Sample Complexity Bounds for Constrained Markov Decision Processes with Linear Function Approximation

by

Xingtu Liu

B.Math., University of Waterloo, 2022

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2025

Declaration of Committee

Name:	**Xingtu Liu**
Degree:	**Master of Science**
Thesis title:	**Sample Complexity Bounds for Constrained Markov Decision Processes with Linear Function Approximation**
Committee:	**Chair:** Andrei Bulatov Professor, Computing Science
	Sharan Vaswani Supervisor Assistant Professor, Computing Science
	Oliver Schulte Committee Member Professor, Computing Science
	Hang Ma Examiner Assistant Professor, Computing Science

Abstract

We consider infinite-horizon γ-discounted (linear) constrained Markov decision processes (CMDPs) where the objective is to find a policy that maximizes the expected cumulative reward subject to expected cumulative constraints. Given access to a generative model, we propose to solve CMDPs with a primal-dual framework that can leverage any black-box unconstrained MDP solver. For linear CMDPs with feature dimension d, we instantiate the framework by using mirror descent value iteration (`MDVI`) an example MDP solver. We provide sample complexity bounds for the resulting CMDP algorithm in two cases: (i) relaxed feasibility, where small constraint violations are allowed, and (ii) strict feasibility, where the output policy is required to exactly satisfy the constraint. For (i), we prove that the algorithm can return an ε-optimal policy with high probability by using $\tilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right)$ samples. For (ii), we show that the algorithm requires $\tilde{O}\left(\frac{d^2}{(1-\gamma)^6\varepsilon^2\zeta^2}\right)$ samples, where ζ is the problem-dependent Slater constant that characterizes the size of the feasible region. Furthermore, we prove a lower-bound of $\Omega\left(\frac{d^2}{(1-\gamma)^5\varepsilon^2\zeta^2}\right)$ for the strict feasibility setting.

Keywords: Reinforcement Learning; Sample Complexity; Constrained MDPs; Linear Function Approximation

Acknowledgements

I am profoundly grateful to the many individuals whose guidance and support have been indispensable throughout my research journey and the completion of this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor, Sharan Vaswani, for his unwavering guidance, insightful feedback, and constant support. The journey has been truly remarkable—it's inspiring to see how the discussions we had during my application interview have come to fruition in this work. I am also sincerely thankful to my supervisor and collaborator, Lin F. Yang, for his invaluable support and thoughtful feedback throughout this process.

I am deeply indebted to my parents, whose unconditional love, faith, and endless encouragement have been my foundation. Thank you for always standing by me, for every word of support, and for instilling in me the belief that I could achieve my goals.

Finally, I wish to extend my heartfelt thanks to my undergraduate mentors, Gautam Kamath and Huanyu Zhang. Their early guidance introduced me to academia and the field of machine learning theory. Their mentorship set me on this path, and for that, I remain deeply grateful.

This accomplishment would not have been possible without all of your unwavering support and sacrifices.

Preface

The main matter of this thesis is based on the paper that is under review and available as a preprint:

Xingtu Liu, Lin F. Yang, Sharan Vaswani. Sample Complexity Bounds for Linear Constrained MDPs with a Generative Model. arXiv preprint arXiv:2507.02089, 2025.

Xingtu Liu is the major contributor on this paper in terms of the writing and theoretical results. The work was supervised by Sharan Vaswani and Lin F. Yang who constantly discussed and helped in refining all aspects of the paper.

A previous version of the work was published in NeurIPS 2025 Workshop on Constrained Optimization for Machine Learning:

Xingtu Liu, Lin F. Yang, Sharan Vaswani. Sample Complexity Bounds for Linear Constrained MDPs with a Generative Model. NeurIPS 2025 Workshop on Constrained Optimization for Machine Learning.

Table of Contents

Chapter 1

Introduction

Reinforcement learning (RL) [Sutton et al., 1998] is a machine learning paradigm aimed at building learning agents capable of making sequential decisions in an (unknown) environment. RL algorithms have found applications in games such as Atari [Mnih et al., 2015] or Go [Silver et al., 2016], robot manipulation tasks [Tan et al., 2018, Zeng et al., 2020], clinical trials [Schaefer et al., 2005] and more recently, aligning large language models to human preferences [Ouyang et al., 2022, Shao et al., 2024]. Typical RL algorithms only focus on optimizing an unconstrained objective, although in many real-world applications, agents are often required to not only maximize cumulative rewards but also to satisfy constraints imposed by safety, fairness, or resource usage. RL with such side-constraints is typically formulated within the framework of constrained Markov decision processes (CMDPs) [Altman, 1999], where the goal is to optimize an expected reward function while ensuring that the expected cumulative cost (or utility) satisfies a given threshold. For example, in wireless sensor networks [Buratti et al., 2009, Julian et al., 2002], the agent aims to deploy a policy that maximizes the bitrate with a constraint on its average power consumption.

Given the practical importance of constrained RL, there is a vast literature [Brantley et al., 2020, Ding et al., 2021, Efroni et al., 2020, Gattami et al., 2021, Kalagarla et al., 2021, Miryoosefi and Jin, 2022, Mondal and Aggarwal, 2024, Qiu et al., 2020, Yu et al., 2021, Zheng and Ratliff, 2020] that aims to obtain a near-optimal policy in unknown tabular CMDPs with finite states and actions. These works simultaneously tackle the exploration, estimation and planning problems and aim to minimize the regret and constraint violation in the online setting. On the other hand, recent works [Bai et al., 2021, HasanzadeZonuzy et al., 2021, Vaswani et al., 2022, Wei et al., 2021] consider an easier, but even more fundamental problem of obtaining a near-optimal policy with access to a simulator or *generative model* [Agarwal et al., 2020, Kakade, 2003, Kearns and Singh, 1999, Sidford et al., 2018, Yang and Wang, 2019]. In particular, these works assume that the agent has access to a sampling oracle (the generative model) that returns a sample of the next state when given any state-action pair as input.

Depending on the application of interest, such a generative model is often available either directly for the task at hand (for example, in Atari games where the aim is to win the game) or as an proxy to the task (for example, the CARLA simulator [Dosovitskiy et al., 2017] for training autonomous vehicles). Moreover, from a theoretical perspective, since the generative model setting removes the need for exploration it has been used to characterize the statistical complexity of obtaining near-optimal policies for (C)MDPs [Agarwal et al., 2020, Azar et al., 2013, Li et al., 2020, Vaswani et al., 2022]. In online RL, agents must do exploration, which means trying new actions to gather information about the unknown environment. However, in the generative model setting, the agent can query a simulator for any state-action pair, effectively removing the need for this costly and potentially unsafe real-world exploration. This “planning with a simulator" setting is naturally analyzed within the Probably Approximately Correct (PAC) framework. PAC-RL provides formal guarantees on an algorithm’s learning outcome. Instead of minimizing regret (mistakes made during online learning), the goal of PAC-RL is to use a finite number of samples to identify a policy that is approximately optimal with high probability. This framework is meaningful and important because it provides statistical confidence in a policy’s performance before it is deployed. The central measure of efficiency in the PAC framework is sample complexity. In machine learning, this term generally refers to the number of training examples an algorithm needs to learn a concept or function to a desired level of accuracy. In PAC-RL, sample complexity specifically measures the number of samples (i.e., queries to the generative model) required to find a probably approximately correct policy. Analyzing the sample complexity is important because it quantifies the fundamental data-efficiency of an algorithm. It allows for a rigorous, theoretical comparison between different methods and reveals the inherent statistical difficulty of a learning problem.

For unconstrained MDPs, the linear MDP assumption (e.g., [Jin et al., 2020, Yang and Wang, 2019]) is a common formalization to analyze algorithms that have access to state-action features and can incorporate linear function approximation. The assumption implies that both the rewards and transition probabilities (approximately) lie in the span of the given d-dimensional feature representation, and can be used to obtain sample complexity bounds independent of the size of the state-action space. Intuitively, the linear MDP assumption means that the environment’s dynamics and rewards are not arbitrarily complex. It is one of the simplest function approximation assumptions that makes the resulting algorithms amenable to analysis. This assumption is beneficial for several reasons. First, it makes large or infinite state-space problems computationally and statistically tractable, as the problem’s complexity now scales with the feature dimension d rather than the enormous number of states. Second, it aligns with many real-world applications where complex states (e.g., a robot’s sensor readings) can be effectively summarized by a compact feature vector. It thus provides a powerful theoretical framework that bridges the gap between the overly simple tabular case and the intractability of general function approximation.

A core goal of this thesis is to establish such sample complexity bounds for linear CMDPs, thereby characterizing how problem-specific factors—like the *feature dimension*, *effective horizon*, and the *size of the feasible region*—impact the amount of data required to find a near-optimal and feasible policy.

1.1 Related Work

For CMDPs, Vaswani et al. [2022] established near-optimal upper and lower-bounds on the sample complexity in two settings: (i) relaxed feasibility, where small constraint violations are allowed, and (ii) strict feasibility, where the output policy is required to exactly satisfy the constraint. For tabular CMDPs, the proposed algorithms and resulting bounds depend on the cardinality of the state-action space, and hence do not apply to modern applications involving large or infinite state spaces. Consequently, it is essential to develop provably efficient algorithms that can incorporate function approximation and go beyond the tabular case.

Unconstrained linear MDPs have been extensively studied in the context of both finite-horizon regret minimization [Hu et al., 2022, Jin et al., 2020, Liu et al., 2023, Sherman et al., 2023, Weisz et al., 2022] and with access to a generative model [Kitamura et al., 2023, Taupin et al., 2023]. Following the linear MDP literature, recent works consider CMDPs with linear function approximation [Ding et al., 2021, Ghosh et al., 2022, 2024, Jain et al., 2022, Liu et al., 2022, Miryoosefi and Jin, 2022, Tian et al., 2024] and assume that (in addition to the rewards and transition probabilities), the costs or utilities can also be expressed using the given features. However, all previous work on linear CMDPs considers the online regret minimization setting and the statistical complexity of the problem remains unclear.

1.2 Contributions

Motivated by Vaswani et al. [2022], we aim to *study the sample complexity of solving linear CMDPs with access to a generative model.* In particular, we make the following contributions.

(1) Generic primal-dual algorithm framework: In chapter 3, we provide a generic primal-dual algorithmic framework (algorithm 1) that can be used to achieve both the *relaxed* and *strict* feasibility objectives, for both *tabular* and *linear* CMDPs. As model-based approaches [Vaswani et al., 2022] are not applicable in the linear CMDP setting, algorithm 1 is designed to be model-free and relies on three black-box subroutines: a `DataCollection` procedure, a black-box `MDP-Solver` and a `PolicyEvaluation` oracle. We prove a meta-theorem (theorem 7) to quantify the sample complexity of algorithm 1 in terms of that of the `MDP-Solver` and `PolicyEvaluation` oracle.

(2) Instantiating the framework for linear CMDPs: In section 4.2, we instantiate the linear `MDP-Solver` with a variant of the mirror-descent value iteration (`MDVI`) algorithm [Kitamura et al., 2023, Kozuno et al., 2022]. In contrast to the existing `MDVI` variants, the proposed algorithm 2 does not use entropy regularization and outputs a stationary policy, thus simplifying the algorithm design. We develop a new theoretical analysis for algorithm 2 and characterize its sample complexity for solving unconstrained linear MDPs. In section 4.3, we instantiate the `PolicyEvaluation` oracle with least-squares policy evaluation (algorithm 3) and analyze the sample complexity required to evaluate the performance of a (data-dependent) policy.

(3) Upper-bound on sample complexity for linear CMDPs: In section 4.5, we leverage our meta-theorem and analyze the sample complexity for the resulting CMDP algorithm that uses algorithms 2 and 3. In particular, if d is the dimension of the feature mapping, we prove that the proposed algorithm requires no more than $\tilde{O}\left(\frac{d^2}{(1-\gamma)^4\varepsilon^2}\right)$ samples to obtain an ε-optimal policy in the relaxed feasibility setting. Since the lower-bound on the sample complexity for solving unconstrained linear MDP is $\Omega\left(\frac{d^2}{(1-\gamma)^3\varepsilon^2}\right)$ [Weisz et al., 2022], our sample complexity achieves the near-optimal dependence on d and ε, and is away from the lower bound by atmost a multiplicative factor of $\tilde{O}\left(1/1-\gamma\right)$. Under strict feasibility, our algorithm requires no more than $\tilde{O}\left(\frac{d^2}{(1-\gamma)^6\varepsilon^2\zeta^2}\right)$ samples, where ζ is the problem-dependent Slater constant that characterizes the size of the feasible region and dictates the difficulty of the problem. Given the lower-bounds for tabular CMDPs in Vaswani et al. [2022], we conjecture that the dependence on d, ε, and ζ in our bounds is tight, with suboptimality arising only in the multiplicative dependence on $O(1/1-\gamma)$. To the best of our knowledge, *these are the first such sample complexity bounds with the near-optimal dependence on both d and ε* . In section C.5, we alternatively instantiate the linear `MDP-Solver` to be the G-Sampling-and-Stop (`GSS`) algorithm [Taupin et al., 2023] and analyze the sample complexity of the resulting CMDP algorithm, thus demonstrating the flexibility of our framework.

(4) Lower-bound on sample complexity for linear CMDPs: In section 4.5, we prove a problem-dependent $\Omega\left(\frac{d^2}{(1-\gamma)^5\varepsilon^2\zeta^2}\right)$ lower-bound on the sample-complexity in the strict feasibility setting. Our results thus demonstrate that the proposed algorithm is near-optimal in terms of d, ε, and ζ, with a suboptimality only in the multiplicative dependence on H. The lower bound also indicates that under strict feasibility solving linear CMDPs is inherently more difficult than solving unconstrained linear MDPs, and the problem difficulty increases as the size of the feasible region (measured in terms of ζ) decreases. To the best of our knowledge, it is the first result characterizing the difficulty of solving linear CMDPs with access to a generative model.

(5) Sample complexity bounds for tabular CMDPs: Finally, in chapter F, we utilize our framework for tabular CMDPs. In particular, we instantiate algorithm 1 with tabular variants of algorithms 2 and 3 (obtained by setting $d = SA$ and considering one-hot features) and analyze the resulting CMDP algorithm. Under the relaxed and strict feasibility settings, the resulting algorithm attains sample complexity bounds of $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ and $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2\zeta^2}\right)$, respectively. These results match the near-optimal bounds attained by the model-based algorithm in Vaswani et al. [2022], and improve upon the sample-complexity of the model-free approach proposed in Bai et al. [2021].

Chapter 2

Problem Formulation

An infinite-horizon discounted constrained tabular Markov decision process (CMDP) [Altman, 1999] is denoted by $\mathcal{M}$, and is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, b, \rho, \gamma \rangle$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the action set, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $\rho \in \Delta_{\mathcal{S}}$ is the initial distribution of states and $\gamma \in [0, 1)$ is the discount factor. The primary reward to be maximized is denoted by $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, whereas the constraint reward is denoted by $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$[1]. If $\Delta_{\mathcal{A}}$ denotes the simplex over the action space, the expected discounted return or *reward value function* of a stationary, stochastic policy[2] $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is defined as $V_r^\pi(\rho) = \mathbb{E}_{s_0, a_0, \ldots} \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \Big]$, where $s_0 \sim \rho, a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. For each state-action pair (s, a) and policy π, the reward action-value function is defined as $Q_r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and satisfies the relation: $V_r^\pi(s) = \langle \pi(\cdot|s), Q_r^\pi(s, \cdot) \rangle$, where $V_r^\pi(s)$ is the reward value function when the starting state is equal to s. Analogously, the *constraint value function* and constraint action-value function of policy π is denoted by $V_c^\pi(\rho)$ and Q_c^π respectively. Throughout, it will be convenient to present our results in terms of the effective horizon $H := {}^{1}\!/\!{}_{(1-\gamma)}$.

In addition to the tabular CMDPs with a finite state-action space, we also consider linear [Jin et al., 2020] CMDPs where the state space can be large or possibly infinite. In this case, we assume access to a feature representation ϕ such that r, c and the transition probabilities $\mathcal{P}$ (approximately) lie in the span of the given d-dimensional feature representation.

Assumption 1 (Linear Constrained MDP)**.** *For the CMDP $\mathcal{M}$ with the state-action space $\mathcal{S} \times \mathcal{A}$, we have access to a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ that satisfies the following condition: there exist vectors $\psi_r, \psi_c \in \mathbb{R}^d$ and signed measures $\mu := (\mu_1, \ldots, \mu_d)$ on $\mathcal{S}$ such that $P(\cdot|s, a) = \langle \phi(s, a), \mu \rangle$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $r = \langle \phi, \psi_r \rangle$, and $c = \langle \phi, \psi_c \rangle$. Let*

[1] These ranges for r and c are chosen for simplicity. Our results can be easily extended to handle other ranges.

[2] The performance of an optimal policy in a CMDP can always be achieved by a stationary, stochastic policy [Altman, 1999]. On the other hand, for an MDP, it suffices to only consider stationary, deterministic policies [Puterman, 2014].

$\Phi := \{\phi(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\} \subset \mathbb{R}^d$ *be the set of all feature vectors. We assume that* Φ *is compact and spans* $\mathbb{R}^d$*.*

The objective is to return a policy that maximizes $V_r^\pi(\rho)$, while ensuring that $V_c^\pi(\rho) \geq b$. Formally,

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad V_c^\pi(\rho) \geq b. \tag{2.1}$$

The optimal stochastic policy for the above CMDP is denoted by π^* and the corresponding reward value function is denoted by $V_r^*(\rho)$. We also define $\zeta := \max_\pi V_c^\pi(\rho) - b > 0$ as the problem-dependent quantity referred to as the Slater constant [Bai et al., 2021, Ding et al., 2021]. The Slater constant is a measure of the size of the feasible region and determines the difficulty of solving eq. (2.1).

For simplicity of exposition, we assume that the rewards r and constraint rewards c are known, but the transition probabilities $\mathcal{P}$ are unknown. We note that assuming the knowledge of the rewards does not affect the leading terms of the sample complexity since learning these is an easier problem [Azar et al., 2013, Sidford et al., 2018]. Following Azar et al. [2013], Vaswani et al. [2022], we assume access to a *generative model* or simulator that allows the agent to obtain samples from the $\mathcal{P}(\cdot|s,a)$ distribution for any (s,a).

Definition 2 (Generative Model)**.** *A generative model* Gen *for an MDP is an oracle that, given any state-action pair* (s,a)*, returns an independent sample of the next state* $s' \sim P(\cdot \mid s,a)$*.*

Assuming access to such a generative model, we aim to characterize the sample complexity (number of times Gen is queried) required to return a near-optimal policy $\bar{\pi}$. Specifically, given a target error $\varepsilon > 0$, we consider two different definitions of optimality.

Relaxed feasibility: We require $\bar{\pi}$ to achieve an approximately optimal reward value, while allowing it to have a small constraint violation. Formally, we aim to find a $\bar{\pi}$ such that,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - \varepsilon \quad \text{and} \quad V_c^{\bar{\pi}}(\rho) \geq b - \varepsilon. \tag{2.2}$$

Strict feasibility: We require $\bar{\pi}$ to achieve an approximately optimal reward value, while simultaneously demanding zero constraint violation. Formally,we aim to find a $\bar{\pi}$ such that,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - \varepsilon \quad \text{and} \quad V_c^{\bar{\pi}}(\rho) \geq b. \tag{2.3}$$

In the next section, we design a generic algorithmic framework to achieves these objectives.

Chapter 3

A Generic Framework for Solving CMDPs

We first present a generic primal-dual algorithmic framework for solving CMDPs, and subsequently present a meta-theorem that quantifies its sample-complexity in the relaxed and strict feasibility settings. For this, we frame the CMDP problem in eq. (2.1) as an equivalent saddle-point problem,

$$\max_{\pi} \min_{\lambda \geq 0} \left[V_r^{\pi}(\rho) + \lambda \left(V_c^{\pi}(\rho) - b \right) \right] , \tag{3.1}$$

where, λ is the Lagrange multiplier. The solution to eq. (3.1) is (π^*, λ^*) where π^* is the optimal policy to the CMDP and λ^* is the optimal Lagrange multiplier. We solve eq. (3.1) iteratively, by alternatively updating the policy (primal variable) and the Lagrange multiplier (dual variable) [Ding et al., 2021, Vaswani et al., 2022].

Algorithm 1 Primal-dual CMDP framework with a generative model

Input: r (rewards), c (constraint rewards), b' (constraint RHS), U (projection upper bound), K (number of iterations), η (step-size), $\lambda_0 = 0$ (initialization), Gen (generative model), $\mathcal{C}$ (subset of $\mathcal{S} \times \mathcal{A}$), N (sample size for each (s, a) pair in $\mathcal{C}$), ϕ (feature map).

Output: Mixture policy $\bar{\pi} = \frac{1}{K} \sum_{k=0}^{K-1} \pi_k$.

```
procedure CMDPF($r, c, b', U, K, \eta, \mathsf{Gen}, \mathcal{C}, N, \phi$)
    $\mathcal{B}$ = DataCollection($\mathsf{Gen}, \mathcal{C}, N$).    ▷ Data collection procedure to populate buffer
    for $k = 0, \ldots, K-1$ do
        Let $\pi_k$ = MDP-Solver($r + \lambda_k c, \mathcal{B}, \phi$)    ▷ Updating the primal variable
        Let $\hat{V}_c^k$ = PolicyEvaluation($\pi_k, c, \mathcal{B}, \phi$)    ▷ Policy Evaluation
        $\lambda_{k+1} = \mathbb{P}_{[0,U]} \left[ \lambda_k - \eta \, (\hat{V}_c^k(\rho) - b') \right]$.    ▷ Updating the dual variable
    end for
end procedure
```

Remark 3. *The model-based framework for solving tabular CMDPs in Vaswani et al. [2022] does not directly extend to the linear MDP setting for the following reasons. In the PAC-RL framework (with access to a simulator), model-based methods build a model for the transition matrix by sampling next states for each state-action pair. In the linear (C)MDP setting, the transition kernel and reward functions are assumed to be approximated by a linear function. Specifically, the $SA \times S$ transition matrix can be factorized into a known $SA \times d$ matrix of features and an unknown $d \times S$ matrix to be estimated. Hence, naively estimating the transition matrix will require $O(S)$ samples, resulting in high sample complexity.*

The primal and dual updates in algorithm 1 rely on three oracles, which we instantiate subsequently.

