stable in episode k_a, optimal action becomes infeasible. Since estimates converge according to Definition 5, linear regret accumulates based on $\max_{a\in A} Q_h^*(s_1^{k_a}, a) - Q_h^{\pi_{k_a}}(s_h^{k_a}, a_h^{k_a})$ for the rest of time.

Putting everything together completes the proof. □

Theorem 2 suggests a trade-off between performance and safety guarantee. At each step in hard constrained RL, a feasible action set is being generated to exclude actions with high costs. In the case where optimal action is cost-consuming, only sub-optimal actions are valid for consideration to guarantee safety, leading to extra regret compared to soft-constrained RL.

Chapter 6

Summary and Conclusions

In this thesis, we propose two constrained RL algorithms with provable sample efficiency in model-free settings under budget constraint. In autonomous systems, hard constraints are considered preferably where violations of constraints may lead to serious consequences. In some cases, however, overconstraints may result in infeasible solution set, and instead of imposing strict conditions, soft constraints should be used to establish desired properties and to model the agent behaviors. Our work takes one step forward to theoretically understand the relation between these two types of constraint incorporation. In future work, it is of interest to relax the assumption of action set stability for general study and connect regret bounds with budget constraint.

Appendix A

Proof of Non-recursive Update Rule

Proof of Lemma 4.1. The recursive update rule for constrained action-value function F_h at step h in Algorithm 1 (UCB-SCQ) is:

$$F_h^{k+1}(\lambda, s, a) \leftarrow (1-\alpha_t)F_h^k(\lambda, s, a) + \alpha_t[w_h(\lambda, s, a) + W_{h+1}^k(\lambda, s_{h+1}^k) + U_{1t}]. \tag{A.1}$$

Correspondingly, update rules for action-value function Q_h, and action-cost function P_h in Algorithm 2 (UCB-HCQ) are:

$$Q_h^{k+1}(s,a) \leftarrow (1-\alpha_t)Q_h^k(s,a) + \alpha_t[r_h(s,a) + V_{h+1}^k(s_{h+1}^k) + U_{2t}], \tag{2}$$

$$P_h^{k+1}(s,a) \leftarrow (1-\nu_t)P_h^k(s,a) + \nu_t[c_h(s,a) + C_{h+1}^k(s_{h+1}^k) - U_{3t}], \tag{3}$$

where $t := N_h(s,a)$ counts the number of times state-action pair (s,a) being visited at step h, and U_{1t}, U_{2t}, U_{3t} are the upper confidence bounds that encourage exploration.

In the remaining of the proof, we demonstrate how to derive the non-recursive update rule for F_h, and results for Q_h and P_h follow naturally. With equation (1) and $F_h^{k_i}$ as the constrained action-value function at the beginning of episode k_i, starting with k_1, recursively we have:

$$\begin{aligned}
F_h^{k_1}(\lambda,s,a) &= H,\\
F_h^{k_2}(\lambda,s,a) &= (1-\alpha_1)F_h^{k_1}(\lambda,s,a)+\alpha_1[w_h(\lambda,s,a)+W_{h+1}^{k_1}(\lambda,s_{h+1}^{k_1})+U_{11}]\\
&= (1-\alpha_1)H+\alpha_1[w_h(\lambda,s,a)+W_{h+1}^{k_1}(\lambda,s_{h+1}^{k_1})+U_{11}],\\
F_h^{k_3}(\lambda,s,a) &= (1-\alpha_2)F_h^{k_2}(\lambda,s,a)+\alpha_2[w_h(\lambda,s,a)+W_{h+1}^{k_2}(\lambda,s_{h+1}^{k_2})+U_{12}]\\
&= (1-\alpha_2)(1-\alpha_1)H+(1-\alpha_2)\alpha_1[w_h(\lambda,s,a)+W_{h+1}^{k_1}(\lambda,s_{h+1}^{k_1})+U_{11}]\\
&+\alpha_2[w_h(\lambda,s,a)+W_{h+1}^{k_2}(\lambda,s_{h+1}^{k_2})+U_{12}],\\
F_h^{k_4}(\lambda,s,a) &= (1-\alpha_3)F_h^{k_3}(\lambda,s,a)+\alpha_3[w_h(\lambda,s,a)+W_{h+1}^{k_3}(\lambda,s_{h+1}^{k_3})+U_{13}]\\
&= (1-\alpha_3)(1-\alpha_2)(1-\alpha_1)H\\
&+(1-\alpha_3)(1-\alpha_2)\alpha_1[w_h(\lambda,s,a)+W_{h+1}^{k_1}(\lambda,s_{h+1}^{k_1})+U_{11}]\\
&+(1-\alpha_3)\alpha_2[w_h(\lambda,s,a)+W_{h+1}^{k_2}(\lambda,s_{h+1}^{k_2})+U_{12}]\\
&+\alpha_3[w_h(\lambda,s,a)+W_{h+1}^{k_3}(\lambda,s_{h+1}^{k_3})+U_{13}]\\
&= \prod_{j=1}^{3}(1-\alpha_j)H+\sum_{i=1}^{3}\left[\alpha_i\prod_{j=i+1}^{3}(1-\alpha_j)\left[w_h(\lambda,s,a)+W_{h+1}^{k_i}(\lambda,s_{h+1}^{k_i})+U_{1i}\right]\right],\\
&\dots\\
F_h^{k}(\lambda,s,a) &= \prod_{j=1}^{t}(1-\alpha_j)H+\sum_{i=1}^{t}\left[\alpha_i\prod_{j=i+1}^{t}(1-\alpha_j)\left[w_h(\lambda,s,a)+W_{h+1}^{k_i}(\lambda,s_{h+1}^{k_i})+U_{1i}\right]\right]. \qquad \square
\end{aligned}$$

Appendix B

Learning Rate Properties

In this chapter, we introduce the properties of selected learning rate that are auxiliary in the theoretical analyses of the presented algorithms.

B.1 Proof of Properties for Update Weight Coefficients

The following lemma regarding to the update weights would be used for decomposing optimal values for regret analysis.

Lemma B.1. *When learning rates are specified with the form in (4.5) or (4.6), the following properties hold for the update weights defined in (4.4) with any valid* β *and* ω*:*

1. $\alpha_t^0 = 1 \, and \, \sum_{i=1}^t \alpha_t^i = 0 \, for \, t = 0.$

2. $\alpha_t^0 = 0 \, and \, \sum_{i=1}^t \alpha_t^i = 1 \, for \, t \geq 1.$

In particular, the general form of learning rate schedule in (4.6) recovers (4.5) by setting $\beta = 1/(H+1)$.

Proof of Lemma B.1. Let $\beta = 1/(H+1)$ in (4.6), we have:

$$\alpha_t = \frac{1}{1 - \frac{1}{H+1} + \frac{t^\omega}{H+1}} = \frac{H+1}{H+t^\omega},$$

which recovers the learning rate in (4.5). Without loss of generality, we take the general form of learning rate schedule defined in (4.6) to continue the proof.

When $t = 0$, action value functions are initialized: $F_h^{k_1}(\lambda, s, a) = Q_h^{k_1}(s, a) = H$. According to the update rules in Lemma 4.1, coefficient for item H is 1, and the sum of coefficients for the remaining items is 0 *i.e.*, $\alpha_t^0 = 1$ and $\sum_{i=1}^t \alpha_t^i = 0$.