Data Collection Oracle: We first describe the mechanism of the `DataCollection` oracle in algorithm 1. This oracle takes as input a generative model Gen, a subset of state-action pairs $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{A}$, and a sample size N. For each $(s, a) \in \mathcal{C}$, it queries the generative model Gen to obtain N independent next-state samples $(s'_i)_{i=1}^N$ from the distribution $\mathsf{Gen}(\cdot \mid s, a)$. It then stores the resulting triplets $(s, a, s'_i)_{i=1}^N$ in a buffer $\mathcal{B}$. After all state-action pairs in $\mathcal{C}$ are processed, the buffer $\mathcal{B}$ contains N samples for each pair and is returned as the output.

MDP-Solver: The primal update at iteration k uses the `MDP-Solver`, which takes as input a buffer $\mathcal{B}$ of samples and returns a policy π_k satisfying the following assumption.

Assumption 4. *We have access to a black-box algorithm `MDP-Solver`$(\square, \mathcal{B}, \phi)$ for which the inputs are the feature map ϕ and an arbitrary but bounded reward function $\square \in [0, R]$, and the output is a policy $\tilde{\pi}$ satisfying the following condition with probability $1 - \delta$,*

$$\max_{\pi} V_{\square}^{\pi}(\rho) - V_{\square}^{\tilde{\pi}}(\rho) \leq R \, f_{\text{mdp}}(\mathcal{B}) \, ,$$

where, $f_{\text{mdp}}(\mathcal{B})$ denotes an upper bound on the sub-optimality when given access to buffer $\mathcal{B}$.

Policy Evaluation Oracle: The dual update at iteration k in algorithm 1 is given as:

$$\lambda_{k+1} = \mathbb{P}_{[0,U]} \left[\lambda_k - \eta \, (\hat{V}_c^k(\rho) - b') \right] \, ,$$

where $\mathbb{P}_{[0,U]}$ denotes the projection onto the interval $[0, U]$, and b' is a relaxed constraint parameter that depends on b, f_{mdp} and the problem setting (relaxed or strict). The term $\hat{V}_c^k$ is an estimate of $V_c^{\pi_k}$, computed via the `PolicyEvaluation` oracle which satisfies following assumption.

Assumption 5. *We have access to a black-box algorithm `PolicyEvaluation`$(\pi, \diamond, \mathcal{B}, \phi)$ for which the inputs are a possibly data-dependent (one that depends on the buffer $\mathcal{B}$) policy π, the feature map ϕ, a reward function $\diamond \in [0, 1]$, and the output is a value function $\hat{V}_{\diamond}$ satisfying the following condition with probability $1 - \delta$,*

$$|\hat{V}_{\diamond}(\rho) - V_{\diamond}^{\pi}(\rho)| \leq f_{\text{eva}}(\mathcal{B}) \, ,$$

where, $f_{\text{eva}}(\mathcal{B})$ denotes an upper bound on the sub-optimality when given access to buffer $\mathcal{B}$.

Remark 6. *We note that while there are existing methods that update the dual variable using empirical utility value functions obtained from the MDP solver, this approach does not result in near-optimal statistical guarantees. This is because the MDP solver can only imply a concentration guarantee on the value function for the combined reward function $r + \lambda_k c$ and not for the individual value functions V_r and V_c. Proving near-optimal statistical guarantees necessitates the development of the proposed policy evaluation subroutine. Alternative ways to handle such concentration include using a uniform concentration bound [Ghosh et al., 2022] (i.e., building an ε-cover over the function class and using a union bound). However, such an analysis results in a sub-optimal $O(d^3)$ dependence on the dimension. Furthermore, in order to guarantee concentration, Ghosh et al. [2022] needs to guarantee Lipschitzness in the policy update. Our technique does not require this, and consequently, we can use greedy policy updates that are not Lipschitz.*

After K iterations of primal and dual updates, algorithm 1 returns a mixture policy $\bar{\pi}$ which is a policy drawn uniformly at random from the set $\{\pi_0, \dots, \pi_{K-1}\}$. Given access to these oracles, we state a meta-theorem (proved in chapter B) to characterize the sub-optimality of the algorithm.

Theorem 7. *Suppose theorems 4 and 5 hold and let $f(\mathcal{B}) := \max\{f_{\text{mdp}}(\mathcal{B}), f_{\text{eva}}(\mathcal{B})\}$. For $\delta \in (0,1)$, algorithm 1 with $U = \frac{2}{\zeta(1-\gamma)}$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = \frac{U^2}{[f(\mathcal{B})]^2(1-\gamma)^2}$ and $b' = b - 2f(\mathcal{B})$, returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - 4f(\mathcal{B}) \quad , \quad V_c^{\bar{\pi}}(\rho) \geq b - 6f(\mathcal{B}). \quad (\textbf{\textit{Relaxed Feasibility Setting}})$$

With the same algorithm parameters, but with $b' = b + 4f(\mathcal{B})$ for $f(\mathcal{B}) \leq \frac{\zeta}{6}$, algorithm 1 returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - \frac{16 f(\mathcal{B})}{\zeta(1-\gamma)} \quad , \quad V_c^{\bar{\pi}}(\rho) \geq b. \quad (\textbf{\textit{Strict Feasibility Setting}})$$

The above theorem implies that, provided we can adequately control the terms $f_{\text{mdp}}(\mathcal{B})$ and $f_{\text{eva}}(\mathcal{B})$ via the three oracle procedures, both the relaxed feasibility condition (2.2) and the strict feasibility conditions (2.3) can be satisfied. Furthermore, we note that similar to [Vaswani et al., 2022], the error for the strict feasibility setting is inflated by an $O\left(\frac{1}{\zeta\,(1-\gamma)}\right)$ factor.

Hence, in the next section, we instantiate the subroutines `DataCollection`, `MDP-Solver` and `PolicyEvaluation` such that the quantities $f_{\text{mdp}}(\mathcal{B})$ and $f_{\text{eva}}(\mathcal{B})$ are sufficiently small.

Chapter 4

Instantiating the Framework for Linear Constrained MDPs

We first describe the construction of the coreset $\mathcal{C}$, which serves as input to the `DataCollection` procedure. We then introduce a model-free algorithm, `LS-MDVI`, as an instantiation of the `MDP-Solver`. Finally, we present `LS-PE`, which serves as the instantiation of the `PolicyEvaluation` subroutine.

4.1 Data Collection via Core Set Construction

Recall that the `DataCollection` procedure requires as input a subset of $\mathcal{S} \times \mathcal{A}$. In the linear setting, we provide a coreset $\mathcal{C}$ as this input. We now describe the construction of the coreset [Kitamura et al., 2023, Lattimore et al., 2020]. The key properties of the coreset are that it has few elements (independent of the cardinality of $\mathcal{S}$ and $\mathcal{A}$), while the features corresponding to the $(x, b) \in \mathcal{C}$ provide a good coverage of the feature space. For a distribution $\tilde{\rho}$ over $\mathcal{S} \times \mathcal{A}$, let $G \in \mathbb{R}^{d \times d}$ and $g(\tilde{\rho}) \in \mathbb{R}$ be defined as:

$$G := \sum_{(x,b) \in \mathcal{C}} \tilde{\rho}(x, b)\, \phi(x, b)\phi(x, b)^\top \qquad \text{and} \qquad g(\tilde{\rho}) := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \langle \phi(s, a), G^{-1}\phi(s, a) \rangle.$$

We refer to $\tilde{\rho}$ as the design, G as the corresponding design matrix, and define the coreset of $\tilde{\rho}$ as its support, $\mathcal{C} := \mathrm{Supp}(\tilde{\rho})$. The task of identifying a design that minimizes g is known as the *G-optimal design problem*. We assume that we can construct near-optimal experimental design.

Assumption 8 (Optimal Design). *We have access to an oracle called `ComputeOptimalDesign` which returns $\tilde{\rho}$, $\mathcal{C}$ and G such that $g(\tilde{\rho}) \leq 2d$ and the coreset of $\tilde{\rho}$ has size at most $\tilde{O}(d)$.*

Such a design can be obtained using the Frank-Wolfe algorithm [Todd, 2016] described in chapter A. Accordingly, we first use the `ComputeOptimalDesign` procedure to construct $\tilde{\rho}$, $\mathcal{C}$, and the associated design matrix G, and then utilize the resulting coreset $\mathcal{C}$ to collect data. For each state-action pair in $\mathcal{C}$, we collect N independent samples and store them in the buffer $\mathcal{B}$. Hence, the total sample complexity is $N\,|\mathcal{C}|$. In the subsequent section, it is convenient to consider $\mathcal{B}$ as a union of T disjoint subsets $B_0 \cup \cdots \cup B_{T-1}$, where each B_i consists of M independent samples for every state-action pair in $\mathcal{C}$. Consequently, we have $N = T\,M$.

4.2 Instantiating the MDP-Solver: Least-Squares Mirror Descent Value Iteration

We now introduce a model-free algorithm referred to as least-squares mirror descent value iteration (LS-MDVI) which serves as an instantiation of the MDP-Solver.

LS-MDVI is a generalization of MDVI [Geist et al., 2019, Kozuno et al., 2022, Vieillard et al., 2020] to the linear function approximation setting and is related to the algorithm proposed in Kitamura et al. [2023]. In particular, LS-MDVI corresponds to a limiting case of policy mirror descent [Lan, 2023] when the KL regularization tends to zero (or equivalently, the step-size tends to infinity). This results in a value iteration method which we describe below.

Define $\mathcal{H}(\pi(\cdot|s))$ as the entropy of the policy π in state s and $\mathrm{KL}(\pi(\cdot|s)||\pi'(\cdot|s))$ as the KL divergence between policies $\pi(\cdot|s)$ and $\pi'(\cdot|s)$ in state s. With a slight abuse of notation, we consider π to be an operator such that $(\pi Q)(s) := \sum_{a\in\mathcal{A}} \pi(a|s)Q(s,a)$. At iteration $t \in [T]$, LS-MDVI requires the corresponding action-value function to update the policy. Specifically, if τ is the strength of the KL regularization and κ is the entropy regularization coefficient s.t. $\alpha = \frac{\tau}{\tau+\kappa}$, $\beta = \frac{1}{\tau+\kappa}$, given Q^{t+1} for some reward function, the entropic mirror descent and LS-MDVI updates can be written as:

$$\textbf{Entropic Mirror Descent}: \pi_{t+1}(a|s) \propto [\pi_t(a|s)]^{\alpha} \exp\left(\beta Q^{t+1}(s,a)\right),$$

$$V^{t+1}(s) = (\pi_{t+1}Q^{t+1})(s) - \tau \mathrm{KL}(\pi_{t+1}(\cdot|s)\|\pi_t(\cdot|s)) + \kappa\mathcal{H}(\pi_{t+1}(\cdot|s)).$$

$$\textbf{LS-MDVI}: \pi_{t+1}(\cdot|s) = \arg\max_a \sum_{i=0}^{t+1} Q^i(s,a)\ ;\ V^{t+1}(s) = \Big(\pi_{t+1}\sum_{i=0}^{t+1} Q^i\Big)(s) - \Big(\pi_t \sum_{i=0}^{t} Q^i\Big)(s).$$

Starting from entropic mirror descent, for $\kappa = 0$ and as $\tau \to 0$, implying $\alpha = 1$, we recover the LS-MDVI update (see [Kozuno et al., 2022, App. B] for the derivation).

Remark 9. *In contrast, Kitamura et al. [2023] consider both $\kappa \to 0$, $\tau \to 0$ while keeping α fixed and effectively consider an entropy-regularized update. This proposed change simplifies the algorithm design for LS-MDVI. Furthermore, while the algorithm in Kitamura et al. [2023] produces non-stationary policies, LS-MDVI outputs a stationary policy.*

Next, we present algorithm 2 which implements the above LS-MDVI update, but uses the linear CMDP structure and the data collected in the buffer $\mathcal{B}$ to estimate Q^{t+1}. Specifically,

θ^{t+1} estimation in algorithm 2 corresponds to the Q^{t+1} estimation using linear regression and the last line in algorithm 2 corresponds to the above update. Similar to approximate value iteration, the $\hat{Q}^{t+1}$ update depends on $\hat{V}^t$ via the Bellman equation, however, π_{t+1} depends on $\tilde{Q}^{t+1}$, the "soft" Q function formed by using the estimates up to iteration $t+1$.

Algorithm 2 Least-Squares Mirror Descent Value Iteration (`LS-MDVI`)

Input: T (number of iterations), M (number of next-state samples obtained per state-action pair in each iteration), $\square$ (rewards in MDP), $\mathcal{B} = \mathcal{B}_0 \cup \dots \cup \mathcal{B}_{T-1}$ (Buffer), $\tilde{\rho}$ (design), $\mathcal{C}$ (coreset),
ϕ (feature map).

Output: π_T where $\forall s \in \mathcal{S}$, $\pi_T(\cdot|s) \in \arg\max_a \tilde{Q}^T_\square(s,a)$.

Define $\hat{V}^0_\square = \mathbf{0}$, $\theta^0_\square = \mathbf{0}$.

```
procedure LS-MDVI($T$, $M$, $\square$, $\mathcal{B}$, $\tilde{\rho}, \mathcal{C}, \phi$)
    for $t = 0, 1, 2 \dots, T-1$ do
        $\forall (s,a) \in \mathcal{C}$ : Access $(s, a, s'_m)_{m=1}^M$ from the buffer $\mathcal{B}_t$.
        Define regression target $\hat{Q}^{t+1}_\square(s,a) := \square(s,a) + \gamma \frac{1}{M} \sum_{m=1}^M \hat{V}^t_\square(s'_m)$.
        $\theta^{t+1}_\square = \arg\min_{\theta \in \mathbb{R}^d} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)(\langle \phi(x,b), \theta \rangle - \hat{Q}^{t+1}_\square(x,b))^2$
        Define $\tilde{Q}^{t+1}_\square := \langle \phi, \sum_{i=0}^{t+1} \theta^i_\square \rangle$ ; $\hat{V}^{t+1}_\square(s) := \max_a \left\{ \tilde{Q}^{t+1}_\square(s,a) \right\} - \max_a \left\{ \tilde{Q}^t_\square(s,a) \right\}$
    end for
end procedure
```

In each iteration $t \in [T]$, algorithm 2 uses the buffer B_t consisting of M samples per state-action pair in $\mathcal{C}$. However, since $\hat{V}^{t+1}$ and $\tilde{Q}^{t+1}$ depend on all the past θ^i vectors and hence, on the data collected in the previous iterations, the algorithm can effectively leverage all the data in $\mathcal{B}$. Furthermore, using the difference between the consecutive $\tilde{Q}$ functions can be viewed as a form of variance reduction. This enables us to prove an $O(1/\sqrt{N})$ concentration result for $\tilde{Q}^T$. Moreover, since the `DataCollection` procedure constructs a coreset which ensures good coverage across the feature space, the resulting sample complexity is independent of the size of the state-action space. Formally, in section C.2, we prove the following sub-optimality bound for $\square = r + \lambda_k \, c$ at iteration k of algorithm 1.

Lemma 10. *For a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, and any $k \in [K]$, when using algorithm 2 at iteration k of algorithm 1 with $\square = r + \lambda_k c$, $M = \tilde{O}\left(\frac{dH^2}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output policy π_T satisfies the following condition with probability* $1-\delta$,

$$\max_\pi V^\pi_{r+\lambda_k c}(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) \le O((1+\lambda_k)\varepsilon)$$

Hence, with a buffer $\mathcal{B}$ of size $T\,M|\mathcal{C}| = \tilde{O}\left(\frac{d^2 H^4}{\varepsilon^2}\right)$, algorithm 2 guarantees an optimality gap of $f_{\text{mdp}}(\mathcal{B}) = O(\varepsilon)$, thereby satisfying theorem 4. We note that the entropy-regularized variant of the above linear `MDVI` algorithm [Kitamura et al., 2023] also attains a similar guarantee but for a non-stationary policy output by the corresponding algorithm. Furthermore, in contrast to theorem 10, the guarantee in Kitamura et al. [2023] only holds for a more restricted range of $\varepsilon \in (0, 1/H]$. In the next section, we instantiate the `PolicyEvaluation` oracle.

4.3 Instantiating the `PolicyEvaluation` oracle: Least-Squares Policy Evaluation

To understand the need for an explicit `PolicyEvaluation` oracle, note that in each iteration k, we can prove that algorithm 2 ensures a concentration guarantee for the value function corresponding to $r + \lambda_k c$. However, this does not directly imply a concentration guarantee on the individual value functions corresponding to the reward and constraint rewards. This is in contrast to model-based approaches [Vaswani et al., 2022] for tabular CMDPs that guarantee concentration for the empirical transition probabilities, and use that to ensure concentration for both the reward and constraint reward value functions. However, since such model-based approaches cannot be used for linear MDPs, we require an additional algorithm that can compute the empirical value functions satisfying theorem 5. To that end, we present algorithm 3 that can be used as an instantiation of the `PolicyEvaluation` oracle in algorithm 1. The algorithm is also based on least-squares and uses the same coreset constructed in section 4.1. Furthermore, we note that algorithm 3 can be viewed as a special case of algorithm 2 for a fixed policy.

Algorithm 3 Least-Squares Policy Evaluation (`LS-PE`)

Input: T (number of iterations), M (number of next-state samples obtained per state-action pair in each iteration), $\diamond$ (either r or c), $\mathcal{B} = \mathcal{B}_0 \cup \cdots \cup \mathcal{B}_{T-1}$ (Buffer), π (policy to be evaluated),
$\tilde{\rho}$ (design), $\mathcal{C}$ (coreset), ϕ (feature map).

Output: $\bar{\mathcal{V}}^T_\diamond(\rho) = \frac{1}{T}\sum_{i=1}^{T} \hat{\mathcal{V}}^i_\diamond(\rho)$.

Define $\hat{\mathcal{V}}^0_\diamond = \mathbf{0}$.

```
procedure LS-PE($T$, $M$, $\diamond$, $\mathcal{B}$, $\pi$, $\tilde{\rho}, \mathcal{C}, \phi$)
    for $t = 0, 1, 2 \ldots, T-1$ do
        $\forall (s,a) \in \mathcal{C}$ : Access $(s, a, s'_m)_{m=1}^{M}$ from the buffer $\mathcal{B}_t$.
        Define regression target $\hat{\mathcal{Q}}^{t+1}_\diamond(s,a) := \diamond(s,a) + \gamma \frac{1}{M}\sum_{m=1}^{M} \hat{\mathcal{V}}^t_\diamond(s'_m)$.
        $\omega^{t+1}_\diamond = \arg\min_{\omega \in \mathbb{R}^d} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)(\langle \phi(x,b), \omega\rangle - \hat{\mathcal{Q}}^{t+1}_\diamond(x,b))^2$.
        Define $\hat{\mathcal{V}}^{t+1}_\diamond(s) := (\pi \langle \phi, \omega^{t+1}_\diamond \rangle)(s)$.
    end for
end procedure
```

Remark 11. *There exists prior CMDP literature that uses PE as a subroutine [Ding et al., 2021, Efroni et al., 2020], but our PE subroutine differs in several aspects. First, Ding et al. [2021], Efroni et al. [2020] use value functions estimated via model-based policy evaluation to update both the primal and dual variables; we utilize the PE subroutine solely for updating* *the dual variable. This decoupling enables a more reductionist algorithm and modular proof. Moreover, the resulting algorithmsin Ding et al. [2021], Efroni et al. [2020] do not achieve the near-optimal regret even for tabular CMDPs (in terms of their dependence on both S and H), and consequently sample-complexity in the generative model setting. In contrast, our method leverages averaging within PE and returns an averaged value function, which plays a crucial role in attaining the near-optimal (in d and ε for the linear setting and in all parameters for the tabular setting) sample complexity. An additional technical challenge lies in integrating the PE-induced error into the primal-dual analysis (i.e., Lemma 15) without losing statistical complexity.*

Note that `LS-PE` uses a fixed dataset (the buffer $\mathcal{B}$) to evaluate a fixed policy, and is similar to the policy evaluation algorithms in offline reinforcement learning [Duan et al., 2020]. The theoretical guarantees for such offline algorithms depend on the quality of the dataset, measured in terms of metrics such as coverage or concentrability. However, in our case, we curate the dataset and choose the buffer $\mathcal{B}$ such that it has good coverage properties that allow for fine-grained control on the algorithm's sub-optimality. In particular, we prove the following result in section C.3.

Lemma 12. *For a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, algorithm 3 with $M = \tilde{O}\left(\frac{dH^2}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output $\bar{\mathcal{V}}_\diamond^T$ satisfies the following condition with probability $1-\delta$,*

$$|\bar{\mathcal{V}}_\diamond^T(\rho) - V_\diamond^\pi(\rho)| \leq O\left(\varepsilon\right).$$

Hence, with a buffer $\mathcal{B}$ of size $T\, M|\mathcal{C}| = \tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$, algorithm 3 guarantees an optimality gap of $f_{\text{eva}}(\mathcal{B}) = O(\varepsilon)$, thereby satisfying theorem 5.

4.4 Putting everything together

We have seen that algorithms 2 and 3 use the buffer $\mathcal{B}$ constructed by the `DataCollection` procedure to provide control over the terms $f_{\text{mdp}}(\mathcal{B})$ and $f_{\text{eva}}(\mathcal{B})$ appearing in theorem 7. Combining these results, we prove the following corollary in section C.4.

Corollary 13. *Using `LS-MDVI` (algorithm 2) and `LS-PE` (algorithm 3) as instantiations of the `MDP-Solver` and `PolicyEvaluation` in algorithm 1 and using the `DataCollection` oracle described in section 4.1 has the following guarantee: for a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, algorithm 1 with $\tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$ samples, $U = O\left(\frac{1}{\zeta(1-\gamma)}\right)$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = O\left(\frac{1}{\varepsilon^2\,(1-\gamma)^2}\right)$, and $b' = b - O(\varepsilon)$, returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b - O(\varepsilon).$$

With the same algorithm parameters, but with $b' = b + O(\varepsilon)$ and $\tilde{O}\left(\frac{d^2H^6}{\zeta^2\varepsilon^2}\right)$ samples, algorithm 1 returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b.$$

Hence, the total sample complexity required to achieve the relaxed feasibility objective in eq. (2.2) and the strict feasibility objective in eq. (2.3) is $\tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$ and $\tilde{O}\left(\frac{d^2H^6}{\varepsilon^2\zeta^2}\right)$ respectively. Since the lower bound for unconstrained linear MDPs is $\Omega\left(\frac{d^2H^3}{\varepsilon^2}\right)$ [Weisz et al., 2022], our sample complexity achieves the optimal dependence on d and ε in the relaxed setting.

Flexibility of algorithm 1. Note that instead of LS-MDVI, we can use other unconstrained linear MDP solvers. For example, the G-Sampling and Stop (GSS) algorithm from Taupin et al. [2023] uses a different DataCollection procedure and algorithm to return an ε-optimal policy. It requires $\tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$ samples to do so, thus matching the sample complexity of LS-MDVI. We describe this algorithm in detail and formally instantiate algorithm 1 in section C.5.

4.5 Lower Bound

As shown by Vaswani et al. [2022], under the relaxed feasibility setting the statistical complexity of solving CMDPs matches that of unconstrained MDPs. In contrast, solving CMDPs under the strict setting is more challenging. In the tabular case, Vaswani et al. [2022] established a lower bound using a linear-programming argument. In this section, we establish a lower bound for solving linear CMDPs under strict feasibility. The proof technique of the following result is simpler and fundamentally different from that of Vaswani et al. [2022]. The proof is presented in Section D.

Theorem 14. *Let $\delta \in (0, 0.08]$, $\gamma \in [7/12, 1)$, $H = 1/(1-\gamma)$, $\varepsilon \in (0, 0.002)$, $\zeta \in (0, 49/2280)$, $b \in [H/2, H]$, and $d \geq 6$. There exists a class of linear constrained MDPs such that any (ε, δ)-sound algorithm requires $\Omega\left(d^2H^5/\varepsilon^2\zeta^2\right)$ samples from the generative model in the worst case.*

Thus, the dependence on d, ε, ζ in our bounds for the strict feasibility setting is also tight, with a suboptimality arising only in the multiplicative dependence on H.

4.6 Discussion

Why variance-weighted least squares fails in our setting. For unconstrained linear MDPs, Kitamura et al. [2023] provide an alternative entropy-regularized algorithm that

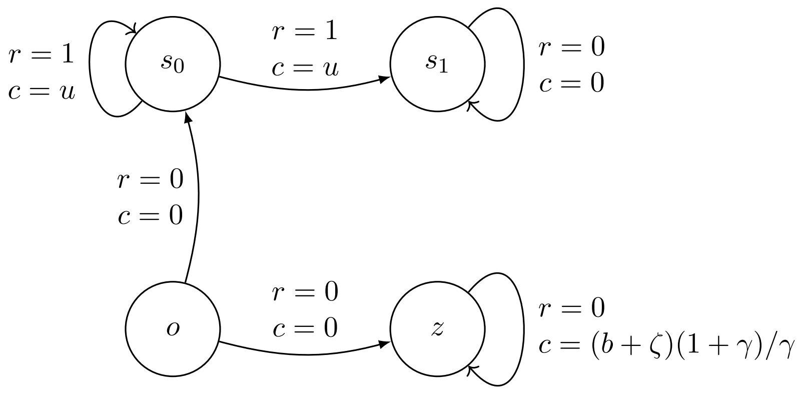

Figure 4.1: The lower bound instance consists of CMDPs with four states. o is the fixed starting state. At state o, taking an action will either transition to state s_0 or to the "safe" state z. At state s_0, taking an action will either transition to state s_1 or stay in s_0. States z and s_1 are absorbing.

constructs coresets that depend on the estimated empirical variance in the value function. The resulting algorithm uses variance-weighted least squares and is able to attain the near-optimal $O\left(\frac{d^2 H^3}{\varepsilon^2}\right)$ sample complexity for unconstrained linear MDPs. To the best of our knowledge, this is the only algorithm that can achieve such an optimal bound. Unfortunately, using such an idea for linear CMDPs fails. This is because in the linear CMDP setting, since the MDP reward function $r + \lambda_k c$ (and hence the MDP value function) change in every iteration k of algorithm 1, using variance-aware coresets implies that we need to construct a distinct coreset in every such iteration. This prevents the resulting algorithm from reusing data similar to algorithm 1, and actually increases the corresponding sample complexity. Resolving this issue and attaining the optimal dependence on H is an important direction for future work.

Comparing with online methods. In order to further contextualize our results, we use the state-of-the-art regret guarantees for the finite-horizon online setting [Ghosh et al., 2022] and use the reduction in Bai et al. [2021] to our problem setting. The reduction implies that the algorithm in [Ghosh et al., 2022] (designed and analyzed for the more difficult online regret minimization) results in an $\tilde{O}\left(\frac{d^3 H^4}{\varepsilon^2}\right)$ and $\tilde{O}\left(\frac{d^3 H^6}{\varepsilon^2 \zeta^2}\right)$ sample complexity for the relaxed and strict settings respectively. Hence, our results have a better dimension dependence. Interestingly, the analysis in [Ghosh et al., 2022] has a worse dependence on d because it uses a uniform concentration argument to get a handle on the concentration for the individual value functions corresponding to the (constraint) rewards. Recall that in section 4.3, we encountered a similar issue and resolved it by using policy evaluation. We believe that our technique might be useful even for online regret minimization.

Comparing with offline methods. Since we store all the samples at the beginning (see algorithm 1), an alternative approach is to treat these stored samples as an offline dataset, and use a standard offline RL algorithm [Hong and Tewari, 2024, Neu and Okolo, 2024] for the CMDP setting. It is worth noting that the sample complexity for such a reduction from the offline constrained RL setting can be substantially worse or even vacuous. In particular, the result on linear CMDP (e.g., Theorem 9) in Hong and Tewari [2024] depends on the concentrability coefficient C^* (see Assumption 2 therein), which can become unbounded if we do not collect data that covers the optimal action in every state. We note that constructing the $O(d^2)$ size coreset as in our paper does not directly ensure such coverage for each optimal action. When using the result from Hong and Tewari [2024], this can lead to unbounded values of C^* and a vacuous bound. For the sake of argument, even if we assume that C^* is bounded by an absolute constant (independent of d and H), the bounds for the offline CMDP setting in Hong and Tewari [2024] are worse than ours. Specifically, Hong and Tewari [2024, Theorem 9] is established under the relaxed feasibility setting; however, their sample complexity bound exhibits a quadratic dependence on the Slater constant, while our bounds in this setting do not have such a dependence. It is known that the Slater constant is only relevant in the strict feasibility regime [Vaswani et al., 2022]. Additionally, the dependence on the feature dimension d is d^3 in Hong and Tewari [2024], whereas our approach achieves the near-optimal dependence of d^2.

Chapter 5

Conclusion

Given access to a generative model, we proposed a generic primal-dual framework for reducing the (linear) CMDP problem to the (linear) MDP problem. Using (linear) `MDVI` as the `MDP-Solver` enabled us to obtain sample complexity bounds for both tabular and linear CMDPs with either $O(\varepsilon)$ or zero constraint violation. We obtained the first near-optimal (in d, ε, and ζ) guarantees for linear CMDPs, whereas for tabular CMDPs, we matched the existing near-optimal guarantees. We also provide a lower bound for solving linear CMDPs under strict feasibility. For linear CMDPs, improving the dependence of the sample complexity on the effective horizon H is an important direction for future work. Besides, it would also be useful to check if the bounds provided in this paper hold for other linear MDP assumptions, such as linear mixture MDPs [Zhou et al., 2021]. The sample complexity analyzed in this paper assumes access to a generative model. Although there has been prior work on linear CMDPs with exploration [Ghosh et al., 2022], the result exhibits suboptimal dependence on d and H. It would be interesting to explore whether the techniques developed in this work can be adapted to address this more challenging setting.

Bibliography

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

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

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *arXiv preprint arXiv:2109.06332*, 2021.

Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*, 2020.

Chiara Buratti, Andrea Conti, Davide Dardari, and Roberto Verdone. An overview on wireless sensor networks technology and evolution. *Sensors*, 9(9):6869–6896, 2009.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Ather Gattami, Qinbo Bai, and Vaneet Aggarwal. Reinforcement learning for constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2656–2664. PMLR, 2021.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pages 2160–2169. PMLR, 2019.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2024.

Aria HasanzadeZonuzy, Dileep M. Kalathil, and Srinivas Shakkottai. Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes. 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 2519–2525. ijcai.org, 2021.

Kihyuk Hong and Ambuj Tewari. A primal-dual algorithm for offline constrained reinforcement learning with linear mdps. *arXiv preprint arXiv:2402.04493*, 2024.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR, 2022.

Arushi Jain, Sharan Vaswani, Reza Babanezhad, Csaba Szepesvari, and Doina Precup. Towards painless policy optimization for constrained mdps. *arXiv preprint arXiv:2204.05176*, 2022.

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.

Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. *arXiv preprint arXiv:2405.12952*, 2024.

David Julian, Mung Chiang, Daniel O'Neill, and Stephen Boyd. Qos and fairness constrained convex optimization of resource allocation for wireless cellular and ad hoc networks. In *Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 477–486. IEEE, 2002.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning.* University of London, University College London (United Kingdom), 2003.

Krishna Chaitanya Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 8030–8037. AAAI Press, 2021.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002, 1999.

Toshinori Kitamura, Tadashi Kozuno, Yunhao Tang, Nino Vieillard, Michal Valko, Wenhao Yang, Jincheng Mei, Pierre Ménard, Mohammad Gheshlaghi Azar, Rémi Munos, et al. Regularization and variance-weighted regression achieves minimax optimality in linear mdps: theory and practice. In *International Conference on Machine Learning*, pages 17135–17175. PMLR, 2023.

Tadashi Kozuno, Eiji Uchibe, and Kenji Doya. Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2995–3003. PMLR, 2019.

Tadashi Kozuno, Wenhao Yang, Nino Vieillard, Toshinori Kitamura, Yunhao Tang, Jincheng Mei, Pierre Ménard, Mohammad Gheshlaghi Azar, Michal Valko, Rémi Munos, et al. Kl-entropy-regularized rl with a generative model is minimax optimal. *arXiv preprint arXiv:2205.14211*, 2022.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1): 1059–1106, 2023.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International conference on machine learning*, pages 5662–5670. PMLR, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36:3560–3577, 2023.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Policy optimization for constrained mdps with provable fast global convergence, 2022.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *International Conference on Machine Learning*, pages 15666–15698. PMLR, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Washim U Mondal and Vaneet Aggarwal. Sample-efficient constrained reinforcement learning with general parameterization. *Advances in Neural Information Processing Systems*, 37: 68380–68405, 2024.

Gergely Neu and Nneka Okolo. Offline rl via feature-occupancy gradient ascent. *arXiv preprint arXiv:2405.13755*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. Modeling medical treatment using markov decision processes. In *Operations research and health care*, pages 593–612. Springer, 2005.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5192–5202, 2018.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. *Naval Research Logistics (NRL)*, 70(5):423–442, 2023.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

Jerome Taupin, Yassir Jedra, and Alexandre Proutiere. Best policy identification in linear mdps. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2023.

Tian Tian, Lin Yang, and Csaba Szepesvári. Confident natural policy gradient for local planning in q_π -realizable constrained mdps. *Advances in Neural Information Processing Systems*, 37:76139–76176, 2024.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122, 2022.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in rl. *arXiv preprint arXiv:2003.14089*, 2020.

Mengdi Wang. Randomized linear programming solves the discounted markov decision problem in nearly-linear (sometimes sublinear) running time. *arXiv preprint arXiv:1704.01869*, 2017.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

Gellért Weisz, András György, Tadashi Kozuno, and Csaba Szepesvári. Confident approximate policy iteration for efficient local planning in q^π-realizable mdps. *Advances in Neural Information Processing Systems*, 35:25547–25559, 2022.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pages 6995–7004. PMLR, 2019.

Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. In *International Conference on Machine Learning*, pages 12167–12176. PMLR, 2021.

Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 36(4):1307–1319, 2020.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Appendix A

An Instantiation of `ComputeOptimalDesign`

In this section, we present an instantiation of the `ComputeOptimalDesign` oracle using the `Frank-Wolfe` algorithm [Todd, 2016].

Algorithm 4 `InitializeDesign`

```
Choose an arbitrary nonzero $c_0 \in \mathbb{R}^d$.        ▷ an auxiliary direction vector
Output: $\tilde{\rho}$.
procedure InitializeDesign
    for $j = 0, 1, 2 \dots, d-1$ do
        $(\bar{s}_j, \bar{a}_j) = \arg\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} c_j^\top \phi(s, a)$.
        $(s_j, a_j) = \arg\min_{(s,a) \in \mathcal{S} \times \mathcal{A}} c_j^\top \phi(s, a)$.
        $x_j = \phi(\bar{s}_j, \bar{a}_j) - \phi(s_j, a_j)$.
        Choose an arbitrary nonzero $c_{j+1}$ orthogonal to $x_0, \dots, x_j$.
    end for
Let $\mathcal{Z} := \{(\bar{s}_j, \bar{a}_j), (s_j, a_j) \mid j = 0, \dots, d-1\}$.
Choose $\tilde{\rho}$ to put equal weight on each of the distinct points of $\mathcal{Z}$.
end procedure
```

Below we present the classical `Frank-Wolfe` algorithm for experimental design.

Algorithm 5 `Frank-Wolfe`

```
Input: $\varepsilon^{FW}$.                                ▷ Tolerance for algorithm
Output: $\tilde{\rho}, \mathcal{C}, G$.                    ▷ Coreset, optimal design and covariance matrix
procedure Frank-Wolfe($\varepsilon^{FW}$)
    $\tilde{\rho}$ = InitializeDesign by Algorithm 4.
    Define $\mathcal{U} : \tilde{\rho} \mapsto \mathrm{diag}(\tilde{\rho}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$, where $\mathrm{diag}(\tilde{\rho})$ is a diagonal matrix with elements of $\tilde{\rho}$.
    For $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ be a matrix such that its $(s|\mathcal{A}| + a)$th row is $\phi(s, a)$.
    Define $\mathcal{I} : \tilde{\rho} \mapsto (\Phi^\top \mathcal{U}(\tilde{\rho})\, \Phi)^{-1}$.        ▷ defines the inverse of the covariance matrix
    Let $\nu : (s, a, \tilde{\rho}) \mapsto \phi(s, a)^\top \mathcal{I}(\tilde{\rho}) \phi(s, a)$.        ▷ measures the variance proxy for $(s, a)$
    Let $\delta : \tilde{\rho} \mapsto \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} (\nu(s, a, \tilde{\rho}) - d)/d$
                ▷ computes the relative difference between the worst-case variance and $d$
    while $\delta(\tilde{\rho}) > \varepsilon^{FW}$ do
        Let $(x, b) := \arg\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \nu(s, a, \tilde{\rho})$.
        Let $\eta^* := (\nu(x, b, \tilde{\rho}) - d)/((d - 1)\nu(x, b, \tilde{\rho}))$.
        $\tilde{\rho}(x, b) \leftarrow \tilde{\rho}(x, b) + \eta*$.
        $\tilde{\rho} \leftarrow \tilde{\rho}/(1 + \eta^*)$
    end while
Let $\mathcal{C} := \left\{ (s, a) \mid \nu(s, a, \tilde{\rho}) \geq d \left( 1 + \frac{\delta(\tilde{\rho})d}{2} - \sqrt{\delta(\tilde{\rho})(d - 1) + \frac{\delta(\tilde{\rho})^2 d^2}{4}} \right) \right\}$.
    ▷ form the coreset containing state-action pairs with sufficiently high variance value
Let $G := \sum_{(x,b) \in \mathcal{C}} \tilde{\rho}(x, b) \phi(x, b) \phi(x, b)^\top$. ▷ calculate the corresponding covariance matrix
end procedure
```

The subroutine `InitializeDesign` returns an initial design to be used in `Frank-Wolfe`. `InitializeDesign` is a deterministic procedure for constructing a core set of state-action pairs that provides good coverage of the feature space in linear MDPs. The algorithm sequentially identifies informative directions in the feature space by iteratively computing difference vectors between state-action pairs with maximal and minimal feature projections along a given search direction. The algorithm iteratively updates the search direction to be orthogonal to the span of the previously discovered directions. Specifically, the vector $c_j \in \mathbb{R}^d$ is an auxiliary direction vector used to sequentially identify maximally informative state-action pairs. The next vector c_{j+1} is then chosen to be orthogonal to all previous $x_0, \ldots, x_j$ ensuring that the design explores linearly independent directions in feature space. The resulting set of state-action pairs is then used as the support for a design distribution in regression.

Appendix B

Proof of Theorem 7

Theorem 7. *Suppose theorems 4 and 5 hold and let* $f(\mathcal{B}) := \max\{f_{\text{mdp}}(\mathcal{B}), f_{\text{eva}}(\mathcal{B})\}$*. For* $\delta \in (0,1)$*, algorithm 1 with* $U = \frac{2}{\zeta(1-\gamma)}$*,* $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$*,* $K = \frac{U^2}{[f(\mathcal{B})]^2(1-\gamma)^2}$ *and* $b' = b - 2f(\mathcal{B})$*, returns a mixture policy* $\bar{\pi}$ *satisfying the following condition with probability* $1-\delta$*,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - 4f(\mathcal{B}) \quad , \quad V_c^{\bar{\pi}}(\rho) \geq b - 6f(\mathcal{B}). \quad (\textbf{\textit{Relaxed Feasibility Setting}})$$

With the same algorithm parameters, but with $b' = b + 4f(\mathcal{B})$ *for* $f(\mathcal{B}) \leq \frac{\zeta}{6}$*, algorithm 1 returns a mixture policy* $\bar{\pi}$ *satisfying the following condition with probability* $1-\delta$*,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - \frac{16f(\mathcal{B})}{\zeta(1-\gamma)} \quad , \quad V_c^{\bar{\pi}}(\rho) \geq b. \quad (\textbf{\textit{Strict Feasibility Setting}})$$

Proof. We denote $\bar{\mathcal{V}}_\diamond^{\bar{\pi}} = \frac{1}{K}\sum_{k=0}^{K-1} \hat{V}_\diamond^k$ where $\diamond = r$ or c. We first prove the relaxed feasibility statement. By Lemma 15, we have $\bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \geq b - 5f(\mathcal{B})$. Hence,

$$\begin{aligned}
V_c^{\bar{\pi}}(\rho) &= V_c^{\bar{\pi}}(\rho) - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) + \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \\
&\geq b - 5f(\mathcal{B}) - |V_c^{\bar{\pi}}(\rho) - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)| \\
&\geq b - 5f(\mathcal{B}) - f(\mathcal{B}) && \text{(By Assumption 5 for each policy } \{\pi_k\}_{k=0}^{K-1}) \\
&= b - 6f(\mathcal{B}).
\end{aligned}$$

Next, we prove $V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}}(\rho) \leq 4f(\mathcal{B})$. We have

$$\begin{aligned}
V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}}(\rho) &= [V_r^{\pi^*}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho)] + [\bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) - V_r^{\bar{\pi}}(\rho)] \\
&\leq 3f(\mathcal{B}) + |\bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) - V_r^{\bar{\pi}}(\rho)| && \text{(By Lemma 15)} \\
&\leq 3f(\mathcal{B}) + f(\mathcal{B}) && \text{(By Assumption 5 for each policy } \{\pi_k\}_{k=0}^{K-1}) \\
&= 4f(\mathcal{B}).
\end{aligned}$$

Now we prove the strict feasibility statement. By Lemma 15, we have $\bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \geq b + f(\mathcal{B})$, and thus,

$$V_c^{\bar{\pi}}(\rho) = V_c^{\bar{\pi}}(\rho) - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) + \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)$$

$$\begin{aligned}
&\geq b + f(\mathcal{B}) - |V_c^{\bar{\pi}}(\rho) - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)| \\
&\geq b + f(\mathcal{B}) - f(\mathcal{B}) && \text{(By Assumption 5 for each policy } \{\pi_k\}_{k=0}^{K-1}\text{)} \\
&\geq b,
\end{aligned}$$

which satisfies the constraint. Next, we prove $V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}}(\rho) \leq 28 f(\mathcal{B})$. We define $\pi^{*+} \in \text{argmax}_\pi V_r^\pi(\rho)$ s.t. $V_c^\pi(\rho) \geq b + 6f(\mathcal{B})$. Note that such a policy exists by the definition of ζ and the assumption that $f(\mathcal{B}) \leq \frac{\zeta}{6}$. By Lemma 70 and Lemma 69, we know that

$$|V_r^{\pi^*}(\rho) - V_r^{\pi^{*+}}(\rho)| \leq 12 f(\mathcal{B}) \lambda^* \leq \frac{12 f(\mathcal{B})}{\zeta(1-\gamma)}.$$

Applying Lemma 15 and Assumption 5 as before, we have

$$\begin{aligned}
V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}}(\rho) &= [V_r^{\pi^*}(\rho) - V_r^{\pi^{*+}}(\rho)] + [V_r^{\pi^{*+}}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho)] + [\bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) - V_r^{\bar{\pi}}(\rho)] \\
&\leq \frac{12 f(\mathcal{B})}{\zeta(1-\gamma)} + 3f(\mathcal{B}) + f(\mathcal{B}) \\
&\leq \frac{16 f(\mathcal{B})}{\zeta(1-\gamma)}. && (\zeta(1-\gamma) \leq \tfrac{1-\gamma}{1-\gamma} = 1)
\end{aligned}$$

This completes the proof. □

B.1 Proof of Lemma 15 (Primal-Dual Guarantees for Algorithm 1)