On the other hand, when $t \geq 1$, for any valid ω and β, we have:

$$\alpha_t^0 = \prod_{j=1}^t (1-\alpha_j) = (1-\alpha_1)\dots(1-\alpha_t) = \left(1 - \frac{1}{1-\beta+\beta}\right)\dots\left(1 - \frac{1}{1-\beta+\beta t^\omega}\right) = 0,$$

$$\begin{aligned}
\sum_{i=1}^t \alpha_t^i &= \sum_{i=1}^t \left[\alpha_i \prod_{j=i+1}^t (1-\alpha_j)\right], \ i \geq 1 \\
&= \alpha_1(1-\alpha_2)\dots(1-\alpha_t) + \alpha_2(1-\alpha_3)\dots(1-\alpha_t) + \dots + \alpha_t \\
&= \frac{1}{1-\beta+\beta}(1-\alpha_2)\dots(1-\alpha_t) + \alpha_2(1-\alpha_3)\dots(1-\alpha_t) + \dots + \alpha_t \\
&= (1-\alpha_2)\left[(1-\alpha_3)\dots(1-\alpha_t)\right] + \alpha_2\left[(1-\alpha_3)\dots(1-\alpha_t)\right] + \dots + \alpha_t \\
&= (1-\alpha_2+\alpha_2)\left[(1-\alpha_3)\dots(1-\alpha_t)\right] + \dots + \alpha_t \\
&\dots \\
&= (1-\alpha_t+\alpha_t) = 1.
\end{aligned}$$

□

B.2 Computational Properties of Learning Rate

As we shall see later, Lemma B.2 helps proof bounded optimism in Theorem 1, Lemma B.3 helps derive the upper confidence bound for exploration bonus used in UCB-SCQ and UCB-HCQ, and Lemma B.4 helps us justify that the increase of localized regret in each step can be upper bounded as a constant.

Lemma B.2. *With the selected learning rate in (4.6), the following inequalities specified* *in [Jin et al., 2018] hold for any valid* $\beta \in (0, \frac{1}{2})$ *when* $\omega = 1$*:*

$$\frac{1}{\sqrt{t}} \leq \sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}} \; for\ every\ t \geq 1.$$

Proof of Lemma B.2. Using induction on t, for the base case when $t = 1$, we have $\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} = \alpha_1^1 = \alpha_1 = 1$. For $t \geq 2$, note that:

$$\alpha_t^i = \alpha_i \prod_{j=i+1}^{t} (1-\alpha_j) = (1-\alpha_t)\big[\alpha_i \prod_{j=i+1}^{t-1} (1-\alpha_j)\big] = (1-\alpha_t)\alpha_{t-1}^i,\ i \in [1,\ t-1],$$

we thus have the following recursive relationship:

$$\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} = \sum_{i=1}^{t-1} \frac{\alpha_t^i}{\sqrt{i}} + \frac{\alpha_t^t}{\sqrt{t}} = (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} + \frac{\alpha_t}{\sqrt{t}}.$$

Using induction, assume $\frac{1}{\sqrt{t-1}} \leq \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t-1}}$ holds for $t-1$. Substitute the inequalities and $\alpha_t = \frac{H+1}{H+t}$ to the above equation, we have:

$$\begin{aligned}
\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} &= (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{i}} + \frac{\alpha_t}{\sqrt{t}} \\
&\leq \frac{2(1-\alpha_t)}{\sqrt{t-1}} + \frac{\alpha_t}{\sqrt{t}} = \frac{2\beta(t-1)}{[1+\beta(t-1)]\sqrt{t-1}} + \frac{1}{[1+\beta(t-1)]\sqrt{t}} \\
&\leq \frac{2\beta\sqrt{t}}{[1+\beta(t-1)]} + \frac{1}{[1+\beta(t-1)]\sqrt{t}} \\
&= \frac{2\beta t + 1}{\sqrt{t}[1+\beta(t-1)]} = \frac{2(\beta t - \beta + 1) + 2\beta - 1}{\sqrt{t}[1+\beta(t-1)]} \\
&= \frac{2}{\sqrt{t}} + \frac{2\beta - 1}{\sqrt{t}[1+\beta(t-1)]} \leq \frac{2}{\sqrt{t}},
\end{aligned}$$

where the final inequality holds as $\beta \in (0, \frac{1}{2})$ and $t \geq 1$.

Following the induction, we also obtain the lower bound:

$$(1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{\sqrt{i}}+\frac{\alpha_t}{\sqrt{t}}\geq\frac{1-\alpha_t}{\sqrt{t-1}}+\frac{\alpha_t}{\sqrt{t}}\geq\frac{1-\alpha_t}{\sqrt{t}}+\frac{\alpha_t}{\sqrt{t}}=\frac{1}{\sqrt{t}}.$$

□

Lemma B.3. *With the selected learning rate in (4.6), the update weights in the non-recursive update rule have the following properties for any valid* β *when* $\omega=1$*:*

$$\max_{i\in[t]}\alpha_t^i\leq\frac{1}{\beta t}\ \textit{and}\ \sum_{i=1}^{t}(\alpha_t^i)^2\leq\frac{1}{\beta t}\ \textit{for every}\ t\geq 1.$$

Proof of Lemma B.3. With the definition of α_t^i and by manipulation, we have:

$$\begin{aligned}\alpha_t^i&=\alpha_i\prod_{j=i+1}^{t}(1-\alpha_j)=\alpha_i(1-\alpha_{i+1})(1-\alpha_{i+2})\dots(1-\alpha_t)\\&=\frac{1}{1-\beta+\beta i}(1-\frac{1}{1-\beta+\beta(i+1)})(1-\frac{1}{1-\beta+\beta(i+2)})\dots(1-\frac{1}{1-\beta+\beta t})\\&=\frac{1}{1-\beta+\beta i}\cdot\frac{\beta i}{1-\beta+\beta(i+1)}\cdot\frac{\beta(i+1)}{1-\beta+\beta(i+2)}\cdots\frac{\beta(t-1)}{1-\beta+\beta t}\\&=\frac{1}{1-\beta+\beta t}\cdot\frac{\beta i}{(1-\beta)+\beta i}\cdot\frac{\beta(i+1)}{(1-\beta)+\beta(i+1)}\cdots\frac{\beta(t-1)}{(1-\beta)+\beta(t-1)}\\&\leq\frac{1}{(1-\beta)+\beta t}\leq\frac{1}{\beta t}.\end{aligned}$$

The last inequality holds as each remaining term involved i is strictly less than 1 with $\beta\in(0,\frac{1}{2})$.

With the fact that $\sum_{i=1}^{t}(\alpha_t^i)=1$, we also have:

$$\sum_{i=1}^{t}(\alpha_t^i)^2=\sum_{i=1}^{t}(\alpha_t^i\times\alpha_t^i)\leq\sum_{i=1}^{t}(\max_{i\in[t]}\alpha_t^i\times\alpha_t^i)=\max_{i\in[t]}\alpha_t^i\sum_{i=1}^{t}(\alpha_t^i)\leq\frac{1}{\beta t}\times 1=\leq\frac{1}{\beta t}.$$

□

Lemma B.4. *Let* $\beta=\frac{1}{2H}$ *and* $\omega=1$ *in (4.6), where* $H\geq 1$*, the following summation result* *holds for the resulting learning rate:*

$$\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{2H-1},\ for\ every\ i \geq 1.$$

Proof of Lemma B.4. Starting with the definition of α_t^i and α_i, we have:

$$\begin{aligned}
\sum_{t=i}^{\infty} \alpha_t^i &= \alpha_i^i + \alpha_{i+1}^i + \alpha_{i+2}^i + \dots \\
&= \alpha_i + \alpha_i(1-\alpha_{i+1}) + \alpha_i(1-\alpha_{i+1})(1-\alpha_{i+2}) + \dots \\
&= \alpha_i\left[1 + (1-\alpha_{i+1}) + (1-\alpha_{i+1})(1-\alpha_{i+2}) + \dots\right] \\
&= \frac{2H}{2H-1+i}\left[1 + (1-\frac{2H}{2H+i}) + (1-\frac{2H}{2H+i})(1-\frac{2H}{2H+i+1}) + \dots\right] \\
&= \frac{2H}{2H-1+i}\left[1 + \frac{i}{2H+i} + \frac{i}{2H+i}\cdot\frac{i+1}{2H+i+1} + \dots\right] \\
&= \frac{2H}{2H-1+i}\sum_{j=0}^{\infty} x_j,
\end{aligned}$$

where, x_j is defined as:

$$x_j = \begin{cases} 1, & \text{if } j = 0; \\ \displaystyle\prod_{k=1}^{j} \frac{i+k-1}{2H-1+i+k}, & \text{if } j \geq 1. \end{cases}$$

Next, we will show that the sum of the infinite series $\sum_{j=0}^{\infty} x_j$ converges to $\frac{2H-1+i}{2H-1}$, i.e.:

$$\sum_{j=0}^{\infty} x_j \to \frac{i+H}{H}.$$

To do so, we first prove the following relationship using induction on t:

$$\frac{2H-1+i}{2H-1} - \sum_{j=0}^{t} x_j = \frac{i}{2H-1}\prod_{j=1}^{t}\frac{i+j}{2H-1+i+j}.$$

As the base case, when $t = 0$, we have $\frac{2H-1+i}{2H-1} - 1 = \frac{i}{2H-1}$. The above relationship holds. Assume for $t-1$, the following holds:

$$\frac{2H-1+i}{2H-1} - \sum_{j=0}^{t-1} x_j = \frac{i}{2H-1} \prod_{j=1}^{t-1} \frac{i+j}{2H-1+i+j}.$$

Then for t, we have:

$$\begin{aligned}
\frac{2H-1+i}{2H-1} - \sum_{j=0}^{t} x_j &= \frac{2H-1+i}{2H-1} - \sum_{j=0}^{t-1} x_j - x_t \\
&= \frac{i}{2H-1} \prod_{j=1}^{t-1} \frac{i+j}{2H-1+i+j} - \prod_{j=1}^{t} \frac{i+j-1}{2H-1+i+j} \\
&= \frac{i}{2H-1} \prod_{j=1}^{t-1} \frac{i+j}{2H-1+i+j} - \frac{i+t-1}{2H-1+i+t} \prod_{j=1}^{t-1} \frac{i+j}{2H-1+i+j} \\
&= \prod_{j=1}^{t-1} \frac{i+j-1}{2H-1+i+j} \cdot \left[\frac{i}{2H-1} - \frac{i}{2H-1+i+t} \right] \\
&= \prod_{j=1}^{t-1} \frac{i+j-1}{2H-1+i+j} \cdot \frac{i(i+t)}{(2H-1)(2H-1+i+t)} \\
&= \frac{i}{2H-1} \prod_{j=1}^{t} \frac{i+j}{2H-1+i+j}.
\end{aligned}$$

As each item involved in the product is less then 1, when $t \to \infty$, we have the following limit:

$$\lim_{t\to\infty} \left[\frac{2H-1+i}{2H-1} - \sum_{j=0}^{t} x_j \right] = \lim_{t\to\infty} \left[\frac{i}{2H-1} \prod_{j=1}^{t} \frac{i+j}{2H-1+i+j} \right] = 0.$$

That is to say, the sum of the infinite series $\sum_{j=0}^{\infty} x_j$ converges to $\frac{2H-1+i}{2H-1}$. Thus, we have:

$$\sum_{t=i}^{\infty} \alpha_t^i = \frac{2H}{2H-1+i} \sum_{j=0}^{\infty} x_j = \frac{2H}{2H-1+i} \cdot \frac{2H-1+i}{2H-1} = 1 + \frac{1}{2H-1}, \quad \text{where } H \geq 1.$$

□

Appendix C

Proof of Constrained Functions Decomposition

Proof of Lemma 5.1. In UCB-SCQ, from Definition 3 and the Bellman optimal equation, we have:

$$
\begin{aligned}
F_h^*(\lambda,s,a) &= Q_h^*(s,a) - \lambda P_h^*(s,a) \\
&= r_h(s,a) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} V_{h+1}^*(s') - \lambda\big[c_h(s,a) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} C_{h+1}^*(s')\big] \\
&= w_h(\lambda,s,a) + \textstyle\sum_{s'} \mathbb{P}_h(s'|s,a) W_{h+1}^*(\lambda,s') = w_h(\lambda,s,a) + \mathbb{P}_h W_{h+1}^*(\lambda,s,a).
\end{aligned}
$$

With Lemma B.1, we have $\sum_{i=0}^{t} \alpha_t^i = 1$, and thus are able to decompose F_h^* as follows:

$$
\begin{aligned}
&F_h^*(\lambda,s,a) = \left[\textstyle\sum_{i=0}^{t} \alpha_t^i\right] F_h^*(\lambda,s,a) = \alpha_t^0 F_h^*(\lambda,s,a) + \textstyle\sum_{i=1}^{t} \alpha_t^i F_h^*(\lambda,s,a) \\
&= \alpha_t^0 F_h^*(\lambda,s,a) + \textstyle\sum_{i=1}^{t} \alpha_t^i \big[w_h(\lambda,s,a) + \mathbb{P}_h W_{h+1}^*(\lambda,s,a)\big] \\
&= \alpha_t^0 F_h^*(\lambda,s,a) + \textstyle\sum_{i=1}^{t} \alpha_t^i \big[w_h(\lambda,s,a) + \mathbb{P}_h W_{h+1}^*(\lambda,s,a) + W_{h+1}^*(\lambda,s_{h+1}^{k_i}) - W_{h+1}^*(\lambda,s_{h+1}^{k_i})\big] \\
&= \alpha_t^0 F_h^*(\lambda,s,a) + \textstyle\sum_{i=1}^{t} \alpha_t^i \big[w_h(\lambda,s,a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k_i}) W_{h+1}^*(\lambda,s,a) + W_{h+1}^*(\lambda,s_{h+1}^{k_i})\big]
\end{aligned}
$$

The third equality holds because of the empirical definition of constrained value function W^*: $\hat{\mathbb{P}}_h^{k_i} W_{h+1}^*(\lambda, s, a) := W_{h+1}^*(\lambda, s_{h+1}^{k_i})$. Similarly, for action-value function Q, and action-cost function P in UCB-HCQ, we have:

$$Q_h^*(s,a) = \alpha_t^0 Q_h^*(x,a) + \sum\nolimits_{i=1}^{t} \alpha_t^i \left[r_h(s,a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k_i}) V_{h+1}^*(s,a) + V_{h+1}^*(s_{h+1}^{k_i}) \right]$$

$$P_h^*(s,a) = \nu_t^0 P_h^*(s,a) + \sum\nolimits_{i=1}^{t} \nu_t^i \left[c_h(s,a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k_i}) C_{h+1}^*(s,a) + C_{h+1}^*(s_{h+1}^{k_i}) \right]$$

Expression in Lemma 5.1 can then be acquired by subtracting (4.1), (4.2), (4.3) from the above equations respectively. □

Appendix D

Proofs of Optimism Under Uncertainty

D.1 Proof of Lemma 5.2 (Concentrated bonus)

Proof of Lemma 5.2. In this proof, we derive the concentrated bonus U_{1t} used in UCB-SCQ, and results of U_{2t} and U_{3t} in UCB-HCQ follow exactly the same derivation. To obtain nice learning properties for UCB-SCQ and UCB-HCQ, we resort to concentration inequalities, in which the deviation probability decay exponentially in the distance from the expected value.