Lemma 15 (Primal-Dual Guarantees for Algorithm 1)**.** *Suppose theorems 4 and 5 hold and let* $f(\mathcal{B}) := \max\{f_{\text{mdp}}(\mathcal{B}), f_{\text{eva}}(\mathcal{B})\}$. *For* $\delta \in (0,1)$, *when algorithm 1 is run with* $U = \frac{2}{\zeta(1-\gamma)}$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = \frac{U^2}{[f(\mathcal{B})]^2 (1-\gamma)^2}$ *and* $b' = b - 2f(\mathcal{B})$, *the following condition holds with probability* $1-\delta$,

$$\frac{1}{K}\sum_{k=0}^{K-1} \hat{V}_r^k(\rho) \geq V_r^{\pi^*}(\rho) - 3f(\mathcal{B}) \quad , \quad \frac{1}{K}\sum_{k=0}^{K-1} \hat{V}_c^k(\rho) \geq b - 5f(\mathcal{B}).$$

With the same algorithm parameters, but with $b' = b + 4f(\mathcal{B})$, *the following condition holds with probability* $1-\delta$,

$$\frac{1}{K}\sum_{k=0}^{K-1} \hat{V}_r^k(\rho) \geq V_r^{\pi^{*+}}(\rho) - 3f(\mathcal{B}) \quad , \quad \frac{1}{K}\sum_{k=0}^{K-1} \hat{V}_c^k(\rho) \geq b + f(\mathcal{B}).$$

Proof. We begin by proving the first part of the lemma. Since both r and c are bounded by 1, we note that $r(s,a) + \lambda_k c(s,a) \leq 1 + \lambda_k$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Define $\pi_k^* := \arg\max_\pi V_{r+\lambda_k c}^\pi$ as an optimal policy in the MDP with rewards $r + \lambda_k c$. For each iteration k in algorithm 1, by Assumption 4 with $R = 1 + \lambda_k$, we have

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - V_{r+\lambda_k c}^{\pi_k} \leq f_{\text{mdp}}(\mathcal{B})(1+\lambda_k).$$

By Assumption 5 for policy π_k, we have

$$\begin{aligned}V_{r+\lambda_k c}^{\pi_k}(\rho) - \hat{V}_r^k(\rho) - \lambda_k \hat{V}_c^k(\rho) &= V_r^{\pi_k}(\rho) + \lambda_k V_c^{\pi_k}(\rho) - \hat{V}_r^k(\rho) - \lambda_k \hat{V}_c^k(\rho) \\ &= V_r^{\pi_k}(\rho) - \hat{V}_r^k(\rho) + \lambda_k(V_c^{\pi_k}(\rho) - \hat{V}_c^k(\rho)) \\ &\leq f_{\text{eva}}(\mathcal{B})(1+\lambda_k).\end{aligned}$$

Combining the above inequalities and letting $f(\mathcal{B}) = \max\{f_{\text{mdp}}(\mathcal{B}), f_{\text{eva}}(\mathcal{B})\}$, we obtain

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - (\hat{V}_r^k(\rho) + \lambda_k \hat{V}_c^k(\rho)) \leq (f_{\text{mdp}}(\mathcal{B}) + f_{\text{eva}}(\mathcal{B}))(1+\lambda_k) \leq 2f(\mathcal{B})\,(1+\lambda_k).$$

By the definition of π_k^*,

$$V_r^{\pi^*}(\rho) + \lambda_k V_c^{\pi^*}(\rho) \leq V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho).$$

Therefore, by combining the above inequalities,

$$\begin{aligned}&V_r^{\pi^*}(\rho) + \lambda_k V_c^{\pi^*}(\rho) \leq \hat{V}_r^k(\rho) + \lambda_k \hat{V}_c^k(\rho) + 2f(\mathcal{B})\,(1+\lambda_k) \\ \Longrightarrow\ &V_r^{\pi^*}(\rho) - \hat{V}_r^k(\rho) \leq \lambda_k(\hat{V}_c^k(\rho) - V_c^{\pi^*}(\rho) + 2f(\mathcal{B})) + 2f(\mathcal{B}).\end{aligned} \tag{B.1}$$

Since $V_c^{\pi^*}(\rho) \geq b$ and $\lambda_k \geq 0$, we obtain

$$V_r^{\pi^*}(\rho) - \hat{V}_r^k(\rho) \leq \lambda_k(\hat{V}_c^k(\rho) - b + 2f(\mathcal{B})) + 2f(\mathcal{B}).$$

By taking the average, letting $b' = b - 2f(\mathcal{B})$, and adding both sides by the same term $\frac{\lambda}{K}\sum_{k=0}^{K-1}\left[b' - \hat{V}_c^k(\rho)\right]$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\left[V_r^{\pi^*}(\rho) - \hat{V}_r^k(\rho)\right] + \frac{\lambda}{K}\sum_{k=0}^{K-1}\left[b' - \hat{V}_c^k(\rho)\right] \leq \frac{1}{K}\sum_{k=0}^{K-1}(\lambda_k - \lambda)(\hat{V}_c^k(\rho) - b') + 2f(\mathcal{B}).$$

Now we define $R(\lambda, K) := \sum_{k=0}^{K-1}(\lambda_k - \lambda)(\hat{V}_c^k(\rho) - b')$ as the dual regret and denote $\bar{\mathcal{V}}_\diamond^{\bar{\pi}} = \frac{1}{K}\sum_{k=0}^{K-1}\hat{V}_\diamond^k$ (where $\diamond = r$ or c). Thus, for any $\lambda \in [0, U]$,

$$V_r^{\pi^*}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) + \lambda(b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)) \leq \frac{R(\lambda, K)}{K} + 2f(\mathcal{B}). \tag{B.2}$$

Below we show that for any $\lambda \in [0, U]$, the following bound holds for the dual regret:

$$R(\lambda, K) \leq \frac{U\sqrt{K}}{1-\gamma}.$$

Using the dual update in algorithm 1, we observe that,

$$\begin{aligned}|\lambda_{k+1} - \lambda|^2 &\leq \left|\lambda_k - \eta\left(\hat{V}_c^k(\rho) - b'\right) - \lambda\right|^2 \qquad \text{(by non-expansiveness of projection)} \\ &= |\lambda_k - \lambda|^2 - 2\eta\,(\lambda_k - \lambda)\left(\hat{V}_c^k(\rho) - b'\right) + \eta^2\left(\hat{V}_c^k(\rho) - b'\right)^2 \\ &\overset{(a)}{\leq} |\lambda_k - \lambda|^2 - 2\eta\,(\lambda_k - \lambda)\left(\hat{V}_c^k(\rho) - b'\right) + \frac{\eta^2}{(1-\gamma)^2},\end{aligned}$$

where (a) follows because b and the constraint value are in the $[0, 1/(1-\gamma)]$ interval. Rearranging and dividing by 2η, we get

$$(\lambda_k - \lambda)\left(\hat{V}_c^k(\rho) - b'\right) \leq \frac{|\lambda_t - \lambda|^2 - |\lambda_{k+1} - \lambda|^2}{2\eta} + \frac{\eta}{2(1-\gamma)^2}.$$

Summing from $k = 0$ to $K - 1$ and using the definition of the dual regret,

$$R(\lambda, K) \leq \frac{1}{2\eta} \sum_{k=0}^{K-1} \left[|\lambda_k - \lambda|^2 - |\lambda_{k+1} - \lambda|^2\right] + \frac{\eta K}{2(1-\gamma)^2}.$$

Telescoping, bounding $|\lambda_0 - \lambda|$ by U and dropping a negative term gives

$$R(\lambda, K) \leq \frac{U^2}{2\eta} + \frac{\eta K}{2(1-\gamma)^2},$$

Setting $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$,

$$R(\lambda, K) \leq \frac{U\sqrt{K}}{1-\gamma}. \tag{B.3}$$

Next, in order to bound the reward optimality gap, setting $\lambda = 0$ in eq. (B.2) and using the above bound on the dual regret, we obtain

$$V_r^{\pi^*}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) \leq \frac{U}{(1-\gamma)\sqrt{K}} + 2f(\mathcal{B}). \tag{B.4}$$

In order to bound the constraint violation, we consider two cases. The first case is when $b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \leq 0$. Consequently, $b - 2f(\mathcal{B}) - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \leq 0$ and hence, $\bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \geq b - 2f(\mathcal{B}) \geq b - 5f(\mathcal{B})$, which completes the proof.

The second case is when $b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) > 0$. In this case, using the notation $[x]_+ = \max\{x, 0\}$ and eq. (B.2) with $\lambda = U$, we have

$$V_r^{\pi^*}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) + U\left[b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)\right]_+ \leq \frac{R(U, K)}{K} + 2f(\mathcal{B}).$$

Since U has been set such that $U > \lambda^*$, we can use Lemma 68 and obtain that,

$$\left[b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)\right]_+ \leq \frac{R(U, K)}{K(U - \lambda^*)} + \frac{2f(\mathcal{B})}{U - \lambda^*}$$

Combining the above inequality with eq. (B.3) gives

$$b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \leq \left[b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)\right]_+ \leq \frac{U}{(U - \lambda^*)(1-\gamma)\sqrt{K}} + \frac{2f(\mathcal{B})}{U - \lambda^*}. \tag{B.5}$$

By Lemma 69, we know $\lambda^* \leq \frac{1}{\zeta(1-\gamma)}$. By letting $U = \frac{2}{\zeta(1-\gamma)}$, we have $U - \lambda^* \geq \frac{1}{\zeta(1-\gamma)} \geq 1$ as the Slater constant $\zeta \in (0, \frac{1}{1-\gamma}]$. Thus, $\frac{1}{U-\lambda^*} \leq 1$. Now, setting K to to be

$$K = \frac{U^2}{[f(\mathcal{B})]^2(1-\gamma)^2}$$

and substituting into eqs. (B.4) and (B.5), we obtain

$$\bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - 3f(\mathcal{B}), \quad \text{and} \quad \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \geq b' - 3f(\mathcal{B}). \tag{B.6}$$

This establishes the first claim by substituting $b' = b - 2f(\mathcal{B})$.

Next, we prove the second claim. We define $\pi^{*+} \in \operatorname{argmax}_\pi V_r^\pi(\rho)$ s.t. $V_c^\pi(\rho) \geq b + 6f(\mathcal{B})$. From eq. (B.7), recall that

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - (\hat{V}_r^k(\rho) + \lambda_k \hat{V}_c^k(\rho)) \leq 2f(\mathcal{B})\,(1 + \lambda_k)$$

As before, using the definition of π_k^*, we have

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) \geq V_r^{\pi^{*+}}(\rho) + \lambda_k V_c^{\pi^{*+}}(\rho),$$

Therefore, by combining the above inequalities,

$$\begin{aligned} & V_r^{\pi^{*+}}(\rho) + \lambda_k V_c^{\pi^{*+}}(\rho) \leq \hat{V}_r^k(\rho) + \lambda_k \hat{V}_c^k(\rho) + 2f(\mathcal{B})\,(1 + \lambda_k) \\ \implies & V_r^{\pi^{*+}}(\rho) - \hat{V}_r^k(\rho) \leq \lambda_k(\hat{V}_c^k(\rho) - V_c^{\pi^{*+}}(\rho) + 2f(\mathcal{B})) + 2f(\mathcal{B}). \end{aligned} \tag{B.7}$$

Since $V_c^{\pi^{*+}}(\rho) \geq b + 6f(\mathcal{B})$, we obtain,

$$V_r^{\pi^{*+}}(\rho) - \hat{V}_r^k(\rho) \leq \lambda_k[\hat{V}_c^k(\rho) - (b + 3f(\mathcal{B}))] + 2f(\mathcal{B}).$$

As before, by taking the average, letting $b' = b + 4f(\mathcal{B})$, and adding both sides by the same term $\frac{\lambda}{K}\sum_{k=0}^{K-1}\left[b' - \hat{V}_c^k(\rho)\right]$, we obtain that for $\lambda \in [0, U]$,

$$V_r^{\pi^{*+}}(\rho) - \bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) + \lambda(b' - \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho)) \leq \frac{R(\lambda, K)}{K} + 2f(\mathcal{B}).$$

The remainder of the proof proceeds in the same manner as before. Setting K to to be

$$K = \frac{U^2}{[f(\mathcal{B})]^2(1-\gamma)^2}$$

the algorithm ensures that

$$\bar{\mathcal{V}}_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^{*+}}(\rho) - 3f(\mathcal{B}), \quad \text{and} \quad \bar{\mathcal{V}}_c^{\bar{\pi}}(\rho) \geq b' - 3f(\mathcal{B}). \tag{B.8}$$

This establishes the second claim by substituting $b' = b + 4f(\mathcal{B})$. □

Appendix C

Proofs for Section 4

The proofs in Section C.1, Section C.2 and Section C.3 are adapted from Kitamura et al. [2023], Kozuno et al. [2022] with modifications to fit our setting. Specifically, the analysis in Kitamura et al. [2023] applies to the *non-stationary policies* returned by `MDVI` *with* entropy regularization. In contrast, our analysis applies to the *stationary policy* returned by MDVI *without* entropy regularization. Furthermore, we also require additional analysis of the value functions returned by the `LS-PE` algorithm.

Throughout, we treat π as an operator that returns an $|\mathcal{S}|$-dimensional vector s.t. for an arbitrary $|\mathcal{S}||\mathcal{A}|$-dimensional vector u such that $(\pi u)(s) := \sum_{a \in \mathcal{A}} \pi(a|s)\, u(s,a)$. Furthermore, we define $P_\pi := \pi P$ where $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and denotes the transition probability matrix induced by policy π.

C.1 Deriving `LS-MDVI` from Entropic Mirror Descent

We show that the `LS-MDVI` update can be derived as a limiting case of entropic mirror descent. At iteration t, given Q_t, if κ is the entropy regularization parameter and τ is the KL regularization parameter, then, the entropic mirror descent policy update Kitamura et al. [2023] is:

$$\pi_t(\cdot|s) = \arg\max_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \left(Q^t(s,a) - \tau \log \frac{p(a)}{\pi_{t-1}(a|s)} - \kappa \log p(a) \right), \quad \text{for all } s \in \mathcal{S},$$

The above policy update can be rewritten in a closed-form solution as follows [Kozuno et al., 2019, Equation 5]),

$$\pi_t(a|s) = \frac{[\pi_{t-1}(a|s)]^\alpha \exp\left(\beta Q^t(s,a)\right)}{\sum_{b \in \mathcal{A}} [\pi_{t-1}(b|s)]^\alpha \exp\left(\beta Q^t(s,b)\right)}, \text{ where } \alpha := \tau/(\tau+\kappa),\ \beta := 1/(\tau+\kappa)$$

$$\implies \pi_t(a|s) = \frac{\exp(\beta \sum_{i=0}^{t} \alpha^{t-i}\, Q^i(s,a))}{\sum_{b \in \mathcal{A}} \exp(\beta \sum_{i=0}^{t} \alpha^{t-i}\, Q^i(s,b))}.$$

Since `LS-MDVI` does not use entropy regularization $\kappa = 0$ implying $\alpha = 1$, the resulting update is:

$$\pi_t(a|s) = \frac{\exp(\beta \sum_{i=0}^{t} Q^i(s,a))}{\sum_{b \in \mathcal{A}} \exp(\beta \sum_{i=0}^{t} Q^i(s,b)} = \frac{1}{1 + \sum_{b \neq a} \exp(\beta(\bar{Q}^t(s,b) - \bar{Q}^t(s,a))} \qquad \text{(where } \bar{Q}^t := \textstyle\sum_{i=0}^{t} Q^i\text{)}$$

For `LS-MDVI`, we take the limit $\tau \to 0$, $\beta \to \infty$ and consider two cases.

Case 1: If $a = \arg\max_b \bar{Q}^t(s,b)$, then, $\beta(\bar{Q}^t(s,b) - \bar{Q}^t(s,a)) < 0$ for all $b \neq a$. Hence, as $\beta \to \infty$, $\sum_{b \neq a} \exp(\beta(\bar{Q}^t(s,b) - \bar{Q}^t(s,a)) \to 0$ and $\pi_t(a|s) \to 1$.

Case 2: If $a \neq \arg\max_b \bar{Q}^t(s,b)$, then, $\beta(\bar{Q}^t(s,b) - \bar{Q}^t(s,a)) > 0$ for the action b corresponding to the arg max action. Hence, as $\beta \to \infty$, $\sum_{b \neq a} \exp(\beta(\bar{Q}^t(s,b) - \bar{Q}^t(s,a)) \to \infty$ and $\pi_t(a|s) \to 0$.

Hence, as $\kappa = 0$ and $\tau \to 0$, π_t is a greedy policy and for all $s \in \mathcal{S}$, $\pi_t(a|s) = 1$ for $a = \arg\max_b \sum_{i=0}^{t} Q^i(s,b)$, which recovers the policy update for `LS-MDVI`.

For entropic mirror descent, the value update is given as [Kitamura et al., 2023], for all $s \in \mathcal{S}$,

$$\begin{aligned} V^t(s) &= \sum_a \pi_t(a|s) \left(Q^t(s,a) - \tau \log\left(\frac{\pi_t(a|s)}{\pi_{t-1}(a|s)}\right) - \kappa \ln(\pi_t(a|s)) \right) \\ &= (\pi_t Q^t)(s) - \tau \mathrm{KL}(\pi_t(\cdot|s) \| \pi_{t-1}(\cdot|s)) + \kappa\, \mathcal{H}(\pi_t(\cdot|s)). \end{aligned}$$

Plugging the entropic mirror descent policy update and simplifying similar to [Kozuno et al., 2022, App. B], we get,

$$\begin{aligned} V^t(s) &= \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp\left(\beta Q^t(s,a) + \alpha \log \pi_{t-1}(a|s)\right) \\ &= \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp\left(\beta \sum_{i=0}^{t} \alpha^{t-i} Q^i(s,a)\right) - \frac{\alpha}{\beta} \log \sum_{a \in \mathcal{A}} \exp\left(\beta \sum_{i=0}^{t-1} \alpha^{t-i} Q^i(s,a)\right). \end{aligned}$$

Since `LS-MDVI` does not use entropy regularization i.e. $\kappa = 0$ implying $\alpha = 1$, the update is:

$$V^t(s) = \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp\left(\beta \sum_{i=0}^{t} Q^i(s,a)\right) - \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp\left(\beta \sum_{i=0}^{t-1} Q^i(s,a)\right).$$

For `LS-MDVI`, we take the limit $\tau \to 0$, $\beta \to \infty$. Using LH́opital's rule for the two terms, we get that,

$$\begin{aligned} V^t(s) &= \sum_a \pi_t(a|s) \sum_{i=0}^{t} Q^i(s,a) - \sum_a \pi_{t-1}(a|s) \sum_{i=0}^{t-1} Q^i(s,a) \\ &= \left(\pi_t \sum_{i=0}^{t} Q^i\right)(s) - \left(\pi_{t-1} \sum_{i=0}^{t-1} Q^i\right)(s), \end{aligned}$$

which recovers the value update for `LS-MDVI`.

C.2 Proof of Lemma 10 (Optimality Guarantees for Algorithm 2 - Linear CMDP)

Note that for each λ_k where $k \in [K]$, we run Algorithm 2 with $\square = r + \lambda_k\, c$. We define

$$\pi_k^* := \arg\max_{\pi} V^{\pi}_{r+\lambda_k c} \tag{C.1}$$

$$\bar{V}_\square^T := \frac{1}{T}\sum_{i=1}^{T} \hat{V}_\square^i \overset{\text{(by telescoping)}}{=} \frac{1}{T}(\pi_T \tilde{Q}_\square^T) \overset{\text{(by definition)}}{=} \frac{1}{T}\left(\pi_T \left\langle \phi, \sum_{i=0}^{T} \theta_\square^i \right\rangle\right). \tag{C.2}$$

Throughout the proof, for any $|\mathcal{S}||\mathcal{A}|$-dimensional vector z, we let $W(z)$ denote the solution to a weighted linear regression problem over the core set,

$$W(z) := \arg\min_{\theta} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)\left(z(x,b) - \langle \phi(x,b), \theta\rangle\right)^2. \tag{C.3}$$

The above problem can be solved as

$$W(z) = G^{-1} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)\phi(x,b)z(x,b) \tag{C.4}$$

where $G := \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)\,\phi(x,b)\phi(x,b)^\top$.

Using this definition and the definition of $\theta_\square^i$ in algorithm 2, we have $\theta_\square^i = W(\hat{Q}_\square^i)$.

The linear MDP assumption ensures that there exists a vector $\boldsymbol{\theta}_\square^t$ such that $\langle \phi, \boldsymbol{\theta}_\square^t\rangle := \square + \gamma P \hat{V}_\square^{t-1}$. Therefore, using the definition of W, we have $\boldsymbol{\theta}_\square^i = W(\langle \phi, \boldsymbol{\theta}_\square^i\rangle)$.

We now present the proof of Lemma 10.

Lemma 16. *For a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, and any $k \in [K]$, when using algorithm 2 at iteration k of algorithm 1 with $\square = r + \lambda_k c$, $M = \tilde{O}\left(\frac{dH^2}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output policy π_T satisfies the following condition with probability $1-\delta$,*

$$\max_{\pi} V^{\pi}_{r+\lambda_k c}(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) \leq O((1+\lambda_k)\varepsilon)$$

Proof. Using the definition of π_k^* and that $\square = r + \lambda_k\, c$, we decompose the sub-optimality as:

$$V^{\pi_k^*}_{r+\lambda_k c}(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) = [V^{\pi_k^*}_{r+\lambda_k c}(\rho) - \bar{V}_\square^T(\rho)] + [\bar{V}_\square^T(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho)]$$

Bounding the first term by Lemma 17 and the second by Lemma 18,

$$\leq \tilde{O}\left(\frac{H^2(1+\lambda_k)}{T} + H^2(1+\lambda_k)\sqrt{\frac{d}{TM}}\right)$$

with probability at least $1-2\delta$. Setting $M = \tilde{O}\left(\frac{dH^2}{\varepsilon}\right)$, $T = O(\frac{H^2}{\varepsilon})$, and appropriately rescaling the confidence parameter δ completes the proof. □

We now prove theorems 17 and 18.