The key here is to upper bound the probability that the empirical optimal action-value $W^*_{h+1}(\lambda, s^{k_i}_{h+1})$, a.k.a. $\hat{\mathbb{P}}^{k_i}_h W^*_{h+1}(\lambda, s, a)$, at each episode k_i deviates from its expected value $\mathbb{E}_{s\prime \sim \mathbb{P}(\cdot | s^{k_i}_h = s, a^{k_i}_h = a)} W^*_{h+1}(\lambda, s\prime)$, a.k.a. $\mathbb{P}_h W^*_{h+1}(\lambda, s, a)$ by more than a predefined threshold.

Here we consider only the valid episodes where $k_i \in [K]$ for all $i = 1, \ldots, t$. Construct a martingale difference sequence $\{X_i\}_{i \in \mathbb{Z}_+} = \{[(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h) W^*_{h+1}](\lambda, s, a)\}^t_{i=1}$, with respect to the filtration $\{\mathcal{F}_i\}_{i \in \mathbb{Z}_+}$ such that:

1. $\{\mathcal{F}_i\}_{i \in \mathbb{Z}_+}$ is a nondecreasing collection of σ-fields: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \cdots \subseteq \mathcal{F}_n$, where $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$ captures the information that is known at step i.

2. X_i is $\mathcal{F}_i$-measurable, for all $i \geq 0$.

3. $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$, for every $i \geq 0$.

Then we have martingale $Z_i = \sum_{j=1}^{i} X_j$. Note that $|X_i| \leq 2H$ almost surely for all $i = 1, \ldots, t$. Because of linearity of expectation, $\{\alpha_t^i \cdot X_i\}_{i \in \mathbb{Z}_+}$ is also a martingale difference sequence. By Azuma-Hoeffding inequality, for all $\varepsilon > 0$:

$$\mathbb{P}\left[\left|\sum_{i=1}^{t} \alpha_t^i \cdot X_i\right| \geq \varepsilon\right] \leq 2exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^{t}(\alpha_t^i \cdot 2H)^2}\right).$$

Taking probability as $\delta/(SAKH)$, and applying the union bound over episodes K, then with probability at least $1-\delta/(SAH)$, we have:

$$\forall t \in [K] : \left|\sum_{i=1}^{t} \alpha_t^i \cdot X_i\right| \leq 2H\sqrt{\sum_{i=1}^{t}(\alpha_t^i)^2 \cdot 2\ln(2SAT/\delta)} \leq H\sqrt{\frac{8\ln(2SAT/\delta)}{\beta t}}.$$

The last inequality follows the fact that $\sum_{i=1}^{t}(\alpha_t^i)^2 \leq \frac{1}{\beta t}$ (see Lemma B.3 in Appendix B). Applying union bound for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ over all episodes, with probability at least $1-\delta$:

$$\left|\sum_{i=1}^{t} \alpha_t^i \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)W_{h+1}^*](\lambda, s, a)\right| \leq \sqrt{\frac{8H^2\ln(2SAT/\delta)}{\beta t}}, \quad \text{where } t = N_h^k(s,a) \leq K.$$

Similarly, we have:

$$U_{2t} = \sqrt{\frac{2H^2\ln(2SAT/\delta)}{\beta t}}, \quad U_{3t} = \sqrt{\frac{2\gamma^2 B^2\ln(2SAT/\delta)}{\beta t}}.$$

□

D.2 Proof of Theorem 1

Proof of Theorem 1. According to Lemma 5.1 and Lemma 5.2, we have:

$$\begin{aligned}
&(F_h^k - F_h^*)(\lambda, s, a) \\
&\leq \alpha_t^0 (H - F_h^*(\lambda, s, a)) + \textstyle\sum_{i=1}^t \alpha_t^i \left[(W_{h+1}^{k_i} - W_{h+1}^*)(\lambda, s_{h+1}^{k_i}) \right] + U_{1t} + \textstyle\sum_{i=1}^t \alpha_t^i U_{1i} \\
&\leq 2\alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(x_{h+1}^{k_i}) \right] + U_{1t} + \textstyle\sum_{i=1}^t \alpha_t^i U_{1i}.
\end{aligned}$$

The last inequality holds as $F_h^*(\lambda, s, a) \in [-H, H]$. With the selected concentrated bonus:

$$U_{1t} \leq \sum_{i=1}^t \alpha_t^i U_{1i} = \sum_{i=1}^t \alpha_t^i \sqrt{\frac{8H^2 \ln(2SAT/\delta)}{\beta i}} = \sqrt{\frac{8H^2 \ln(2SAT/\delta)}{\beta}} \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq 2U_{1t},$$

Here, we use the fact that $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$ (see Lemma B.2 in Appendix B). We thus have:

$$(F_h^k - F_h^*)(\lambda, s, a) \leq 2\alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(W_{h+1}^{k_i} - W_{h+1}^*)(\lambda, s_{h+1}^{k_i}) \right] + 3U_{1t}.$$

Similarly, we also have:

$$\begin{aligned}
&(Q_h^k - Q_h^*)(s, a) \leq \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) \right] + 3U_{2t}. \\
&(P_h^k - P_h^*)(s, a) \geq -\nu_t^0 \gamma B + \textstyle\sum_{i=1}^t \nu_t^i \left[(C_{h+1}^{k_i} - C_{h+1}^*)(s_{h+1}^{k_i}) \right] - 3U_{3t}.
\end{aligned}$$

On the other hand, in terms of optimism, it is required that $F_H^k(\lambda, s, a) \geq F_H^*(\lambda, s, a)$, $Q_H^k(s, a) \geq Q_H^*(s, a), P_k^h(s, a) \leq P_h^*(s, a)$, as less cost and higher value are desirable. Next, we will show that optimism under uncertainty is always guaranteed for estimations of both value function and cost function in constrained Q-learning, which can be verified with Lemma 5.2 and by induction on $h = H, H-1, \dots, 1$. In episodic MDP, an episode ends at step H. We thus have:

$$W^*_{H+1}(\lambda,s) = W^k_{H+1}(\lambda,s) = C^k_{H+1}(s) = C^*_{H+1}(s) = 0, \quad \forall k \in [K], (s,a) \in \mathcal{S}\times\mathcal{A},$$
$$F^*_H(\lambda,s,a) = w_H(\lambda,s,a) + [\mathbb{P}_H V^*_{H+1}](\lambda,s,a) = w_H(\lambda,s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$
$$P^*_H(s,a) = c_H(s,a) + [\mathbb{P}_h C^*_{H+1}](s,a) = c_H(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

With Lemma 5.2, bounded constrained reward $w_H(\lambda,s,a) \in [-1,1]$ and bounded cost $c_H(s,a) \in [-1,1]$, the base cases hold:

$$\begin{aligned}(F^k_H - F^*_H)(\lambda,s,a) &= \alpha^0_t(H - F^*_H(\lambda,s,a)) + \textstyle\sum_{i=1}^t \alpha^i_t U_{1i} \\ &= \alpha^0_t(H - w_H(\lambda,s,a)) + \sum_{i=1}^t \alpha^i_t U_{1i} \geq \alpha^0_t(H - w_H(\lambda,s,a)) + U_{1t} \geq 0.\end{aligned}$$
$$(P^k_H - P^*_H)(s,a) = -\alpha^0_t P^*_H(s,a) - \textstyle\sum_{i=1}^t \alpha^i_t U_{3i} \leq -\alpha^0_t c_H(s,a) - U_{3t} \leq 0.$$

Next, assume $(F^k_{h+1} - F^*_{h+1})(\lambda,s,a) \geq 0, (P^k_{h+1} - P^*_{h+1})(s,a) \leq 0$ hold at step $h+1$, $\forall (s,a,k) \in \mathcal{S}\times\mathcal{A}\times[K]$. Specifically, both hold for $k_i \in [K]$. With Lemma E.1 (see Appendix E), it also implies $(W^k_{h+1} - W^*_{h+1})(\lambda,s) \geq 0, (C^k_{h+1} - C^*_{h+1})(s) \leq 0$.

With Lemma 5.3, with probability of at least $1-\delta$:

$$\textstyle\sum_{i=1}^t \alpha^i_t \cdot [(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h) W^*_{h+1}](\lambda,s,a) \geq -U_{1t}, \quad \sum_{i=1}^t \alpha^i_t \cdot [(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h) C^*_{h+1}](s,a) \leq U_{3t}.$$

Plug in the above results back to Lemma 5.2, then at step h, with probability of at least $1-\delta$:

$$\begin{aligned}&(F^k_h - F^*_h)(\lambda,s,a) \\ &\geq \alpha^0_t(H - F^*_h(\lambda,s,a)) + \textstyle\sum_{i=1}^t \alpha^i_t \cdot 0 + \sum_{i=1}^t \alpha^i_t \cdot [(\hat{\mathbb{P}}^{k_i}_h - \mathbb{P}_h) W^*_{h+1}](\lambda,s,a) + \sum_{i=1}^t \alpha^i_t U_{1i} \\ &\geq \alpha^0_t(H - F^*_h(\lambda,s,a)) - U_{1t} + \textstyle\sum_{i=1}^t \alpha^i_t U_{1i} \geq 0; \\ &(P^k_h - P^*_h)(s,a)\end{aligned}$$

$$\leq -\alpha_t^0 P_h^*(s,a) + \textstyle\sum_{i=1}^t \alpha_t^i \cdot 0 + \sum_{i=1}^t \alpha_t^i \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) C_{h+1}^*](s,a) - \sum_{i=1}^t \alpha_t^i U_{3i}$$

$$\leq -\alpha_t^0 P_h^*(s,a) + U_{3t} - \textstyle\sum_{i=1}^t \alpha_t^i U_{3i} \leq 0.$$

The last inequality holds because $F_h^*(\lambda, s, a) \in [-H, H]$, $\sum_{i=1}^t \alpha_t^i U_{1i} \geq U_{1t}$, and $P_h^*(s,a) \in [0, \gamma B]$, $\sum_{i=1}^t \alpha_t^i U_{3i} \geq U_{3t}$. Therefore, in constrained Q-learning, the estimation of value functions and cost functions are always bounded optimistic.

The proof of $Q_h(s,a)$ follows exactly the same procedure and thus is omitted.

□

Appendix E

Auxiliary Proofs for Regret Analysis

Lemma E.1. *In constrained RL, when optimism under uncertainty holds for action-value functions (action-cost function), it also holds for value functions (cost function):*

$$
\begin{aligned}
&(F_h^k - F_h^*)(\lambda, s, a) \geq 0 \Rightarrow (W_h^k - W_h^*)(\lambda, s) \geq 0, \\
&(Q_h^k - Q_h^*)(s, a) \geq 0 \Rightarrow (V_h^k - V_h^*)(s) \geq 0, \\
&(P_h^k - P_h^*)(s, a) \leq 0 \Rightarrow (C_h^k - C_h^*)(s) \leq 0.
\end{aligned}
$$

Proof of Lemma E.1. Recall that $W_h^k(\lambda, s) \leftarrow min\{H, max_{a\in\mathcal{A}} F_h^k(\lambda, s, a)\}$, $V_h^k(s_h^k) \leftarrow \min\{H,\ \max_{a\in\mathcal{A}_h} Q_h^k(s, a)\}$, and $C_h^k(s) \leftarrow \max\{0,\ \min_{a\in A_h} P_h^k(s, a)\}$:

$$
\forall s \in \mathcal{S},\ (W_h^k - W_h^*)(\lambda, s) = \begin{cases} \max_{a\in\mathcal{A}}(F_h^k - F_h^*)(\lambda, s, a), & \text{if} \max_{a\in\mathcal{A}} F_h^k(\lambda, s, a) < H \\ H - \max_{a\in\mathcal{A}} F_h^*(\lambda, s, a), & \text{if} \max_{a\in\mathcal{A}} F_h^k(\lambda, s, a) \geq H \end{cases};
$$

$$
\forall s \in \mathcal{S},\ (V_h^k - V_h^*)(s) = \begin{cases} \max_{a\in\mathcal{A}_h}(Q_h^k - Q_h^*)(s, a), & \text{if} \max_{a\in\mathcal{A}_h} Q_h^k(s, a) < H \\ H - \max_{a\in\mathcal{A}_h} Q_h^k(s, a), & \text{if} \max_{a\in\mathcal{A}_h} Q_h^k(s, a) \geq H \end{cases};
$$

$$\forall s \in \mathcal{S},\, (C_h^k - C_h^*)(s) = \begin{cases} 0 - \min_{a \in A_h} P_h^*(s,a), & \text{if} \min_{a\in A_h} P_h^k(s,a) < 0 \\ \min_{a\in A_h}(P_h^k(s,a) - P_{h+1}^*(s,a)), & \text{if} \min_{a\in A_h} P_h^k(s,a) \geq 0 \end{cases}.$$

Thus, when $(F_h^k - F_h^*)(\lambda, s, a) \geq 0, (Q_h^k - Q_h^*)(s) \geq 0, (P_h^k - P_h^*)(s,a) \leq 0$ hold at step h, $\forall (s,a,k) \in \mathcal{S} \times \mathcal{A} \times [K]$, with the facts that F and Q are upper bounded by H, P is lower bounded by 0, it also implies $(W_h^k - W_h^*)(\lambda, s) \geq 0, (V_h^k - V_h^*)(s) \geq 0, (C_h^k - C_h^*)(s) \leq 0$:

□

Lemma E.2. *In constrained Q-learning, for any $(k,h) \in [K] \times [H]$ and s_h^k being the state visited, we have $(V_h^k - V_h^{\pi_k})(s_h^k) \leq (Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k)$ and $(C_h^{\pi_k} - C_h^k)(s_h^k) \leq (P_h^{\pi_k} - P_h^k)(s_h^k, a_h^k)$, where $a_h^k \in A_h$ is the optimal feasible action.*

Proof of Lemma E.2. Recall that $V_h^k(s_h^k) = \min\{H,\ \max_{a\in\mathcal{A}_h} Q_h^k(s_h^k, a)\}$:

$$V_h^k(s_h^k) = \begin{cases} H, & \text{if } H \leq \max_{a\in\mathcal{A}_h} Q_h^k(s_h^k, a) \\ \max_{a\in\mathcal{A}_h} Q_h^k(s_h^k, a), & \text{if } H > \max_{a\in\mathcal{A}_h} Q_h^k(s_h^k, a) \end{cases},$$

we thus have $V_h^k(s_h^k) \leq \max_{a\in\mathcal{A}_h} Q_h^k(s_h^k, a) = Q_h^k(s_h^k, a_h^k)$. Note that for a given state s_h^k at step h in episode k, $V_h^{\pi_k}(s_h^k) = Q_h^{\pi_k}(s_h^k, \pi_k(s_h^k)) = Q_h^{\pi_k}(s_h^k, a_h^k)$. Therefore, $(V_h^k - V_h^{\pi_k})(s_h^k) \leq Q_h^k(s_h^k, a_h^k) - V_h^{\pi_k}(s_h^k) = (Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k)$.