Lemma 17. *Let π_k^* and $\bar{V}_\square^T$ be defined as in eqs.* (C.1) *and* (C.2)*. For any $k \in [K]$, with $\square = r + \lambda_k c$ and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$V_{r+\lambda_k c}^{\pi_k^*}(\rho) - \bar{V}_\square^T(\rho) \leq \tilde{O}\left(\frac{H^2(1+\lambda_k)}{T} + H^2(1+\lambda_k)\sqrt{\frac{d}{TM}}\right)$$

with probability at least $1-\delta$.

Proof. We first recall that $V_\square^{\pi_k^*} = V_{r+\lambda_k c}^{\pi_k^*}$ and $\bar{V}_\square^T = \bar{V}_{r+\lambda_k c}^T$ by the definition of $\square$. By the value difference lemma, we have that,

$$V_\square^{\pi_k^*} - \bar{V}_\square^T = (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*} \bar{V}_\square^T - \bar{V}_\square^T) \tag{C.5}$$

Next, from Line 6 in Algorithm 2, by the telescoping sum, and by the greediness of π_T, we have

$$\bar{V}_\square^T = \frac{1}{T}\,(\pi_T \tilde{Q}_\square^T) \tag{C.6}$$
$$\geq \frac{1}{T}\,(\pi_k^* \tilde{Q}_\square^T). \tag{C.7}$$

Now, we have

$$\begin{aligned}
V_\square^{\pi_k^*} - \bar{V}_\square^T &= (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*} \bar{V}_\square^T - \bar{V}_\square^T) && \text{(By eq. (C.5))}\\
&\leq (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*} \bar{V}_\square^T - \frac{1}{T}\,(\pi_k^* \tilde{Q}_\square^T)) && \text{(By eq. (C.7))}\\
&= (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*} \frac{1}{T}\,(\pi_T \tilde{Q}_\square^T) - \frac{1}{T}\,(\pi_k^* \tilde{Q}_\square^T)) && \text{(By eq. (C.6))}\\
&= (I - \gamma P_{\pi_k^*})^{-1}\,\Bigg[(\pi_k^* \square) + \gamma P_{\pi_k^*} \frac{1}{T}\,(\pi_T \tilde{Q}_\square^T) - (\pi_k^* \square) - \gamma P_{\pi_k^*} \frac{1}{T}(\pi_{T-1} \tilde{Q}_\square^{T-1})\\
&\quad - \left(\pi_k^* \left\langle \phi, W\left(\frac{1}{T}\sum_{i=0}^{T}(\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle)\right)\right\rangle\right)\Bigg] && \text{(Using Lemma 25 for } \tfrac{1}{T}\tilde{Q}_\square^T)\\
&= (I - \gamma P_{\pi_k^*})^{-1}\,\Bigg[\gamma P_{\pi_k^*} \frac{1}{T}\,(\pi_T \tilde{Q}_\square^T) - \gamma P_{\pi_k^*} \frac{1}{T}\,(\pi_{T-1} \tilde{Q}_\square^{T-1})\\
&\quad - \left(\pi_k^* \left\langle \phi, W\left(\frac{1}{T}\sum_{i=0}^{T}(\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle)\right)\right\rangle\right)\Bigg].
\end{aligned}$$

By defining $\mathcal{H}_{\pi_k^*} := (I - \gamma P_{\pi_k^*})^{-1}$, taking the infinity norm and using the triangle inequality, we obtain

$$\left\| V_\square^{\pi_k^*} - \bar{V}_\square^T \right\|_\infty \leq \underbrace{\left\| \gamma \mathcal{H}_{\pi_k^*} P_{\pi_k^*} \left(\frac{1}{T}\,(\pi_T \tilde{Q}_\square^T) - \frac{1}{T}\,(\pi_{T-1} \tilde{Q}_\square^{T-1})\right) \right\|_\infty}_{\text{Term (i)}}$$

$$+ \underbrace{\left\| \mathcal{H}_{\pi_k^*} \left(\pi_k^* \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle) \right) \right\rangle \right) \right\|_\infty}_{\text{Term (ii)}} . \tag{C.8}$$

In order to bound Term (i), we use Holder's inequality i.e. for a matrix A and vector x, $\left\| Ax \right\|_\infty \leq \left\| A \right\|_{1,\infty} \left\| x \right\|_\infty$, and that $\|\mathcal{H}_{\pi_k^*} P_{\pi_k^*}\|_{1,\infty} \leq H$ to obtain,

$$\begin{aligned}
\left\| \gamma\, \mathcal{H}_{\pi_k^*} P_{\pi_k^*} \left(\frac{1}{T} \left(\pi_T \tilde{Q}_\square^T \right) - \frac{1}{T} \left(\pi_{T-1} \tilde{Q}_\square^{T-1} \right) \right) \right\|_\infty &\leq H \left\| \left(\frac{1}{T} \left(\pi_T \tilde{Q}_\square^T \right) - \frac{1}{T} \left(\pi_{T-1} \tilde{Q}_\square^{T-1} \right) \right) \right\|_\infty \\
&\leq \frac{4H^2(1+\lambda_k)}{T} \qquad \text{(Using theorem 21)}
\end{aligned}$$

with probability at least $1 - \delta$. For term (ii),

$$\begin{aligned}
&\left\| \mathcal{H}_{\pi_k^*} \left(\pi_k^* \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle) \right) \right\rangle \right) \right\|_\infty \\
&\leq \left\| \mathcal{H}_{\pi_k^*} \right\|_{1,\infty} \left\| \left(\pi_k^* \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle) \right) \right\rangle \right) \right\|_\infty && \text{(By Holder's inequality)} \\
&\leq H \left\| \left(\pi_k^* \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle) \right) \right\rangle \right) \right\|_\infty && \text{(Since } \|\mathcal{H}_{\pi_k^*}\|_{1,\infty} \leq H\text{)} \\
&\leq H \left\| \phi^\top W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{Q}_\square^i - \langle \phi, \boldsymbol{\theta}_\square^i \rangle) \right) \right\|_\infty && \text{(By definition of the } \pi \text{ operator)} \\
&\leq \tilde{O} \left(H^2 (1+\lambda_k) \sqrt{\frac{d}{TM}} \right) && \text{(By Lemma 22)}
\end{aligned}$$

Combining the above relations,

$$\left\| V_\square^{\pi_k^*} - \bar{V}_\square^T \right\|_\infty \leq \frac{4H^2(1+\lambda_k)}{T} + \tilde{O} \left(H^2 (1+\lambda_k) \sqrt{\frac{d}{TM}} \right)$$

Using that for any $|\mathcal{S}|$-dimensional vector V, $V(\rho) = \mathbb{E}_{s\sim\rho} V(s) \leq \|V\|_\infty$, we get that,

$$V_{r+\lambda_k c}^{\pi_k^*}(\rho) - \bar{V}_\square^T(\rho) \leq \frac{4H^2(1+\lambda_k)}{T} + \tilde{O} \left(H^2 (1+\lambda_k) \sqrt{\frac{d}{TM}} \right)$$

with probability at least $1 - \delta$. □

Lemma 18. *Let $\bar{V}_\square^T$ be defined as in eq. (C.2). For any $k \in [K]$, with $\square = r + \lambda_k\, c$ and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$\bar{V}_\square^T(\rho) - V_{r+\lambda_k c}^{\pi_T}(\rho) \leq \tilde{O} \left(\frac{H^2(1+\lambda_k)}{T} + H^2(1+\lambda_k) \sqrt{\frac{d}{TM}} \right)$$

with probability at least $1 - \delta$.

Proof. The proof is similar as for the above lemma. By the value difference lemma, we have that,

$$\bar{V}_{\Box}^{T} - V_{r+\lambda_k c}^{\pi_T} = (I - \gamma P_{\pi_T})^{-1}(\bar{V}_{\Box}^{T} - (\pi_T \Box) - \gamma P_{\pi_T}\, \bar{V}_{\Box}^{T}) \tag{C.9}$$

Now, we have

$$\begin{aligned}
\bar{V}_{\Box}^{T} - V_{r+\lambda_k c}^{\pi_T} &= (I - \gamma P_{\pi_T})^{-1} \left(\frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) - (\pi_T \Box) - \gamma P_{\pi_T}\, \bar{V}_{\Box}^{T} \right) \\
&= (I - \gamma P_{\pi_T})^{-1} \left(\frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) - (\pi_T \Box) - \gamma P_{\pi_T}\, \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \right) \\
&= (I - \gamma P_{\pi_T})^{-1} \Bigg[(\pi_T \Box) + \gamma P_{\pi_T}\, \frac{1}{T} (\pi_{T-1} \tilde{Q}_{\Box}^{T-1}) + \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i} \rangle) \right) \right\rangle \right) \\
&\quad - (\pi_T \Box) - \gamma P_{\pi_T}\, \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \Bigg] \qquad \text{(Using Lemma 25 for } \tfrac{1}{T}\tilde{Q}^{T}\text{)} \\
&= (I - \gamma P_{\pi_T})^{-1} \Bigg[\gamma P_{\pi_T}\, \frac{1}{T} (\pi_{T-1} \tilde{Q}_{\Box}^{T-1}) - \gamma P_{\pi_T}\, \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \\
&\quad + \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i} \rangle) \right) \right\rangle \right) \Bigg].
\end{aligned}$$

By defining $\mathcal{H}_{\pi_T} := (I - \gamma P_{\pi_T})^{-1}$, taking the infinity norm and using the triangle inequality, we obtain

$$\begin{aligned}
\left\| \bar{V}_{\Box}^{T} - V_{r+\lambda_k\, c}^{\pi_T} \right\|_{\infty} \leq{}& \underbrace{\left\| \gamma \mathcal{H}_{\pi_T} P_{\pi_T} \left(\frac{1}{T}\,(\pi_{T-1} \tilde{Q}_{\Box}^{T-1}) - \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \right) \right\|_{\infty}}_{\text{Term (i)}} \\
&+ \underbrace{\left\| \mathcal{H}_{\pi_T} \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i} \rangle) \right) \right\rangle \right) \right\|_{\infty}}_{\text{Term (ii)}}.
\end{aligned} \tag{C.10}$$

In order to bound Term (i), we use Holder's inequality and that $\|\mathcal{H}_{\pi_T}\, P_{\pi_T}\|_{1,\infty} \leq H$,

$$\begin{aligned}
\left\| \gamma\, \mathcal{H}_{\pi_T} P_{\pi_T} \left(\frac{1}{T}\,(\pi_{T-1} \tilde{Q}_{\Box}^{T-1}) - \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \right) \right\|_{\infty} &\leq H\, \left\| \frac{1}{T}\,(\pi_{T-1} \tilde{Q}_{\Box}^{T-1}) - \frac{1}{T}\,(\pi_T \tilde{Q}_{\Box}^{T}) \right\|_{\infty} \\
&\leq \frac{4H^2(1+\lambda_k)}{T} \qquad \text{(Using theorem 21)}
\end{aligned}$$

with probability at least $1 - \delta$. For term (ii),

$$\begin{aligned}
&\left\| \mathcal{H}_{\pi_T} \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i} \rangle) \right) \right\rangle \right) \right\|_{\infty} \\
&\leq \|\mathcal{H}_{\pi_T}\|_{1,\infty}\, \left\| \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i} \rangle) \right) \right\rangle \right) \right\|_{\infty} \qquad \text{(By Holder's inequality)}
\end{aligned}$$

$$
\begin{aligned}
&\leq H \left\| \left(\pi_T \left\langle \phi, W \left(\frac{1}{T} \sum_{i=0}^{T} (\hat{Q}^i_\square - \langle \phi, \boldsymbol{\theta}^i_\square \rangle) \right) \right\rangle \right) \right\|_\infty && \text{(Since } \|\mathcal{H}_{\pi_T}\|_{1,\infty} \leq H\text{)} \\
&\leq H \left\| \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{Q}^i_\square - \langle \phi, \boldsymbol{\theta}^i_\square \rangle) \right) \right\rangle \right\|_\infty && \text{(By definition of the } \pi \text{ operator)} \\
&\leq \tilde{O} \left(H^2 (1+\lambda_k) \sqrt{\frac{d}{TM}} \right) && \text{(By Lemma 22)}
\end{aligned}
$$

Combining the above relations and using that for any $|\mathcal{S}|$-dimensional vector V, $V(\rho) = \mathbb{E}_{s\sim\rho} V(s) \leq \|V\|_\infty$, we get that,

$$
\bar{V}^T_\square(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) \leq \frac{4H^2(1+\lambda_k)}{T} + \tilde{O}\left(H^2(1+\lambda_k)\sqrt{\frac{d}{TM}} \right)
$$

with probability at least $1-\delta$. □

C.2.1 Auxiliary Lemmas

Lemma 19. *For any $k \in [K]$ and $t \in [T]$, with $\square = r + \lambda_k\, c$ and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$
\|\hat{V}^t_\square\|_\infty \leq 2H(1+\lambda_k)
$$

with probability at least $1-\delta$.

Proof. First, we note that from the last line in Algorithm 2,

$$
\begin{aligned}
\hat{V}^t_\square = {} & (\pi_t \tilde{Q}^t_\square) - (\pi_{t-1} \tilde{Q}^{t-1}_\square) \\
= {} & \left(\pi_t \left\langle \phi, \sum_{i=0}^{t} \theta^i_\square \right\rangle \right) - \left(\pi_{t-1} \left\langle \phi, \sum_{i=0}^{t-1} \theta^i_\square \right\rangle \right) \\
\leq {} & \left(\pi_t \left\langle \phi, \sum_{i=0}^{t} \theta^i_\square \right\rangle \right) - \left(\pi_t \left\langle \phi, \sum_{i=0}^{t-1} \theta^i_\square \right\rangle \right) && \text{(By the greediness of } \pi_{t-1}\text{)} \\
= {} & (\pi_t \langle \phi, \theta^t_\square \rangle). && \text{(C.11)}
\end{aligned}
$$

Next, we bound the term $\langle \phi, \theta^t_\square \rangle$. We have

$$
\begin{aligned}
\left| \langle \phi, \theta^t_\square \rangle \right| &= \left| \langle \phi, W(\hat{Q}^t_\square) \rangle \right| && \text{(By the definition of } W \text{ in eq. (C.4))} \\
&\leq \left| \langle \phi, W(\langle \phi, \boldsymbol{\theta}^t_\square \rangle) \rangle \right| + \left| \langle \phi, W(\hat{Q}^t_\square) - W(\langle \phi, \boldsymbol{\theta}^t_\square \rangle) \rangle \right| && \text{(By triangle inequality)} \\
&= \left| \langle \phi, \boldsymbol{\theta}^t_\square \rangle \right| + \left| \langle \phi, W(\hat{Q}^t_\square - \langle \phi, \boldsymbol{\theta}^t_\square \rangle) \rangle \right| && \text{(Since } W(z) \text{ is linear in } z\text{)} \\
&= \left| \square + \gamma P \hat{V}^{t-1}_\square \right| + \left| \langle \phi, W(\hat{Q}^t_\square - \langle \phi, \boldsymbol{\theta}^t_\square \rangle) \rangle \right| && \text{(By the definition of } \boldsymbol{\theta}^t_\square\text{)} \\
&\leq (1 + \lambda_k + \gamma \|\hat{V}^{t-1}_\square\|_\infty) \mathbf{1} + \left| \langle \phi, W(\hat{Q}^t_\square - \langle \phi, \boldsymbol{\theta}^t_\square \rangle) \rangle \right| && \text{(Since } \square(s,a) \leq 1 + \lambda_k\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq (1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty)\mathbf{1} + \sqrt{2d} \max_{(s,a)\in\mathcal{C}} \left| \hat{Q}_{\square}^t(s,a) - (\langle \phi, \boldsymbol{\theta}_{\square}^t \rangle)(s,a) \right| \mathbf{1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By Lemma 26 with } z = \hat{Q}_{\square}^t - \langle \phi, \boldsymbol{\theta}_{\square}^t \rangle) \\
&= (1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty)\mathbf{1} \\
&+ \sqrt{2d} \max_{(s,a)\in\mathcal{C}} \left| \square(s,a) + \gamma(\hat{P}_{t-1}\hat{V}_{\square}^{t-1})(s,a) - \square(s,a) - \gamma(P\hat{V}_{\square}^{t-1})(s,a) \right| \mathbf{1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By definition of } \hat{Q}_{\square}^t \text{ and } \boldsymbol{\theta}_{\square}^t) \\
&= (1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty)\mathbf{1} + \sqrt{2d} \max_{(s,a)\in\mathcal{C}} \left| \gamma(\hat{P}_{t-1}\hat{V}_{\square}^{t-1})(s,a) - \gamma(P\hat{V}_{\square}^{t-1})(s,a) \right| \mathbf{1}
\end{aligned}
$$

$$
\implies \left| \langle \phi, \theta_{\square}^t \rangle \right| \leq (1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty)\mathbf{1} + \sqrt{2d} \frac{\|\hat{V}_{\square}^{t-1}\|_\infty}{\|\hat{V}_{\square}^{t-1}\|_\infty} \max_{(s,a)\in\mathcal{C}} \left| \gamma(\hat{P}_{t-1}\hat{V}_{\square}^{t-1})(s,a) - \gamma(P\hat{V}_{\square}^{t-1})(s,a) \right| \mathbf{1}
$$

Next, we bound the term

$$
\frac{1}{\|\hat{V}_{\square}^{t-1}\|_\infty} \max_{(s,a)\in\mathcal{C}} \left| \gamma(\hat{P}_{t-1}\hat{V}_{\square}^{t-1})(s,a) - \gamma(P\hat{V}_{\square}^{t-1})(s,a) \right|.
$$

We first note that this term is upper bounded by 2. Now, using the Azuma-Hoeffding inequality (Lemma 62) and taking a union bound over $(s,a) \in \mathcal{C}$ and $t \in [T]$, we have

$$
\mathbb{P}\left(\exists (s,a,t) \in \mathcal{C} \times [T] \ \text{ s.t. } \frac{1}{\|\hat{V}_{\square}^{t-1}\|_\infty} \max_{(s,a)\in\mathcal{C}} \left| \gamma(\hat{P}_{t-1}\hat{V}_{\square}^{t-1})(s,a) - \gamma(P\hat{V}_{\square}^{t-1})(s,a) \right| \geq \tilde{O}\left(\gamma \sqrt{\frac{1}{M}} \right) \right) \leq \delta.
$$

Therefore, with probability at least $1 - \delta$, we have

$$
\left| \langle \phi, \theta_{\square}^t \rangle \right| \leq (1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty)\mathbf{1} + \|\hat{V}_{\square}^{t-1}\|_\infty \tilde{O}\left(\gamma \sqrt{\frac{d}{M}} \right) \mathbf{1}. \tag{C.12}
$$

Given the above inequality, we can prove the claim by induction on t. Since $\hat{V}_{\square}^0 = 0$, the base case is satisfied. We assume that $\|\hat{V}_{\square}^{t-1}\|_\infty \leq 2H(1 + \lambda_k)$. By combining eq. (C.11) and eq. (C.12), we have

$$
\begin{aligned}
\|\hat{V}_{\square}^t\|_\infty &\leq \|(\pi_t \langle \phi, \theta_{\square}^t \rangle)\|_\infty \\
&\leq \|\langle \phi, \theta_{\square}^t \rangle\|_\infty && \text{(By definition of the } \pi \text{ operator)} \\
&\leq \left(1 + \lambda_k + \gamma \|\hat{V}_{\square}^{t-1}\|_\infty + \|\hat{V}_{\square}^{t-1}\|_\infty \tilde{O}\left(\gamma \sqrt{\frac{d}{M}} \right) \right) \\
&\leq \left(1 + 2H\,\gamma + 2H\,\tilde{O}\left(\gamma \sqrt{\frac{d}{M}} \right) \right) (1 + \lambda_k). && \text{(Induction hypothesis)}
\end{aligned}
$$

By taking $M \geq \tilde{O}\left(dH^2\right)$, we have

$$
\begin{aligned}
\|\hat{V}_{\square}^t\|_\infty &\leq (1 + 2H\,\gamma + 1)(1 + \lambda_k) \\
&= (2 + 2H\,\gamma)(1 + \lambda_k) \\
&\leq 2H(1 + \lambda_k) && \text{(Since } H = 1/(1-\gamma))
\end{aligned}
$$

which completes the proof. □

The following corollary is a direct consequence of eq. (C.12) and the above lemma.

Corollary 20. *For any $k \in [K]$ and $t \in [T]$, with $\square = r + \lambda_k\, c$ and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$\left\| \langle \phi, \theta^t_\square \rangle \right\|_\infty \leq 2H(1 + \lambda_k)$$

with probability at least $1 - \delta$ respectively.

Lemma 21. *For any $k \in [K]$, with $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$\left\| \frac{1}{T}(\pi_T \tilde{Q}^T_\square) - \frac{1}{T}(\pi_{T-1} \tilde{Q}^{T-1}_\square) \right\|_\infty \leq \frac{2H(1 + \lambda_k)}{T}$$

with probability at least $1 - \delta$.

Proof. By the definition of $\tilde{Q}^t_\square$ and due to the greediness of π_{T-1}, we have

$$\begin{aligned}
\frac{1}{T}(\pi_T \tilde{Q}^T_\square) - \frac{1}{T}(\pi_{T-1} \tilde{Q}^{T-1}_\square) &\leq \frac{1}{T}(\pi_T \tilde{Q}^T_\square) - \frac{1}{T}(\pi_T \tilde{Q}^{T-1}_\square) \\
&= \left(\pi_T \left\langle \phi, \frac{1}{T} \sum_{i=0}^{T} \theta^i_\square - \frac{1}{T} \sum_{i=0}^{T-1} \theta^i_\square \right\rangle \right) \\
&= \frac{1}{T}(\pi_T \langle \phi, \theta^T_\square \rangle) \\
&\leq \frac{1}{T} \left\| \langle \phi, \theta^T_\square \rangle \right\|_\infty \mathbf{1} && \text{(By definition of the } \pi \text{ operator)} \\
&\leq \frac{2H(1 + \lambda_k)}{T} \mathbf{1} && \text{(By Corollary 20)}
\end{aligned}$$

with probability at least $1 - \delta$. Similarly, by the greediness of π_T, we have

$$\begin{aligned}
\frac{1}{T}(\pi_{T-1} \tilde{Q}^{T-1}_\square) - \frac{1}{T}(\pi_T \tilde{Q}^T_\square) &\leq \frac{1}{T}(\pi_{T-1} \tilde{Q}^{T-1}_\square) - \frac{1}{T}(\pi_{T-1} \tilde{Q}^T_\square) \\
&= \left(\pi_{T-1} \left\langle \phi, \frac{1}{T} \sum_{i=0}^{T-1} \theta^i_\square - \frac{1}{T} \sum_{i=0}^{T} \theta^i_\square \right\rangle \right) \\
&= -\frac{1}{T}(\pi_T \langle \phi, \theta^T_\square \rangle) \leq \frac{1}{T} \left\| (\pi_T \langle \phi, \theta^T_\square \rangle) \right\|_\infty \\
&\leq \frac{1}{T} \left\| \langle \phi, \theta^T_\square \rangle \right\|_\infty \mathbf{1} && \text{(By definition of the } \pi \text{ operator)} \\
&\leq \frac{2H(1 + \lambda_k)}{T} \mathbf{1} && \text{(By Corollary 20)}
\end{aligned}$$

with probability at least $1 - \delta$. □

Lemma 22. *For any $k \in [K]$ and $t \in [T]$, with $\square = r + \lambda_k\, c$ and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$\left\| \left\langle \phi, W\left(\frac{1}{t} \sum_{i=1}^{t} (\hat{Q}^i_\square - \langle \phi, \boldsymbol{\theta}^i_\square \rangle) \right) \right\rangle \right\|_\infty \leq \tilde{O}\left(H(1 + \lambda_k) \sqrt{\frac{d}{tM}} \right)$$

with probability at least $1 - \delta$.

Proof.

$$\begin{aligned}
&\left\| \left\langle \phi, W\left(\frac{1}{t}\sum_{i=1}^{t}(\hat{Q}^i_\Box - \langle\phi,\boldsymbol{\theta}^i_\Box\rangle)\right)\right\rangle \right\|_\infty \\
&\le \sqrt{2d}\max_{(s,a)\in\mathcal{C}}\left|\frac{1}{t}\sum_{i=0}^{t-1}\left[\hat{Q}^i_\Box(s,a) - (\langle\phi,\boldsymbol{\theta}^i_\Box\rangle)(s,a)\right]\right| \\
&\hspace{16em}\text{(By Lemma 26 with } z = \tfrac{1}{t}\textstyle\sum_{i=0}^{t-1}[\hat{Q}^i_\Box - \langle\phi,\boldsymbol{\theta}^i_\Box\rangle]\text{)} \\
&= \sqrt{2d}\max_{(s,a)\in\mathcal{C}}\left|\frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma(\hat{P}_i\hat{V}^i_\Box)(s,a) - \gamma(P\hat{V}^i_\Box)(s,a)\right]\right| \qquad \text{(By definition of } \hat{Q}^i_\Box \text{ and } \boldsymbol{\theta}^i_\Box\text{)}
\end{aligned}$$

By Lemma 19, we have that, with probability at least $1-\delta$, the bound $\left\|\hat{V}^i_\Box\right\|_\infty \le 2H(1+\lambda_k)$ holds for all $i \in [T]$. Now, using Lemma 61 and taking the union bound over $(s,a)\in\mathcal{C}$, we have

$$\mathbb{P}\left(\exists (s,a)\in\mathcal{C} \;\text{ s.t. }\; \frac{1}{t}\sum_{i=0}^{t-1}\left[(\hat{P}_i\hat{V}^i_\Box)(s,a) - (P\hat{V}^i_\Box)(s,a)\right] \ge \tilde{O}\left(H(1+\lambda_k)\sqrt{\frac{1}{tM}}\right)\right) \le \delta.$$

Therefore, by appropriately rescaling δ, we have that with probability at least $1-\delta$,

$$\left\| \left\langle \phi, W\left(\frac{1}{t}\sum_{i=1}^{t}(\hat{Q}^i_\Box - \langle\phi,\boldsymbol{\theta}^i_\Box\rangle)\right)\right\rangle \right\|_\infty \le \tilde{O}\left(H(1+\lambda_k)\sqrt{\frac{d}{tM}}\right).$$

□

Lemma 23. *For any $k \in [K]$ and $i \in [T]$, with $\Box = r + \lambda_k c$ and $M \ge \tilde{O}\left(dH^2\right)$, we have*

$$\left\|\langle\phi, W(\hat{Q}^i_\Box - \langle\phi,\boldsymbol{\theta}^i_\Box\rangle)\rangle\right\|_\infty \le \tilde{O}\left(H(1+\lambda_k)\sqrt{\frac{d}{M}}\right)$$

with probability at least $1-\delta$.

Proof. By following a similar proof as that for the above lemma,

$$\left\|\langle\phi, W(\hat{Q}^i_\Box - \langle\phi,\boldsymbol{\theta}^i_\Box\rangle)\rangle\right\|_\infty \le \sqrt{2d}\max_{(s,a)\in\mathcal{C}}\left|\gamma\hat{P}_i\hat{V}^i_\Box(s,a) - \gamma P\hat{V}^i_\Box(s,a)\right|. \tag{C.13}$$

By Lemma 19, we have that, with probability at least $1-\delta$, $(\hat{P}_i\hat{V}^i_\Box)(s,a) \le 2H(1+\lambda_k)$ holds for all $i\in[T]$ and all (s,a). We note that by the definition of $\hat{P}_i$, $(\hat{P}_i\hat{V}^i_\Box)(s,a)$ is the empirical average of M value functions. Now, using Lemma 62 with $N = M$ and taking the union bound over $(s,a)\in\mathcal{C}$ and $i\in[T]$, we have

$$\mathbb{P}\left(\exists (s,a,t)\in\mathcal{C}\times[T] \;\text{ s.t. }\; (\hat{P}_i\hat{V}^i_\Box)(s,a) - (P\hat{V}^i_\Box)(s,a) \ge \tilde{O}\left(H(1+\lambda_k)\sqrt{\frac{1}{M}}\right)\right) \le \delta$$

Combining the above inequality with eq. (C.13) and appropriately rescaling δ completes the proof. □

Lemma 24. *For any $t \in [T]$, we have*

$$\frac{1}{t}\left\langle \phi, \sum_{i=0}^{t} \boldsymbol{\theta}_{\Box}^{i} \right\rangle = \Box + \gamma\, \frac{1}{t}\, P\,(\pi_{t-1}\tilde{Q}_{\Box}^{t-1}).$$

Proof. We first recall that by definition, $\langle \phi, \boldsymbol{\theta}_{\Box}^{t}\rangle := \Box + \gamma P \hat{V}_{\Box}^{t-1}$, $\hat{V}_{\Box}^{0} = \mathbf{0}$, and $\theta_{\Box}^{0} = \mathbf{0}$. Now we have

$$\begin{aligned}
\frac{1}{t}\left\langle \phi, \sum_{i=0}^{t} \boldsymbol{\theta}_{\Box}^{i} \right\rangle &:= \frac{1}{t}\sum_{i=0}^{t-1}(\Box + \gamma P \hat{V}_{\Box}^{i}) \\
&= \Box + \gamma P\left(\frac{1}{t}\sum_{i=0}^{t-1} \hat{V}_{\Box}^{i}\right) \\
&= \Box + \gamma P\left(\frac{1}{t}\sum_{i=0}^{t-1} \Big[(\pi_i \tilde{Q}_{\Box}^{i}) - (\pi_{i-1}\tilde{Q}_{\Box}^{i-1})\Big]\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(From the last line in Algorithm 2)} \\
&= \Box + \gamma \frac{1}{t}\, P\,(\pi_{t-1}\tilde{Q}_{\Box}^{t-1}). \qquad\qquad \text{(Telescoping Sum)}
\end{aligned}$$

□

Lemma 25. *We have*

$$\frac{1}{T}\tilde{Q}_{\Box}^{T} = \left\langle \phi, W\left(\frac{1}{T}\sum_{i=0}^{T}(\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle)\right)\right\rangle + \Box + \gamma P \frac{1}{T}(\pi_{T-1}\tilde{Q}_{\Box}^{T-1}).$$

Proof. We first recall that by the definition of W, we have $\theta_{\Box}^{i} = W(\hat{Q}_{\Box}^{i})$ and $\boldsymbol{\theta}_{\Box}^{i} = W(\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle)$. Thus,

$$\begin{aligned}
\frac{1}{T}\tilde{Q}_{\Box}^{T} &= \frac{1}{T}\tilde{Q}_{\Box}^{T} - \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle + \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle \\
&= \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \theta_{\Box}^{i}\rangle - \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle + \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle \qquad \text{(By definition of } \tilde{Q}_{\Box}^{T}\text{)} \\
&= \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \theta_{\Box}^{i} - \boldsymbol{\theta}_{\Box}^{i}\rangle + \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle \\
&= \frac{1}{T}\sum_{i=0}^{T}\Big\langle \phi, W(\hat{Q}_{\Box}^{i}) - W(\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle)\Big\rangle + \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle \\
&\qquad\qquad\qquad\qquad \text{(Since } \theta_{\Box}^{i} = W(\hat{Q}_{\Box}^{i}) \text{ and } \boldsymbol{\theta}_{\Box}^{i} = W(\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle))\text{)} \\
&= \left\langle \phi, W\left(\frac{1}{T}\sum_{i=0}^{T}(\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle\right)\right\rangle + \frac{1}{T}\sum_{i=0}^{T}\langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle \qquad \text{(}W(z)\text{ is linear in }z\text{)} \\
&= \left\langle \phi, W\left(\frac{1}{T}\sum_{i=0}^{T}(\hat{Q}_{\Box}^{i} - \langle \phi, \boldsymbol{\theta}_{\Box}^{i}\rangle\rangle\right)\right\rangle + \Box + \frac{1}{T}\,\gamma P\,(\pi_{T-1}\tilde{Q}_{\Box}^{T-1}). \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(By Lemma 24 with } t = T-1\text{)}
\end{aligned}$$

□

The following lemma bounds the extrapolation error due to the least-squares regression. It is the unweighted version (i.e., uniform weighting with $f = \mathbf{1}$) of Lemma 4.3 in Kitamura et al. [2023].

Lemma 26 (KW Bound)**.** *Let z be a function defined over $\mathcal{C}$. Then, there exists $\tilde{\rho} \in \Delta(\mathcal{S} \times \mathcal{A})$ with a finite support $\mathcal{C} := Supp(\tilde{\rho})$ of size less than or equal to $u_{\mathcal{C}}$ such that*

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left[\langle \phi(s,a), W(z)\rangle\right] \leq \sqrt{2d} \max_{(x',b')\in\mathcal{C}} \left|z(x',b')\right|,$$

where $W(z) := G^{-1} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x,b)\phi(x,b)z(x,b)$.

C.3 Proof of Lemma 12 (Optimality Guarantees for Algorithm 3 - Linear CMDP)

We define

$$\bar{\mathcal{V}}_{\diamond}^{T} := \frac{1}{T}\sum_{i=1}^{T} \hat{\mathcal{V}}_{\diamond}^{i} \overset{\text{(a)}}{=} \frac{1}{T}\left(\pi\left\langle \phi, \sum_{i=1}^{T} \omega_{\diamond}^{i}\right\rangle\right) \tag{C.14}$$

$$\langle \phi, \boldsymbol{\omega}_{\diamond}^{t}\rangle := \diamond + \gamma P \hat{\mathcal{V}}_{\diamond}^{t-1} \tag{C.15}$$

where (a) is from the last line in Algorithm 3.

Lemma 27. *For a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, algorithm 3 with $M = \tilde{O}\left(\frac{dH^2}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output $\bar{\mathcal{V}}_{\diamond}^{T}$ satisfies the following condition with probability $1-\delta$,*

$$|\bar{\mathcal{V}}_{\diamond}^{T}(\rho) - V_{\diamond}^{\pi}(\rho)| \leq O\left(\varepsilon\right).$$

Proof. Using the value difference lemma,

$$\bar{\mathcal{V}}_{\diamond}^{T} - V_{\diamond}^{\pi} = (I-\gamma P_{\pi})^{-1}(\bar{\mathcal{V}}_{\diamond}^{T} - (\pi\diamond) - \gamma P_{\pi}\bar{\mathcal{V}}_{\diamond}^{T}).$$

We now have

$$\begin{aligned}
\bar{\mathcal{V}}_{\diamond}^{T} - V_{\diamond}^{\pi} &= (I-\gamma P_{\pi})^{-1}\left[\left(\pi\left\langle \phi, \frac{1}{T}\sum_{i=1}^{T}\omega_{\diamond}^{i}\right\rangle\right) - (\pi\diamond) - \gamma P_{\pi}\bar{\mathcal{V}}_{\diamond}^{T}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(By definition of } \bar{\mathcal{V}}_{\diamond}^{T} \text{ in eq. (C.14))} \\
&= (I-\gamma P_{\pi})^{-1}\left[\left(\pi\left\langle \phi, W\left(\frac{1}{T}\sum_{i=1}^{T}(\hat{\mathcal{Q}}_{\diamond}^{i} - \langle\phi,\boldsymbol{\omega}_{\diamond}^{i}\rangle)\right)\right\rangle\right) + (\pi\diamond) + \gamma P_{\pi}\left(\pi\left\langle \phi, \frac{1}{T}\sum_{i=1}^{T-1}\omega_{\diamond}^{i}\right\rangle\right)\right. \\
&\quad \left. - (\pi\diamond) - \gamma P_{\pi}\bar{\mathcal{V}}_{\diamond}^{T}\right] \qquad\qquad\qquad\qquad\qquad\qquad\text{(By Lemma 29)} \\
&= (I-\gamma P_{\pi})^{-1}\left[\left(\pi\left\langle \phi, W\left(\frac{1}{T}\sum_{i=1}^{T}(\hat{\mathcal{Q}}_{\diamond}^{i} - \langle\phi,\boldsymbol{\omega}_{\diamond}^{i}\rangle)\right)\right\rangle\right) + \gamma P_{\pi}\left(\pi\left\langle \phi, \frac{1}{T}\sum_{i=1}^{T-1}\omega_{\diamond}^{i}\right\rangle\right) - \gamma P_{\pi}\bar{\mathcal{V}}_{\diamond}^{T}\right]
\end{aligned}$$

$$
\begin{aligned}
&= (I - \gamma P_\pi)^{-1} \left[\left(\pi \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right) + \gamma P_\pi \left(\pi \left\langle \phi, \frac{1}{T} \sum_{i=1}^{T-1} \omega_\diamond^i \right\rangle \right) \right. \\
&\quad \left. - \gamma P_\pi \left(\pi \left\langle \phi, \frac{1}{T} \sum_{i=1}^{T} \omega_\diamond^i \right\rangle \right) \right] \qquad \text{(By definition of } \bar{\mathcal{V}}_\diamond^T \text{ in eq. (C.14))} \\
&= (I - \gamma P_\pi)^{-1} \left[\left(\pi \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right) - \gamma P_\pi \frac{1}{T} (\pi \langle \phi, \omega_\diamond^T \rangle) \right].
\end{aligned}
$$

Taking the infinity norm and using the triangle inequality,

$$
\begin{aligned}
\left\| \bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi \right\|_\infty &\le \left\| (I - \gamma P_\pi)^{-1} \left(\pi \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right) \right\|_\infty \\
&\quad + \left\| (I - \gamma P_\pi)^{-1} \gamma P_\pi \frac{1}{T} (\pi \langle \phi, \omega_\diamond^T \rangle) \right\|_\infty \\
&\le \left\| (I - \gamma P_\pi)^{-1} \right\|_{1,\infty} \left\| \pi \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right\|_\infty \\
&\quad + \left\| (I - \gamma P_\pi)^{-1} \gamma P_\pi \right\|_{1,\infty} \left\| \frac{1}{T} (\pi \langle \phi, \omega_\diamond^T \rangle) \right\|_\infty \qquad \text{(By Holder's inequality)} \\
&\le H \left[\left\| \pi \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right\|_\infty + \left\| \frac{1}{T} (\pi \langle \phi, \omega_\diamond^T \rangle) \right\|_\infty \right] \\
&\qquad \text{(Since } \left\| (I - \gamma P_\pi)^{-1} \right\|_{1,\infty} \le H, \ \left\| (I - \gamma P_\pi)^{-1} \gamma P_\pi \right\|_{1,\infty} \le H) \\
&\le H \left[\left\| \left\langle \phi, W \left(\frac{1}{T} \sum_{i=1}^{T} (\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle) \right) \right\rangle \right\|_\infty + \left\| \frac{1}{T} \langle \phi, \omega_\diamond^T \rangle \right\|_\infty \right] \\
&\qquad \text{(By definition of the } \pi \text{ operator)} \\
&\le \tilde{O} \left(H^2 \sqrt{\frac{d}{TM}} + \frac{H^2}{T} \right). \qquad \text{(By Lemma 28 and Lemma 30)}
\end{aligned}
$$

Using that for any $|\mathcal{S}|$-dimensional vector V, $V(\rho) = \mathbb{E}_{s \sim \rho} |V(s)| \le \|V\|_\infty$ completes the proof. □

C.3.1 Auxiliary Lemmas

Since the updates in Algorithm 3 are a special case of those in Algorithm 2, the proofs of the auxiliary lemmas are analogous. We therefore present lemmas analogous to Lemma 19, Lemma 25 and Lemma 22, whose proofs follow by the same reasoning.

Lemma 28. *For any $t \in [T]$, with $\diamond = r$ or c and $M \ge \tilde{O}\left(dH^2\right)$, we have*

$$
\|\langle \phi, \omega_\diamond^t \rangle\|_\infty \le 2H \quad \textit{and} \quad \|\hat{\mathcal{V}}_\diamond^t\|_\infty \le 2H
$$

with probability at least $1 - \delta$.

Lemma 29. *For any $t \in [T]$, $\diamond = r$ or c, we have*

$$\left\langle \phi, \frac{1}{T}\sum_{i=1}^{T} \omega_\diamond^i \right\rangle = \left\langle \phi, W\left(\frac{1}{T}\sum_{i=1}^{T}(\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle)\right)\right\rangle + \diamond + \gamma P\left(\pi\left\langle \phi, \frac{1}{T}\sum_{i=1}^{T-1} \omega_\diamond^i \right\rangle\right).$$

Lemma 30. *With $\diamond = r$ or c and $M \geq \tilde{O}\left(dH^2\right)$, we have*

$$\left\| \left\langle \phi, W\left(\frac{1}{T}\sum_{i=1}^{T}(\hat{\mathcal{Q}}_\diamond^i - \langle \phi, \boldsymbol{\omega}_\diamond^i \rangle)\right)\right\rangle \right\|_\infty \leq \tilde{O}\left(H\sqrt{\frac{d}{TM}}\right)$$

with probability at least $1-\delta$.

C.4 Proof of Corollary 13

Corollary 31. *Using `LS-MDVI` (algorithm 2) and `LS-PE` (algorithm 3) as instantiations of the `MDP-Solver` and `PolicyEvaluation` in algorithm 1 and using the `DataCollection` oracle described in section 4.1 has the following guarantee: for a fixed $\varepsilon \in (0,1]$, $\delta \in (0,1)$, algorithm 1 with $\tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$ samples, $U = O\left(\frac{1}{\zeta(1-\gamma)}\right)$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = O\left(\frac{1}{\varepsilon^2\,(1-\gamma)^2}\right)$, and $b' = b - O(\varepsilon)$, returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b - O(\varepsilon).$$

With the same algorithm parameters, but with $b' = b + O(\varepsilon)$ and $\tilde{O}\left(\frac{d^2H^6}{\zeta^2\varepsilon^2}\right)$ samples, algorithm 1 returns a mixture policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b.$$

Proof. By Lemma 10 and Lemma 12, the sample complexity required to ensure $f(\mathcal{B}) \leq O(\varepsilon)$ is $TM|\mathcal{C}| = \tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$. Therefore, the guarantee for the relaxed feasibility setting follows directly from our meta-theorem (Theorem 7). For the strict feasibility setting, we rescale ε by a factor of $O(\zeta(1-\gamma))$. Since $\varepsilon \leq 1$ and $1-\gamma \leq 1$, the condition of $f(\mathcal{B}) \leq \zeta/6$ in Theorem 7 can be satisfied. The rescaling increases the sample complexity by a multiplicative factor of $\frac{1}{\zeta^2(1-\gamma)^2}$, thereby completing the proof. □

C.5 Instantiating the `MDP-Solver`: G-Sampling-and-Stop

Instead of `LS-MDVI`, we can instantiate the linear `MDP-Solver` in Algorithm 1 with the `GSS` algorithm [Taupin et al., 2023]. The `GSS` algorithm begins by computing a G-optimal sampling distribution over state-action pairs that minimizes the worst-case variance of value estimates. It then repeatedly samples transitions and rewards according to this distribution and uses regularized least-squares estimators to learn the reward and transition parameters of the MDP. For an arbitrary distribution $\tilde{\rho}$ over $\mathcal{S} \times \mathcal{A}$, let $G \in \mathbb{R}^{d \times d}$ and $g(\tilde{\rho}) \in \mathbb{R}$ be defined as:

$$G := \sum_{(x,b) \in \mathcal{C}} \tilde{\rho}(x,b)\, \phi(x,b)\phi(x,b)^\top \qquad \text{and} \qquad g(\tilde{\rho}) := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \langle \phi(s,a), G^{-1}\phi(s,a) \rangle,$$

The `GSS` method samples one state-action pair $(s_t, a_t) \sim \rho^*$ in an iteration t where $\rho^* := \arg\min_{\rho \in \Delta_{\mathcal{S} \times \mathcal{A}}} g(\rho)$. We denote this data collection procedure as `DataCollection-GSS`. Note that this is different than the sampling scheme used in chapter A.