For cost function, recall that $C_h^k(s_h^k) = \max\{0,\ \min_{a\in A_h} P_h^k(s_h^k, a)\}$:

$$C_h^k(s_h^k) = \begin{cases} \min_{a\in A_h} P_h^k(s_h^k, a), & \text{if } 0 \leq \min_{a\in A_h} P_h^k(s_h^k, a) \\ 0, & \text{if } 0 > \min_{a\in A_h} P_h^k(s_h^k, a) \end{cases},$$

i.e., $C_h^k(s_h^k) \geq \min_{a\in A_h} P_h(s_h^k, a)$. Meanwhile, $C_h^{\pi_k}(s_h^k) = P_h^{\pi_k}(s_h^k, \pi_k(s_h^k)) = P_h^{\pi_k}(s_h^k, a_h^k)$. Hence, $(C_h^{\pi_k} - C_h^k)(s_h^k) \leq C_h^{\pi_k}(s_h^k) - P_h^k(s_h^k, \pi_k(s_h^k)) = (P_h^{\pi_k} - P_h^k)(s_h^k, a_h^k)$.

□

Lemma E.3. (Decomposition of Surrogate Regret). *For any fixed $(k,h) \in [K] \times [H]$, let $t = N_h^k(s_h^k, a_h^k)$, and suppose (s_h^k, a_h^k) was previously taken at step h of episodes $k_1, \ldots, k_t < k$. The following upper bound holds for localized surrogate regret $\delta_h^k := (V_h^k - V_h^{\pi_k})(s_h^k)$:*

$$\delta_h^k \leq \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k_i} + 3U_{2t} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k,$$

where $\phi_{h+1}^k := (V_{h+1}^k - V_{h+1}^*)(s_{h+1}^k) \geq 0$, and $\xi_{h+1}^k := \left[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k)$ is a martingale difference sequence w.r.t. the regret.

Proof of Lemma E.3. With Lemma E.2, Bellman equations, and Theorem 1, δ_h^k can be decomposed as follows:

$$\begin{aligned}
\delta_h^k &= (V_h^k - V_h^{\pi_k})(s_h^k) \\
&\leq (Q_h^k - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&= (Q_h^k - Q_h^*)(s_h^k, a_h^k) + (Q_h^* - Q_h^{\pi_k})(s_h^k, a_h^k) \\
&= (Q_h^k - Q_h^*)(s_h^k, a_h^k) + [r_h(s_h^k, a_h^k) + \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k)] - [r_h(s_h^k, a_h^k) + \mathbb{P}_h V_{h+1}^{\pi_k}(s_h^k, a_h^k)] \\
&= (Q_h^k - Q_h^*)(s_h^k, a_h^k) + [\mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\pi_k})](s_h^k, a_h^k) \\
&\leq \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right] + 3U_{2t} + [\mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\pi_k})](s_h^k, a_h^k) \\
&= \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right] + 3U_{2t} + \left[\hat{\mathbb{P}}_h^k(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&\qquad + \left[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&= \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right] + 3U_{2t} + (V_{h+1}^* - V_{h+1}^{\pi_k})(s_{h+1}^k) \\
&\qquad + \left[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&= \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i})\right] + 3U_{2t} + \left[(V_{h+1}^* - V_{h+1}^k + V_{h+1}^k - V_{h+1}^{\pi_k}\right](s_{h+1}^k) \\
&\qquad + \left[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k) \\
&= \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k_i} + 3U_{2t} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k,
\end{aligned}$$

where $\phi_{h+1}^k := (V_{h+1}^k - V_{h+1}^*)(s_{h+1}^k) \geq 0$ (Lemma E.1), and $\xi_{h+1}^k := \big[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\big](s_h^k, a_h^k)$ is a martingale difference sequence. The eighth equality holds as $[\hat{\mathbb{P}}_h^k V_{h+1}^{\pi_k}](s_h^k, a_h^k) = V_{h+1}^{\pi_k}(s_{h+1}^k)$ and $[\hat{\mathbb{P}}_h^k V_{h+1}^*](s_h^k, a_h^k) = V_{h+1}^*(s_{h+1}^k)$. □

Appendix F

Proof of Theorem 2

Proof of Theorem 2. Intuitively, UCB-SCQ subsumes costs into rewards by expanding the original range, and thus has the same order of complexity as proposed in [Jin et al., 2018]. Proof idea of UCB-SCQ can be referred to [Jin et al., 2018]. In this proof, we focus on UCB-HCQ for hard-contained RL. Specifically, we analyze the regret bound for the worst case scenario where optimal action is costly and become infeasible upon convergence.

Assume assumption 1 holds, and let $k_a = \lceil \frac{t}{H} \rceil$ be the episode in which UCB-HCQ becomes stable. We then divide the total regret based on k_a. For time steps before stability, exploration dominates and gradually switch to exploitation when estimates become precise and stable. We will show that with Theorem 1, the first part of the regret is guaranteed to be upper bounded by the approximated regret because of bounded optimism. This makes sense, as in model-free settings without oracles, the optimal values are unknown and it is essential to use the estimated constrained values as surrogate functions.

Before stability, with probability at least $1-\delta$, we have $Q_h^k(s,a) \geq Q_h^*(s,a)$, which holds for all valid state-action pair. Let $\phi_h^k := (V_h^k - V_h^*)(s)$, which measures the upward bias in estimation and is always greater than or equal to zero (Lemma E.1). Define $\delta_h^k := (V_h^k - V_h^{\pi_k})(s_h^k)$, which represents the approximate surrogate regret at each step h. We thus have:

$$\begin{aligned}
\mathcal{G}_r(K) &= \mathbb{E}\Bigg[\sum_{k=1}^{K}\big(V_1^{\pi^*}(s_1^k) - V_1^{\pi_k}(s_1^k)\big)\Bigg] = \sum_{k=1}^{K}\big(\mathbb{E}[V_1^*(s_1^k)] - \mathbb{E}[V_1^{\pi_k}(s_1^k)]\big)\\
&= \sum_{k=1}^{K}\big(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\big) = \sum_{k=1}^{k_a-1}\big(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\big) + \sum_{k=k_a}^{K}\big(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\big)\\
&\le \sum_{k=1}^{k_a-1}\delta_1^k + \sum_{k=k_a}^{K}\big(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\big)\\
&\le \sum_{k=1}^{k_a-1}\delta_1^k + \big(K-k_a\big)\Big[V_1^*(s_1^{k_a}) - V_1^{\pi_{k_a}}(s_1^{k_a}) + \varepsilon\Big]\\
&= \sum_{k=1}^{k_a-1}\delta_1^k + \big(K-k_a\big)\Bigg[\sum_{h=1}^{H}\Big(\max_{a\in A} Q_h^*(s_1^{k_a},a) - Q_h^{\pi_{k_a}}(s_h^{k_a},a_h^{k_a})\Big) + \varepsilon\Bigg].
\end{aligned}$$

The key then becomes to upper bound the surrogate regret $\sum_{k=1}^{k_a-1}\delta_1^k$ before stability. To do so, we decompose it in a recursive manner for each step h, so as to localize errors, and to upper bound each term in the resulting regret decomposition.

According to Lemma E.3. (see Appendix E), at fixed step $h \in [H]$ over all episodes, the sum of localized surrogate regret across episodes can be upper bounded as follows:

$$\begin{align*}
\sum_{k=1}^{k_a-1}\delta_h^k &\le \sum_{k=1}^{k_a-1}\Bigg[\alpha_t^0 H + \sum_{i=1}^{t}\alpha_t^i\phi_{h+1}^{k_i} + 3U_{2t} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k\Bigg] \tag{*}\\
&\le SAH + \sum_{k=1}^{k_a-1}\sum_{i=1}^{t}\Bigg[\alpha_t^i\phi_{h+1}^{k_i} + 3U_{2t} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k\Bigg]\\
&\le SAH + (1+\frac{1}{2H-1})\sum_{k=1}^{k_a-1}\phi_{h+1}^k + \sum_{k=1}^{k_a-1}\big(3U_{2N_h^k} - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k\big)\\
&= SAH + \frac{1}{2H-1}\sum_{k=1}^{k_a-1}\phi_{h+1}^k + \sum_{k=1}^{k_a-1}\delta_{h+1}^k + \sum_{k=1}^{k_a-1}\big(3U_{2N_h^k} + \xi_{h+1}^k\big)\\
&\le SAH + (1+\frac{1}{2H-1})\sum_{k=1}^{k_a-1}\delta_{h+1}^k + \sum_{k=1}^{k_a-1}\big(3U_{2N_h^k} + \xi_{h+1}^k\big), \tag{**}
\end{align*}$$

where $\xi_{h+1}^k := \big[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\big](s_h^k, a_h^k)$ is a martingale difference sequence related to regret. The first inequality is obtained by directly applying Lemma E.3.

For the second inequality, recall that $t = N_h^k(s_h^k, a_h^k)$ is the number of times that (s_h^k, a_h^k) has been visited at step h at the beginning of k, $\alpha_t^0 = 0$ when $t \geq 1$, and $\alpha_t^0 = 1$ when $t = 0$. Hence, the first term in (*) measures the greatest possible regret accumulated due to initialization when state-action pairs are visited for the first time. It is maximized if each time a new state-action pair is selected until all state-action pairs have been traversed at least once, which can be rigorously denoted with indicator function:

$$\sum_{k=1}^{k_a-1} \mathbb{I}[t=0] \leq \begin{cases} k_a, & \text{if}\, k_a \leq \mathcal{S} \times \mathcal{A} \\ SA, & \text{if}\, k_a > \mathcal{S} \times \mathcal{A} \end{cases},$$

i.e., $\sum_{k=1}^{k_a-1} \mathbb{I}[t=0] \leq SA$. Thus, we have:

$$\textstyle\sum_{k=1}^{k_a-1} \alpha_t^0 H = \sum_{k=1}^{k_a-1} \mathbb{I}[t=0] \cdot H \leq SAH,$$

which gives the second inequality. The second term in (*) then measures the cumulative estimation deviations over episodes, which decay along with time. Recall that $k_i(s_h^k, a_h^k)$ is the episode that (s_h^k, a_h^k) was traversed at step h for the i-th time. Note that for every $k' \in [K]$, $\phi_{h+1}^{k'}$ appears in the later episodes with $k > k'$ if and only if $(s_h^k, a_h^k) = (s_h^{k'}, a_h^{k'})$. For the i-th time it appears, we have $t = N_h^{k'}(s_h^{k'}, a_h^{k'}) + i$. Reorganizing the summation with the fact that $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{2H-1}$ for every $i \geq 1$ (Lemma B.4), we have:

$$\sum_{k=1}^{k_a-1} \sum_{i=1}^{t} \alpha_t^i \phi_{h+1}^{k_i} \leq \sum_{k'=1}^{k_a-1} \phi_{h+1}^{k'} \sum_{t=N_h^{k'}+1}^{\infty} \alpha_t^{N_h^{k'}} \leq (1 + \frac{1}{2H-1}) \sum_{k=1}^{k_a-1} \phi_{h+1}^k,$$

which gives the third inequality. The last inequality holds as $V_{h+1}^*(s_h^k) \geq V_{h+1}^{\pi_k}(s_h^k)$ and thus $\delta_{h+1}^k \geq \phi_{h+1}^k$. With the above recursive form of surrogate regret $\sum_{k=1}^{k_a-1} \delta_h^k$, we can now easily obtain its complexity. In episodic MDP, for $k \in [K], h = H+1$, we have:

$$\delta^k_{H+1} = V^k_{H+1}(s^k_{H+1}) - V^{\pi_k}_{H+1}(s^k_{H+1}) = V^k_{H+1}(s^k_{H+1}) = \min\{H, \max_{a\in\mathcal{A}} Q^k_{H+1}(s^k_{H+1}, a)\} \equiv 0.$$

By writing (**) recursively for $h = 1, 2, \ldots, H$, and let $\mathcal{P} = 1 + \frac{1}{2H-1}$, we have:

$$\begin{aligned}
\sum_{k=1}^{k_a-1} \delta^k_1 &\leq SAH + \mathcal{P}\sum_{k=1}^{k_a-1}\delta^k_2 + \sum_{k=1}^{k_a-1}\big(3U_{2N^k_1} + \xi^k_2\big)\\
&\leq SAH + \mathcal{P}\Bigg[SAH + \mathcal{P}\sum_{k=1}^{k_a-1}\delta^k_3 + \sum_{k=1}^{k_a-1}\big(3U_{2N^k_2} + \xi^k_3\big)\Bigg] + \sum_{k=1}^{k_a-1}\big(3U_{2N^k_1} + \xi^k_2\big)\\
&= (1+\mathcal{P})SAH + \mathcal{P}\sum_{k=1}^{k_a-1}\big(3U_{2N^k_2} + \xi^k_3\big) + \sum_{k=1}^{k_a-1}\big(3U_{2N^k_1} + \xi^k_2\big) + \mathcal{P}^2\sum_{k=1}^{k_a-1}\delta^k_3\\
&\leq \sum_{h=1}^{3}\mathcal{P}^{h-1}SAH + \sum_{h=1}^{3}\mathcal{P}^{h-1}\sum_{k=1}^{k_a-1}\big(3U_{2N^k_h} + \xi^k_{h+1}\big) + \mathcal{P}^3\sum_{k=1}^{k_a-1}\delta^k_4\\
&\ldots\\
&\leq O(H + o(\frac{1}{2H-1}))SAH + \sum_{h=1}^{H} O(1)\sum_{k=1}^{k_a-1}\big(3U_{2N^k_h} + \xi^k_{h+1}\big)\\
&= O\Bigg(H^2SA + \sum_{h=1}^{H}\sum_{k=1}^{k_a-1}\big(3U_{2N^k_h} + \xi^k_{h+1}\big)\Bigg).
\end{aligned}$$