For solving a linear MDP, the `GSS` algorithm uses a stopping rule based on confidence bounds derived from matrix concentration inequalities, and determines when the estimates are accurate enough to ensure that the returned policy is ε-optimal for the true MDP with high probability. The stopping time is denoted by

$$\tau = \inf\{t \geq 1 : Z(t) \geq \beta(t)\}$$

where $\beta(t)$ is a certain threshold and $Z(t)$ is the quantity we seek to control in order to achieve the desired sample complexity. Their main result in the setting of infinite-horizon γ-discounted linear unconstrained MDPs is stated below.

Theorem 32 (Theorem 2 and Theorem 3 in [Taupin et al., 2023])**.** *Let $\varepsilon, \delta \in (0,1)$. The `GSS` algorithm returns an ε-optimal policy with probability at least $1-\delta$, and the expected number of samples used is bounded by*

$$O\left(\frac{d}{(1-\gamma)^4\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + d\log\left(\frac{d}{(1-\gamma)^4\varepsilon^2}\right)\right)\right).$$

Using the `GSS` algorithm as an alternative instantiation of `MDP-Solver`$(r + \lambda_k c, \mathcal{B}, \phi)$, we have that, with $N = \tilde{O}\left(\frac{d^2H^4}{\varepsilon^2}\right)$, the `GSS` algorithm satisfies Assumption 4 with $f_{\mathrm{mdp}}(\mathcal{B}) = O(\varepsilon)$. Hence, instantiating the three oracles by `DataCollection-GSS`, the `GSS` algorithm and using the same `PolicyEvaluation` oracle as in algorithm 3, we can use our meta-theorem (Theorem 7) to obtain the same sample complexity bounds as in theorem 13.

Appendix D

Lower Bound

Theorem 14. *Let $\delta \in (0, 0.08]$, $\gamma \in [7/12, 1)$, $H = 1/(1-\gamma)$, $\varepsilon \in (0, 0.002)$, $\zeta \in (0, 49/2280)$, $b \in [H/2, H]$, and $d \geq 6$. There exists a class of linear constrained MDPs such that any (ε, δ)-sound algorithm requires $\Omega\left(d^2 H^5/\varepsilon^2\zeta^2\right)$ samples from the generative model in the worst case.*

Proof. We construct a lower bound by using the ideas from Vaswani et al. [2022], Weisz et al. [2022].

Hard Instance. We construct a class of hard linear CMDPs, where each individual CMDP, denoted as $\mathcal{M}_\beta$, is parameterized by a vector β. Each MDP in $\mathcal{M}_\beta$ has four states: $\mathcal{S} = \{o, s_0, s_1, z\}$ with o being the initial state. The action space is $\mathcal{A} = \bar{\mathcal{A}} \times \{0, 1\} = \{\pm 1/\sqrt{d-5}\}^{d-5} \times \{0, 1\}$, where $\bar{\mathcal{A}}$ is a subset of a $(d-5)$-dimensional hypercube. Thus, each action can be written as $a = (\bar{a}^\top, a')^\top \in \mathcal{A}$ where $\bar{a}$ is a $d-5$ dimensional vector and $a' \in \{0,1\}$ is a scalar. Recall from Assumption 1 that the reward $r(s,a) = \langle \phi(s,a), \psi_r \rangle$, cost $c(s,a) = \langle \phi(s,a), \psi_c \rangle$, and transition probability $\mathcal{P}(\cdot|s,a) = \langle \phi(s,a), \mu \rangle$ are all defined as linear functions. Let $\beta \in \bar{\mathcal{A}}$, $\Delta \in (0, 0.2(1-\gamma)]$, $u = \frac{1-\gamma^2-\gamma\Delta}{\gamma}(b-\zeta)$. Note that the parameter ζ is the Slater constant for this CMDP instance. As can be verified from eq. (D.3), there exists a policy that achieves a constraint value of $b + \zeta$, satisfying the definition. Now, we define the following parameters.

$$\begin{aligned}
\phi(s_0, a) &= (1, 0, 0, 0, 0, \bar{a}^\top)^\top, & \phi(s_1, a) &= (0, 1, 0, 0, 0, \ldots, 0)^\top, \\
\phi(o, a) &= (0, 0, a', 1-a', 0, \ldots, 0)^\top, & \phi(z, a) &= (0, 0, 0, 1, 1, 0, \ldots, 0)^\top \\
\mu(s_0) &= (\gamma, 0, 1, 0, 0, \Delta\beta^\top)^\top, & \mu(s_1) &= (1-\gamma, 1, 0, 0, 0, -\Delta\beta^\top)^\top, \\
\mu(o) &= (0, \ldots, 0)^\top, & \mu(z) &= (0, 0, 0, 1, 0, \ldots, 0)^\top, \\
\psi_r &= (1, 0, \ldots, 0)^\top, & \psi_c &= (u, 0, 0, 0, (b+\zeta)(1-\gamma)/\gamma, 0 \ldots, 0)^\top.
\end{aligned}$$

Under the above parameter settings, the transition dynamics is defined as follows.

$$\begin{aligned}
\mathcal{P}_\beta(s_0|s_0, a) &= \gamma + \Delta\beta^\top \bar{a}, & \mathcal{P}_\beta(s_1|s_0, a) &= 1 - \gamma - \Delta\beta^\top \bar{a}, \\
\mathcal{P}_\beta(s_0|o, a) &= a', & \mathcal{P}_\beta(z|o, a) &= 1 - a',
\end{aligned}$$

States s_1 and z are absorbing. The reward and constraint functions are then specified as follows. For any action a,

$$\begin{aligned} r(s_0, a) &= 1, & r(z, a) &= 0, \\ c(s_0, a) &= u, & c(z, a) &= (b+\zeta)(1-\gamma)/\gamma. \end{aligned}$$

Rewards and costs for other states are zero. By solving the Bellman equation $V^{\pi}_{r,\beta}(s_0) = 1 + \gamma(\gamma + \Delta \mathbb{E}_{\bar{a}\sim\pi}[\bar{a}^\top \beta])V^{\pi}_{r,\beta}(s_0)$, we can define the value functions for a policy π in the MDP $\mathcal{M}_\beta$. The notation $V^{\pi}_{r,\beta}$ represents the reward value function for policy π within the specific MDP instance defined by β. Similarly, $V^{\pi}_{c,\beta}$ is the cost value function. The value functions starting from the initial state o are

$$V^{\pi}_{r,\beta}(o) = \gamma \sum_{a\in\mathcal{A}} a'\pi(a|o)V^{\pi}_{r,\beta}(s_0) = \frac{\gamma \sum_{a\in\mathcal{A}} a'\pi(a|o)}{1-\gamma^2-\gamma\Delta\mathbb{E}_{\bar{a}\sim\pi}[\bar{a}^\top\beta]}, \tag{D.1}$$

$$\begin{aligned} V^{\pi}_{c,\beta}(o) &= \gamma \sum_{a\in\mathcal{A}} a'\pi(a|o)V^{\pi}_{c,\beta}(s_0) + \gamma \sum_{a\in\mathcal{A}} (1-a')\pi(a|o)V^{\pi}_{c,\beta}(z) \\ &= \frac{\gamma u \sum_{a\in\mathcal{A}} a'\pi(a|o)}{1-\gamma^2-\gamma\Delta\mathbb{E}_{\bar{a}\sim\pi}[\bar{a}^\top\beta]} + \sum_{a\in\mathcal{A}} \pi(a|o)(1-a')(b+\zeta). \end{aligned} \tag{D.2}$$

Note that $V^{\pi}_{r,\beta}(s_1) = 0$ and $V^{\pi}_{c,\beta}(z) = \frac{1}{1-\gamma}\cdot\frac{(b+\zeta)(1-\gamma)}{\gamma} = \frac{b+\zeta}{\gamma}$. We now define the probability of a policy π transitioning from o to s_0 as $p := \mathcal{P}^{\pi}_{\beta}[o \to s_0]$. According to the hard instance construction, $\mathcal{P}_\beta(s_0|o,a) = a'$. Thus, p is the marginal probability under π that $a' = 1$ when in state o. This is equivalent to stating that the policy π must satisfy:

$$\sum_{\bar{a}\in\bar{\mathcal{A}}} \pi((\bar{a}^\top, 1)^\top|o) = p \quad \text{and} \quad \sum_{\bar{a}\in\bar{\mathcal{A}}} \pi((\bar{a}^\top, 0)^\top|o) = 1-p.$$

Since we are in the strict feasibility setting, the constraint has to be satisfied as $V^{\pi}_{c,\beta}(o) \ge b$ for a feasible policy π. We denote the optimal feasible policy for the CMDP $\mathcal{M}_\beta$ as π^*_β. Next, we show that the policy π^*_β must satisfy:

$$\sum_{\bar{a}\in\bar{\mathcal{A}}} \pi^*_\beta((\bar{a}^\top, 1)^\top|o) = \frac{1}{2} \quad \text{and} \quad \sum_{a'\in\{0,1\}} \pi^*_\beta((\beta^\top, a')^\top|s_0) = 1.$$

Since states z and s_1 are absorbing, the choice of policies on them is irrelevant. It is clear from eq. (D.1) that for any fixed p, $V^{\pi}_{r,\beta}(o)$ is maximized when $\mathbb{E}_{\bar{a}\sim\pi}[\bar{a}^\top\beta] = \beta^\top\beta = 1$. In this case, $V^{\pi}_{r,\beta}(o) = \frac{\gamma p}{1-\gamma^2-\gamma\Delta}$. Next, when $\mathbb{E}_{\bar{a}\sim\pi}[\bar{a}^\top\beta] = \beta^\top\beta$, we observe that

$$\begin{aligned} V^{\pi}_{c,\beta}(o) &= \frac{\gamma u p}{1-\gamma^2-\gamma\Delta} + (1-p)(b+\zeta) \\ &= p(b-\zeta) + (1-p)(b+\zeta) & \left(u = \tfrac{1-\gamma^2-\gamma\Delta}{\gamma}(b-\zeta)\right) \\ &= b + (1-2p)\zeta. \end{aligned} \tag{D.3}$$

Therefore, for a policy to be feasible, it needs to ensure $p \le 1/2$. The value function is then maximized when $p = 1/2$. It means that, at the initial state o, the optimal policy should maintain an equal probability of moving into s_0 and z. At state s_0, the optimal policy should pick an action from the set $\{(\beta^\top, 0)^\top, (\beta^\top, 1)^\top\}$ with probability 1. Note that for an action

$a = (\bar{a}^\top, a')^\top$, the constraint satisfaction depends only on the action component a', and the value function at state s_0 is determined by $\bar{a}$. Thus, $\bar{a}$ and a' are optimized independently. For this optimal policy π^*_β, we have

$$V_{r,\beta}^{\pi^*_\beta}(o) = \frac{\gamma}{2(1-\gamma^2-\gamma\Delta)}, \quad \text{and} \quad V_{c,\beta}^{\pi^*_\beta}(o) = b.$$

Now, consider an algorithm which incorrectly identifies the true parameter β as a different parameter β'. The algorithm then outputs a policy $\hat{\pi}$. According to this misestimated parameter, the policy satisfies $\mathbb{E}_{\bar{a}\sim\hat{\pi}}[\bar{a}^\top\beta] = \beta'^\top\beta$. Hence, in the true CMDP $\mathcal{M}_\beta$, its corresponding value function involves $\beta'^\top\beta$. Suppose the probability of transition to s_0 from o with this policy is $\hat{p}$. We have

$$\begin{aligned}
V_{c,\beta}^{\hat{\pi}}(o) &= \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta\beta'^\top\beta} + (1-\hat{p})(b+\zeta) \\
&= \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta} + (1-\hat{p})(b+\zeta) - \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta} + \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta\beta'^\top\beta} \\
&= \hat{p}(b-\zeta) + (1-\hat{p})(b+\zeta) - \varepsilon_c \qquad \left(u = \tfrac{1-\gamma^2-\gamma\Delta}{\gamma}(b-\zeta)\right) \\
&= b + (1-2\hat{p})\zeta - \varepsilon_c
\end{aligned}$$

where

$$\varepsilon_c = \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta} - \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta\beta'^\top\beta} = \frac{\gamma^2 u\hat{p}\Delta(1-\beta'^\top\beta)}{(1-\gamma^2-\gamma\Delta)(1-\gamma^2-\gamma\Delta\beta'^\top\beta)}.$$

Therefore, in order for $\hat{\pi}$ to be feasible in the original CMDP $\mathcal{M}_\beta$, we require that

$$(1-2\hat{p})\zeta \geq \varepsilon_c \Rightarrow \hat{p} \leq \frac{1}{2} - \frac{\varepsilon_c}{2\zeta}. \tag{D.4}$$

In the meanwhile, for a policy $\hat{\pi}$ to be (ε,δ)-sound, it must satisfied that, with probability $1-\delta$,

$$\varepsilon \geq V_{r,\beta}^{\pi^*_\beta}(o) - V_{r,\beta}^{\hat{\pi}}(o) = \frac{\gamma}{2(1-\gamma^2-\gamma\Delta)} - \frac{\hat{p}\gamma}{(1-\gamma^2-\gamma\Delta\beta'^\top\beta)}$$

which implies

$$\begin{aligned}
\hat{p} &\geq \frac{(1-\gamma^2-\gamma\Delta\beta'^\top\beta)}{2(1-\gamma^2-\gamma\Delta)} - \varepsilon(1-\gamma^2-\gamma\Delta\beta'^\top\beta)/\gamma \\
&\geq \frac{(1-\gamma^2-\gamma\Delta)}{2(1-\gamma^2-\gamma\Delta)} - \varepsilon(1-\gamma^2-\gamma\Delta\beta'^\top\beta)/\gamma \qquad (\beta'^\top\beta \leq 1) \\
&= \frac{1}{2} - \varepsilon(1-\gamma^2-\gamma\Delta\beta'^\top\beta)/\gamma.
\end{aligned}$$

This lower bound implies that achieving higher rewards requires the policy to transition less frequently into the safe state. Since $\varepsilon \in (0, 0.002)$, $\gamma \in [7/12, 1)$, $\Delta \leq 0.2(1-\gamma)$, and

${\beta'}^\top\beta \geq -1$, we have

$$\varepsilon(1-\gamma^2-\gamma\Delta{\beta'}^\top\beta)/\gamma \leq \varepsilon(1-\gamma^2+0.2\gamma(1-\gamma))/\gamma \leq \frac{3}{2}\varepsilon \leq 0.003$$

and thus

$$\hat{p} \geq 0.5 - 0.003 \geq 0.4. \tag{D.5}$$

Now we analyze parameter u. Recall that $\zeta \in (0, 49/2280)$ $\Delta \leq 1-\gamma$, $b \in [H/2, H]$, and $H = \frac{1}{1-\gamma} \in [12/5, \infty)$. We obtain

$$u = \frac{1-\gamma^2-\gamma\Delta}{\gamma}(b-\zeta) \geq \frac{1-\gamma}{\gamma}\left(\frac{1}{2(1-\gamma)} - \frac{49}{2280}\right) \geq \frac{1}{2} - \frac{49}{2280} \geq \frac{1}{4}.$$

Next, we denote τ as the total number of queries a planner sends to the generative model before it stops its process and outputs a final policy. For any (ε,δ)-sound algorithm outputting a policy $\hat{\pi}$ with query complexity τ, we have (with probability $1-\delta$)

$$\begin{aligned}
\varepsilon &\geq V_{r,\beta}^{\pi_\beta^*}(o) - V_{r,\beta}^{\hat{\pi}}(o) \\
&= \frac{\gamma}{2(1-\gamma^2-\gamma\Delta)} - \frac{\hat{p}\gamma}{(1-\gamma^2-\gamma\Delta{\beta'}^\top\beta)} \\
&\geq \frac{\gamma}{2(1-\gamma^2-\gamma\Delta)} - \frac{\hat{p}\gamma}{(1-\gamma^2-\gamma\Delta)} && ({\beta'}^\top\beta \leq 1) \\
&= \frac{\gamma}{(1-\gamma^2-\gamma\Delta)}\left(\frac{1}{2} - \hat{p}\right) \\
&\geq \frac{\gamma}{(1-\gamma^2)}\left(\frac{1}{2} - \hat{p}\right) && (\Delta \geq 0) \\
&= \frac{\gamma}{(1-\gamma)(1+\gamma)}\left(\frac{1}{2} - \hat{p}\right) \\
&\geq \frac{7}{19(1-\gamma)}\left(\frac{1}{2} - \hat{p}\right) && (\gamma \in [7/12, 1)) \\
&\geq \frac{7\varepsilon_c}{19\zeta(1-\gamma)} && \text{(by eq. (D.4))} \\
&= \frac{7}{19\zeta(1-\gamma)}\left(\frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta} - \frac{\gamma u\hat{p}}{1-\gamma^2-\gamma\Delta{\beta'}^\top\beta}\right) \\
&= \frac{7\gamma u\hat{p}}{19\zeta(1-\gamma)}\left(\frac{1}{1-\gamma^2-\gamma\Delta} - \frac{1}{1-\gamma^2-\gamma\Delta{\beta'}^\top\beta}\right) \\
&= \frac{7\gamma u\hat{p}}{19\zeta(1-\gamma)}\left(V_{r,\beta}^{\pi_\beta^*}(s_0) - V_{r,\beta}^{\hat{\pi}}(s_0)\right) \\
&\geq \frac{49}{2280\zeta(1-\gamma)}\left(V_{r,\beta}^{\pi_\beta^*}(s_0) - V_{r,\beta}^{\hat{\pi}}(s_0)\right). && \text{(by the lower bound on } \gamma, u, \text{ and } \hat{p})
\end{aligned}$$

Note that once the agent enters state s_0, the CMDP framework no longer plays a role. At this point, the problem has been reduced to establishing a lower bound for unconstrained linear MDPs where the starting state distribution is s_0, a result proven in Weisz et al. [2022].

Note that the hard instance constructed in Weisz et al. [2022] is identical to ours after the initial transition from initial state o to s_0.

Specifically, as shown in their Theorem H.3, under the parameter settings $\delta \in (0, 0.08]$, $\gamma \in [7/12, 1)$, $H = 1/(1-\gamma)$, $\alpha \in (0, 0.005\gamma H/(1+\gamma)^2)$, $\Delta \leq 0.2(1-\gamma)$, and $d \geq 3$, if τ is the number of samples, then,

$$\mathbb{P}_\beta \left(V_{r,\beta}^{\pi_\beta^*}(s_0) - V_{r,\beta}^{\hat{\pi}}(s_0) \geq \alpha \right) \geq \frac{3}{35} - \frac{8}{35}\sqrt{1 - \exp\left(-\frac{5\Delta^2 H \mathbb{E}_\beta[\tau]}{(d-2)^2}\right)}$$

which implies that the algorithm is not (α, δ) sound unless $\mathbb{E}_\beta[\tau] \geq \Omega(d^2 H^3/\alpha^2)$. Substituting $\alpha = \frac{2280\varepsilon\zeta(1-\gamma)}{49}$ completes our proof. We note that the range of α in Weisz et al. [2022] is $(0, 0.005\gamma H/(1+\gamma)^2)$ where $0.005\gamma H/(1+\gamma)^2 \geq 0.00279$. Replacing $\alpha = \frac{2280\varepsilon\zeta(1-\gamma)}{49} \leq \varepsilon \in (0, 0.002)$ satisfies their choice of range. Except for the range of ε and the lower bound on d, other parameter choices exactly match those in Weisz et al. [2022]. □

Appendix E

Algorithms for Solving Tabular CMDPs

Algorithm 6 Tabular Mirror Descent Value Iteration (`Tabular-MDVI`)

Input: T (number of iterations), M (number of next-state samples obtained per state-action pair in each iteration), $\square$ (rewards in MDP), $\mathcal{B} = \mathcal{B}_0 \cup \cdots \cup \mathcal{B}_{T-1}$ (Buffer).

Output: π_T where $\forall s \in \mathcal{S} : \pi_T(\cdot|s) \in \arg\max_a \tilde{Q}^T_\square(s,a)$.

Define $\hat{V}^0_\square = \mathbf{0}$, $\hat{Q}^{-1}_\square = \mathbf{0}$.

```
procedure TABULAR-MDVI($T$, $M$, $\square$, $\mathcal{B}$)
    for $t = 0, 1, 2 \ldots, T-1$ do
        $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ : Access $(s, a, s'_m)_{m=1}^M$ from the buffer $\mathcal{B}_t$.
        $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \hat{Q}^t_\square(s,a) = \square(s,a) + \gamma \frac{1}{M} \sum_{m=1}^M \hat{V}^t_\square(s'_m)$.
        Define $\tilde{Q}^t_\square = \sum_{i=0}^t \hat{Q}^i_\square$; $\forall s \in \mathcal{S} : \hat{V}^{t+1}_\square(s) = \max_a\{\tilde{Q}^t_\square(s,a)\} - \max_a\{\tilde{Q}^{t-1}_\square(s,a)\}$.
    end for
end procedure
```

Algorithm 7 Tabular Policy Evaluation (`Tabular-PE`)

Input: T (number of iterations), M (number of next-state samples obtained per state-action pair in each iteration), $\diamond$ (either r or c), $\mathcal{B} = \mathcal{B}_0 \cup \cdots \cup \mathcal{B}_{T-1}$ (Buffer), π (policy to be evaluated).

Output: $\bar{\mathcal{V}}^T_\diamond(\rho) = \frac{1}{T} \sum_{i=1}^T \hat{\mathcal{V}}^i_\diamond(\rho)$.

Define $\hat{\mathcal{V}}^0_\diamond = \mathbf{0}$.

```
procedure TABULAR-PE($T$, $M$, $\diamond$, $\mathcal{B}$, $\pi$)
    for $t = 0, 1, 2 \ldots, T-1$ do
        $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ : Access $(s, a, s'_m)_{m=1}^M$ from the buffer $\mathcal{B}_t$.
        $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \hat{\mathcal{Q}}^t_\diamond(s,a) = \diamond(s,a) + \gamma \frac{1}{M} \sum_{m=1}^M \hat{\mathcal{V}}^t_\diamond(s'_m)$.
        $\hat{\mathcal{V}}^{t+1}_\diamond = \pi \hat{\mathcal{Q}}^t_\diamond$.
    end for
end procedure
```

Appendix F

Instantiating the Framework for Tabular Constrained MDPs

We now instantiate the framework for tabular CMDPs, and prove that the resulting algorithm attains near-optimal sample complexity. In contrast to the linear setting, we set $\mathcal{C} = \mathcal{S} \times \mathcal{A}$ as the input to the `DataCollection` oracle. For the `MDP-Solver` and `PolicyEvaluation`, we adapt algorithms 2 and 3 to the tabular setting. In particular, for both these algorithms, we set the features to be $|\mathcal{S}||\mathcal{A}|$ dimensional one-hot encodings of the state-action space implying that the feature map ϕ is an $|\mathcal{S}||\mathcal{A}|$-dimensional identity matrix. Consequently, the resulting algorithm does not require linear regression to estimate the Q-function. We provide the pseudo-code for these two instantiations is provided in chapter E. Their corresponding optimality guarantees are proved in chapter G and stated below.