Note that for any $h \in [H]$:

$$\begin{aligned}
\sum_{k=1}^{k_a-1}\sum_{h=1}^{H} 3U_{2N^k_h} &\leq 3\sum_{k=1}^{k_a-1}\sum_{h=1}^{H}\sqrt{\frac{4H^3\ln(2SAT/\delta)}{N^k_h(s^k_h, a^k_h)}} = O(1)\sum_{k=1}^{k_a-1}\sum_{h=1}^{H}\sqrt{\frac{H^3\ln(SAT/\delta)}{N^k_h(s^k_h, a^k_h)}}\\
&= O(1)\sum_{s,a}\sum_{n=1}^{N^{k_a-1}_h(s,a)} H\sqrt{\frac{H^3\ln(SAT/\delta)}{n}}\\
&\leq O(1)\sum_{s,a}\sqrt{N^{k_a-1}_h(s,a)H^5\ln(SAT/\delta)}\\
&\leq O(\sqrt{H^5SAk_a\ln(SAT/\delta)}) = O(\sqrt{H^4SAT_a\ln(SAT/\delta)}).
\end{aligned}$$

The third equality holds as each time when a state-action pair (s,a) is visited in an episode $k \in [K]$,

$N_h^k(s,a)$ is increased by one, and $N_h^{k_a-1}(s,a)$ therefore denotes the total number of times that (s,a) has been traversed over (k_a-1) episodes. Hence summing all visiting times over episodes is equivalent to summing over all entries in $N_h^{k_a-1}$ at terminal episode k_a-1, and increasing a counter for each entry beginning from one to $N_h^{k_a-1}(s,a)$. This is true for all steps h, and it also implies for any fixed $h \in [H]$, we have $\sum_{s,a} N_h^{k_a-1}(s,a) = k_a - 1$ in the final table $N_h^{k_a-1}$. The forth inequality holds as $\sum_{n=1}^{N_h^{k_a-1}(s,a)} \sqrt{\frac{1}{n}} = \sqrt{\frac{1}{1}} + \sqrt{\frac{1}{2}} + \cdot + \sqrt{\frac{1}{N_h^{k_a-1}}} \leq 1 \cdot N_h^{k_a-1}$, which is maximized by $N_h^{k_a-1}(s,a) = (k_a-1)/SA$ for all state-action pairs.

On the other hand, since ξ_{h+1}^k is a martingale difference sequence, by Azuma-Hoeffding inequality, with probability at least $1-\delta$, we have:

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{k_a-1} \xi_{h+1}^k\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{k_a-1} \xi_{h+1}^k \left[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^* - V_{h+1}^{\pi_k})\right](s_h^k, a_h^k)\right| \leq H\sqrt{2T_a \ln(2SAT/\delta)}.$$

Thus, we have:

$$\begin{aligned}\sum_{k=1}^{k_a-1} \delta_1^k &\leq O\left(H^2SA + \sqrt{H^4SAT_a \ln(SAT/\delta)} + \sqrt{2H^2T_a \ln(2SAT/\delta)}\right) \\ &= O\left(H^2SA + \sqrt{H^4SAT_a \ln(SAT/\delta)}\right) \leq O\left(\sqrt{H^4SAT_a \ln(SAT/\delta)}\right).\end{aligned}$$

The last inequality holds since when $T_a \geq \sqrt{H^4SAT_a \ln(SAT/\delta)}$, we have$\sqrt{T_a} \geq \sqrt{H^4SA}$, and thus $\sqrt{H^4SAT_a \ln(SAT/\delta)} \geq H^4SA \geq H^2SA$. On the other hand, if $T_a \leq \sqrt{H^4SAT_a \ln(SAT/\delta)}$, we have $\sum_{k=1}^{k_a-1} \delta_1^k \leq H \cdot k_a \leq T_a \leq \sqrt{H^4SAT_a \ln(SAT/\delta)}$. Thus, H^2SAT can be subsumed in the Big-O notation.

When UCB-HCQ converges, bonus brought by exploration is negligible and it switches to exploitation gradually during evolution. As optimal action at each time step is costly, it becomes infeasible in the remaining time steps, accumulating much more regret compared to UCB-SCQ. Since estimates converge according to Definition 5, linear regret accumulates based on $\max_{a \in A} Q_h^*(s_1^{k_a}, a) - Q_h^{\pi_{k_a}}(s_h^{k_a}, a_h^{k_a})$ for the rest of time.

Putting everything together completes the proof. □

Bibliography

[Abbasi-Yadkori et al., 2019] Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702.

[Achiam et al., 2017] Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31. JMLR. org.

[Altman, 1999] Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

[Amodei et al., 2016] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[Auer, 2002] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

[Azar et al., 2017] Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

[Badanidiyuru et al., 2018] Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018). Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55.

[Bohez et al., 2019] Bohez, S., Abdolmaleki, A., Neunert, M., Buchli, J., Heess, N., and Hadsell, R. (2019). Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*.

[Chow et al., 2017] Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. (2017). Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120.

[Chow et al., 2018] Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. (2018). A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*, pages 8092–8101.

[Chow et al., 2019] Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., and Ghavamzadeh, M. (2019). Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*.

[Dalal et al., 2018] Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y. (2018). Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*.

[Du et al., 2019] Du, S. S., Luo, Y., Wang, R., and Zhang, H. (2019). Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8058–8068.

[Even-Dar and Mansour, 2003] Even-Dar, E. and Mansour, Y. (2003). Learning rates for q-learning. *Journal of machine learning Research*, 5(Dec):1–25.

[Feinberg et al., 2018] Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., and Levine, S. (2018). Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*.

[Ge et al., 2019] Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. (2019). The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, pages 14951–14962.

[Gu et al., 2016] Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. (2016). Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838.

[Hessel et al., 2018] Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. (2018). Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Jin et al., 2018] Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

[Jin et al., 2019] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

[Lai and Robbins, 1985] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

[Levine and Abbeel, 2014] Levine, S. and Abbeel, P. (2014). Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems*, pages 1071–1079.

[Levy and Ermon, 2018] Levy, D. and Ermon, S. (2018). Deterministic policy optimization by combining pathwise and score function estimators for discrete action spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Lötjens et al., 2019] Lötjens, B., Everett, M., and How, J. P. (2019). Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE.

[Lykouris et al., 2019] Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2019). Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*.

[Mnih et al., 2013] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[Mossalam et al., 2016] Mossalam, H., Assael, Y. M., Roijers, D. M., and Whiteson, S. (2016). Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*.

[Roijers et al., 2013] Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113.

[Russo, 2019] Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14410–14420.

[Simchowitz and Jamieson, 2019] Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pages 1151–1160.

[Song and Sun, 2019] Song, Z. and Sun, W. (2019). Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475*.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[Tessler et al., 2018] Tessler, C., Mankowitz, D. J., and Mannor, S. (2018). Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

[Tran-Thanh et al., 2010] Tran-Thanh, L., Chapman, A., de Cote, E. M., Rogers, A., and Jennings, N. R. (2010). Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[Tran-Thanh et al., 2012] Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. R. (2012). Knapsack based optimal policies for budget–limited multi–armed bandits. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[Van Hasselt et al., 2016] Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

[Wainwright, 2019] Wainwright, M. J. (2019). Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

[Wen and Topcu, 2018] Wen, M. and Topcu, U. (2018). Constrained cross-entropy method for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7450–7460.

[Wu et al., 2018] Wu, D., Chen, X., Yang, X., Wang, H., Tan, Q., Zhang, X., Xu, J., and Gai, K. (2018). Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1443–1451.

[Yang and Wang, 2019a] Yang, L. and Wang, M. (2019a). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.

[Yang and Wang, 2019b] Yang, L. F. and Wang, M. (2019b). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.

[Yu et al., 2019] Yu, M., Yang, Z., Kolar, M., and Wang, Z. (2019). Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3121–3133.