Lemma 33. *For a fixed $\varepsilon \in (0, 1/H^2]$, $\delta \in (0, 1)$, any $k \in [K]$, and $T \geq 2\log(T)/\gamma$, when using algorithm 6 at iteration k of algorithm 1 with $\square = r + \lambda_k c$, $M = \tilde{O}\left(\frac{H}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output policy π_T satisfies the following condition with probability $1 - \delta$,*

$$\max_{\pi} V^{\pi}_{r+\lambda_k c}(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) \leq O((1 + \lambda_k)\varepsilon)\,,$$

The resulting sample complexity is $N = T\,M|\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$.

Lemma 34. *For a fixed $\varepsilon \in (0, H]$, $\delta \in (0, 1)$, algorithm 7 with $M = \tilde{O}\left(\frac{H}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output $\bar{\mathcal{V}}^T_\diamond$ satisfies the following condition with probability $1 - \delta$,*

$$|\bar{\mathcal{V}}^T_\diamond(\rho) - V^{\pi}_\diamond(\rho)| \leq O\left(\varepsilon\right)\,,$$

The resulting sample complexity is $N = T\,M|\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$.

The proofs of theorems 33 and 34 can use the total variance technique and a Bernstein-type concentration argument [Azar et al., 2013, Kozuno et al., 2022] and result in near-optimal bounds in the tabular setting. Moreover, the corresponding algorithms do not require constructing coresets or using (variance-weighted) linear regression. Consequently, unlike the linear setting in chapter 4, the same buffer $\mathcal{B}$ can be reused across all iterations of algorithm 1. This allows the near-optimal sample complexities of both algorithms 6 and 7 to be preserved for tabular CMDPs. In particular, we prove the following result in section G.4.

Corollary 35. *Let algorithm 6 and algorithm 7 be the instantiations of the `MDP-Solver` and `PolicyEvaluation` in algorithm 1. For a fixed $\varepsilon \in (0, 1/H^2]$, $\delta \in (0,1)$, algorithm 1 with $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$ samples, $U = O\left(\frac{1}{\zeta(1-\gamma)}\right)$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = O\left(\frac{1}{\varepsilon^2\,(1-\gamma)^2}\right)$, and $b' = b - O(\varepsilon)$, returns a policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \text{and} \quad V_c^{\bar{\pi}}(\rho) \geq b - O(\varepsilon).$$

Under the same conditions, but with $b' = b + O(\varepsilon)$ and $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^5}{\zeta^2\varepsilon^2}\right)$ samples, algorithm 1 returns a policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \text{and} \quad V_c^{\bar{\pi}}(\rho) \geq b.$$

The above result matches the near-optimal sample complexity bounds attained by the model-based algorithm in Vaswani et al. [2022]. Furthermore, instantiating the `MDP-Solver` to be the model-based algorithm [Agarwal et al., 2020, Li et al., 2020] and using algorithm 1 will result in a near-optimal sample complexity for solving tabular CMDPs (see section G.5 for details). Note that the `MDP-Solver` can also be instantiated by a range of model-free algorithms for solving unconstrained MDPs with access to a generative model [Azar et al., 2013, Jin et al., 2024, Sidford et al., 2018, 2023, Wang, 2017]. Consequently, our framework can be interpreted as a generalization of the the primal-dual approach in [Vaswani et al., 2022] to handle model-free algorithms and linear function approximation.

Appendix G

Proofs for Section F

Throughout, we treat π as an operator that returns an $|\mathcal{S}|$-dimensional vector s.t. for an arbitrary $|\mathcal{S}||\mathcal{A}|$-dimensional vector u such that $(\pi u)(s) := \sum_{a \in \mathcal{A}} \pi(a|s)\, u(s,a)$. Furthermore, we define $P_\pi := \pi P$ where $P_\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and denotes the transition probability matrix induced by policy π. We also recall that $\pi_k^* := \arg\max_\pi V^\pi_{r+\lambda_k c}$ and define $\bar{V}^T_\square := \frac{1}{T}\sum_{i=1}^T \hat{V}^i_\square$. We define $y_{t,m,s,a}$ to be the m-th next-state sample s'_m corresponding to the state-action pair (s,a) at iteration t. For a value function V, $\mathrm{Var}(V)$ denote the function

$$\mathrm{Var}(V) : (s,a) \mapsto (PV^2)(s,a) - (PV)^2(s,a)$$

and $\sigma(V) := \sqrt{\mathrm{Var}(V)}$.

G.1 Proof of Lemma 33 (Optimality Guarantees for Algorithm 6 - Tabular CMDP)

Lemma 36. *For a fixed $\varepsilon \in (0, 1/H^2]$, $\delta \in (0,1)$, any $k \in [K]$, and $T \geq 2\log(T)/\gamma$, when using algorithm 6 at iteration k of algorithm 1 with $\square = r + \lambda_k c$, $M = \tilde{O}\left(\frac{H}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output policy π_T satisfies the following condition with probability $1-\delta$,*

$$\max_\pi V^\pi_{r+\lambda_k c}(\rho) - V^{\pi_T}_{r+\lambda_k c}(\rho) \leq O((1+\lambda_k)\varepsilon)\,,$$

The resulting sample complexity is $N = T\,M|\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$.

Proof. By Lemma 39 and Lemma 40, we have

$$\begin{aligned} V^{\pi_k^*}_\square(\rho) - V^{\pi_T}_\square(\rho) &= V^{\pi_k^*}_\square(\rho) - \bar{V}^T_\square(\rho) + \bar{V}^T_\square(\rho) - V^{\pi_T}_\square(\rho) \\ &\leq \frac{7H^2(1+\lambda_k)}{T} + \sqrt{\frac{6H^3}{TM}} + \sqrt{\frac{6H^6(1+\lambda_k)^2}{TM}\left(\frac{50H^2}{T^2} + \frac{4\iota^2}{M}\right)} \end{aligned}$$

with probability at least $1-\delta$. By letting $M = \frac{6H\iota^2}{\varepsilon}$, $T = \frac{82H^2}{\varepsilon}$, and $\varepsilon \in (0, 1/H^2]$, we have

$$\begin{aligned}
V_{\square}^{\pi_k^*}(\rho) - V_{\square}^{\pi_T}(\rho) &\leq (1+\lambda_k)\varepsilon/12 + \varepsilon/9\iota + (1+\lambda_k)H^{3/2}\varepsilon\left(\frac{\varepsilon}{9H} + \sqrt{\varepsilon/81H}\right) \\
&= (1+\lambda_k)\varepsilon/12 + \varepsilon/9\iota + (1+\lambda_k)\varepsilon^2\sqrt{H}/9 + (1+\lambda_k)H\varepsilon^{3/2}/9 \\
&\leq (1+\lambda_k)\varepsilon/12 + \varepsilon/9 + (1+\lambda_k)\varepsilon/9 + (1+\lambda_k)\varepsilon/9 \\
&\qquad\qquad (\varepsilon \in (0, 1/H^2] \text{ and } \iota = \log(2|\mathcal{S}||\mathcal{A}|/\delta) \geq 1) \\
&\leq (1+\lambda_k)\varepsilon
\end{aligned}$$

which completes the proof. □

G.1.1 Proof of Lemma 37 and Lemma 38 (Proofs with Hoeffding's Inequality)

Lemma 37. *Let π_k^* be defined as in eq.* (C.1)*, and let $\bar{V}_{\square}^T$ denote the averaged empirical value function in Algorithm 6 when run with λ_k. For any $k \in [K]$, we have*

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - \bar{V}_r^T(\rho) - \lambda_k \bar{V}_c^T(\rho) \leq \frac{3H^2(1+\lambda_k)}{T} + 2H(1+\lambda_k)\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{TM}}$$

with probability at least $1-\delta$.

Proof. Since $(I - \gamma P_{\pi_k^*})V_{\square}^{\pi_k^*} = (\pi_k^* \square)$, we have

$$\begin{aligned}
(I - \gamma P_{\pi_k^*})(V_{\square}^{\pi_k^*} - \bar{V}_{\square}^T) &= (\pi_k^* \square) - (\bar{V}_{\square}^T - \gamma P_{\pi_k^*}\bar{V}_{\square}^T) \\
&= (\pi_k^* \square) + \gamma P_{\pi_k^*}\bar{V}_{\square}^T - \bar{V}_{\square}^T \\
\Longrightarrow V_{\square}^{\pi_k^*} - \bar{V}_{\square}^T &= (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*}\bar{V}_{\square}^T - \bar{V}_{\square}^T) \qquad \text{(G.1)}
\end{aligned}$$

By Lemma 42 and due to the greediness of π_t, for all $t \in [T]$, we have

$$\begin{aligned}
\bar{V}_{\square}^t &= \frac{1}{t}\sum_{i=0}^{t-1}(\pi_t \hat{Q}_{\square}^i) \\
&\geq \frac{1}{t}\sum_{i=0}^{t-1}(\pi_k^* \hat{Q}_{\square}^i). \qquad \text{(G.2)}
\end{aligned}$$

Now, we have

$$\begin{aligned}
V_{\square}^{\pi_k^*} - \bar{V}_{\square}^T &= (I - \gamma P_{\pi_k^*})^{-1}((\pi_k^* \square) + \gamma P_{\pi_k^*}\bar{V}_{\square}^T - \bar{V}_{\square}^T) && \text{(By eq. (G.1))} \\
&\leq (I - \gamma P_{\pi_k^*})^{-1}\left((\pi_k^* \square) + \gamma P_{\pi_k^*}\bar{V}_{\square}^T - \frac{1}{T}\sum_{i=0}^{T-1}(\pi_k^* \hat{Q}_{\square}^i)\right) && \text{(By eq. (G.2))} \\
&= (I - \gamma P_{\pi_k^*})^{-1}\left((\pi_k^* \square) + \gamma P_{\pi_k^*}\bar{V}_{\square}^T - (\pi_k^* \square) - \gamma P_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_{\square}^i) - \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}_{\pi_k^*}^i \hat{V}_{\square}^i - \gamma P_{\pi_k^*}\hat{V}_{\square}^i\right]\right) \\
& && \text{(By Lemma 43)}
\end{aligned}$$

$$
\begin{aligned}
&= (I-\gamma P_{\pi_k^*})^{-1}\left(\gamma P_{\pi_k^*}\bar{V}_\square^T - \gamma P_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) - \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}^i_{\pi_k^*}\hat{V}_\square^i - \gamma P_{\pi_k^*}\hat{V}_\square^i\right]\right)\\
&= (I-\gamma P_{\pi_k^*})^{-1}\left(\gamma P_{\pi_k^*}\bar{V}_\square^T - \gamma P_{\pi_k^*}\frac{1}{T-1}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) - \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}^i_{\pi_k^*}\hat{V}_\square^i - \gamma P_{\pi_k^*}\hat{V}_\square^i\right]\right)\\
&+ (I-\gamma P_{\pi_k^*})^{-1}\left(\frac{1}{T(T-1)}\gamma P_{\pi_k^*}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i)\right).
\end{aligned}
$$

We note that $\frac{1}{T-1}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) = \bar{V}_\square^{T-1}$ by Lemma 42. By defining $\mathcal{H}_{\pi_k^*} := \gamma(I - \gamma P_{\pi_k^*})^{-1}\pi_k^* \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, we obtain

$$
V_\square^{\pi_k^*} - \bar{V}_\square^T \le \underbrace{\mathcal{H}_{\pi_k^*}P(\bar{V}_\square^T - \bar{V}_\square^{T-1})}_{\text{Term (i)}} + \underbrace{\mathcal{H}_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-1}\left[P\hat{V}_\square^i - \hat{P}_i\hat{V}_\square^i\right]}_{\text{Term (ii)}} + \underbrace{\mathcal{H}_{\pi_k^*}\left(\frac{1}{T(T-1)}P\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i)\right)}_{\text{Term (iii)}}. \tag{G.3}
$$

Note that for any vector $Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$,

$$
\begin{aligned}
\|\mathcal{H}_{\pi_k^*}Q\|_\infty &= \|\gamma(I-\gamma P_{\pi_k^*})^{-1}\pi_k^*Q\|_\infty\\
&\le \|\gamma(I-\gamma P_{\pi_k^*})^{-1}\|_1\|\pi_k^*Q\|_\infty && \text{(By Holder's inequality)}\\
&\le H\|\pi_k^*Q\|_\infty && \text{(Since } \|\gamma(I-\gamma P_{\pi_k^*})^{-1}\|_1 \le H)\\
&\le H\|Q\|_\infty. && \text{(By definition of the } \pi \text{ operator)}
\end{aligned}
$$

In order to bound Term (i), using theorem 44, we have

$$
\left\|\mathcal{H}_{\pi_k^*}P(\bar{V}_\square^T - \bar{V}_\square^{T-1})\right\|_\infty \le \frac{2H^2(1+\lambda_k)}{T}.
$$

For bounding Term (ii), letting $t = T$ in Lemma 61 and invoking it twice for r and c, we have

$$
\left\|\mathcal{H}_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-1}\left[P\hat{V}_\square^i - \hat{P}_i\hat{V}_\square^i\right]\right\|_\infty \le 2H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}
$$

with probability at least $1-\delta$.

Finally, we bound Term (iii) by noting that $\|\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i)\|_\infty \le (T-1)H(1+\lambda_k)$ due to Lemma 41. Hence,

$$
\left\|\mathcal{H}_{\pi_k^*}\left(\frac{1}{T(T-1)}P\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i)\right)\right\|_\infty \le \frac{H^2(1+\lambda_k)}{T}.
$$

Note that for any vector V, $V(\rho) \le \|V\|_\infty$. Putting everything together, we have

$$
V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - \bar{V}_r^T(\rho) - \lambda_k\bar{V}_c^T(\rho) \le \frac{3H^2(1+\lambda_k)}{T} + 2H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}
$$

with probability at least $1-\delta$.

□

Lemma 38. *Let π_T be the output policy, and let $\bar{V}_\diamond^T$ denote the averaged empirical value function in Algorithm 6 when run with λ_k. For any $k \in [K]$, we have*

$$\bar{V}_r^T(\rho) + \lambda_k \bar{V}_c^T(\rho) - V_r^{\pi_T}(\rho) - \lambda_k V_c^{\pi_T}(\rho) \leq \frac{2H^2(1+\lambda_k)}{T} + 2H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}$$

with probability at least $1-\delta$.

Proof. The proof follows similar steps as before. Since $(I - \gamma P_{\pi_T})V_\square^{\pi_T} = \pi_T \square$, we have

$$\begin{aligned}
(I - \gamma P_{\pi_T})(\bar{V}_\square^T - V_\square^{\pi_T}) &= (\bar{V}_\square^T - \gamma P_{\pi_T}\bar{V}_\square^T) - \pi_T \square \\
&= \bar{V}_\square^T - (\pi_T \square) + \gamma P_{\pi_T}\bar{V}_\square^T \\
\implies \bar{V}_\square^T - V_\square^{\pi_T} &= (I - \gamma P_{\pi_T})^{-1}(\bar{V}_\square^T - (\pi_T \square) + \gamma P_{\pi_T}\bar{V}_\square^T) \qquad \text{(G.4)}
\end{aligned}$$

Recall that for all $t \in [T]$, we have

$$\bar{V}_\diamond^t = \frac{1}{t}\sum_{i=1}^{t} \hat{V}_\diamond^i = \frac{1}{t}\sum_{i=0}^{t-1}(\pi_t \hat{Q}_\diamond^i). \qquad \text{(G.5)}$$

Now, we have

$$\begin{aligned}
\bar{V}_\square^T - V_\square^{\pi_T} &= (I - \gamma P_{\pi_T})^{-1}(\bar{V}_\square^T - (\pi_T \square) + \gamma P_{\pi_T}\bar{V}_\square^T) && \text{(By eq. (G.4))} \\
&= (I - \gamma P_{\pi_T})^{-1}\left(\frac{1}{T}\sum_{i=0}^{T-1}(\pi_T \hat{Q}_\square^i) - (\pi_T \square) + \gamma P_{\pi_T}\bar{V}_\square^T\right) && \text{(By eq. (G.5))} \\
&= (I - \gamma P_{\pi_T})^{-1}\left(\frac{1}{T}\sum_{i=0}^{T-1}(\pi_T \hat{Q}_\square^i) - (\pi_T \square) - \gamma P_{\pi_T}\bar{V}_\square^T\right) \\
&= (I - \gamma P_{\pi_T})^{-1}\Bigg((\pi_T \square) + \gamma P_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}_{\pi_T}^i \hat{V}_\square^i - \gamma P_{\pi_T}\hat{V}_\square^i\right] \\
&\quad -(\pi_T \square) - \gamma P_{\pi_T}\bar{V}_\square^T\Bigg) && \text{(By Lemma 43)} \\
&= (I - \gamma P_{\pi_T})^{-1}\left(\gamma P_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}_{\pi_T}^i \hat{V}_\square^i - \gamma P_{\pi_T}\hat{V}_\square^i\right] - \gamma P_{\pi_T}\bar{V}_\square^T\right) \\
&\leq (I - \gamma P_{\pi_T})^{-1}\left(\gamma P_{\pi_T}\frac{1}{T-1}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}_{\pi_T}^i \hat{V}_\square^i - \gamma P_{\pi_T}\hat{V}_\square^i\right] - \gamma P_{\pi_T}\bar{V}_\square^T\right).
\end{aligned}$$

We note that $\frac{1}{T-1}\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i) = \bar{V}_\square^{T-1}$. By letting $\mathcal{H}_{\pi_T} = \gamma(I - \gamma P_{\pi_T})^{-1}\pi_T$, we obtain

$$\bar{V}_\square^T - V_\square^{\pi_T} \leq \mathcal{H}_{\pi_T} P(\bar{V}_\square^{T-1} - \bar{V}_\square^T) + \mathcal{H}_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i \hat{V}_\square^i - P\hat{V}_\square^i\right]. \qquad \text{(G.6)}$$

Note that for any vector $Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$,

$$\begin{aligned}\|\mathcal{H}_{\pi_T} Q\|_\infty &= \|\gamma(I-\gamma P_{\pi_T})^{-1}\pi_T Q\|_\infty \\ &\le \|\gamma(I-\gamma P_{\pi_T})^{-1}\|_1 \|\pi_T Q\|_\infty && \text{(By Holder's inequality)} \\ &\le H\|\pi_T Q\|_\infty && \text{(Since } \|\gamma(I-\gamma P_{\pi_T})^{-1}\|_1 \le H) \\ &\le H\|Q\|_\infty. && \text{(By definition of the } \pi \text{ operator)}\end{aligned}$$

Thus, letting $t = T$ in Lemma 61, we have

$$\left\|\mathcal{H}_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i\hat{V}_\square^i - P\hat{V}_\square^i\right]\right\|_\infty \le 2H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}$$

with probability at least $1-\delta$. By Lemma 44,

$$\left\|\mathcal{H}_{\pi_T}P(\bar{V}_\square^{T-1} - \bar{V}_\square^T)\right\|_\infty \le \frac{2H^2(1+\lambda_k)}{T}.$$

Note that for any vector V, $V(\rho) \le \|V\|_\infty$. Putting everything together, we have

$$\bar{V}_r^T(\rho) + \lambda_k\bar{V}_c^T(\rho) - V_r^{\pi_T}(\rho) - \lambda_k V_c^{\pi_T}(\rho) \le \frac{2H^2(1+\lambda_k)}{T} + 2H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}$$

with probability at least $1-\delta$.

□

G.1.2 Proof of Lemma 39 and Lemma 40 (Proofs with Bernstein's Inequality)

Lemma 39. *Let π_k^* be defined as in eq.* (C.1)*, and let $\bar{V}_\diamond^T$ denote the averaged empirical value function in Algorithm 6 when run with λ_k. For any $k \in [K]$ and $T \ge 2\log(T)/\gamma$, we have*

$$V_r^{\pi_k^*}(\rho) + \lambda_k V_c^{\pi_k^*}(\rho) - \bar{V}_r^T(\rho) - \lambda_k\bar{V}_c^T(\rho) \le \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)} + \sqrt{\frac{3H^3}{TM}} + \frac{4H^2(1+\lambda_k)}{T}.$$

with probability at least $1-\delta$.

Proof. From eq. (G.3), we have

$$V_\square^{\pi_k^*} - \bar{V}_\square^T \le \underbrace{\mathcal{H}_{\pi_k^*}P(\bar{V}_\square^T - \bar{V}_\square^{T-1})}_{\text{Term (i)}} + \underbrace{\mathcal{H}_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-1}\left[P\hat{V}_\square^i - \hat{P}_i\hat{V}_\square^i\right]}_{\text{Term (ii)}} + \underbrace{\mathcal{H}_{\pi_k^*}\left(\frac{1}{T(T-1)}P\sum_{i=0}^{T-2}(\pi_{T-1}\hat{Q}_\square^i)\right)}_{\text{Term (iii)}}$$

We bound Term (i) and Term (iii) the same way as before. Thus, we only have Term (ii) remains. By Lemma 50, we know

$$\frac{1}{t}\sum_{i=1}^{t}\left[\hat{P}_i\hat{V}^i_\square - P\hat{V}^i_\square\right](s,a) \le \frac{H(1+\lambda_k)\iota}{tM} + \sqrt{Z}$$

where

$$Z := \frac{3H^2(1+\lambda_k)^2}{tM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right) + \frac{3\mathrm{Var}(V_\square^{\pi_k^*}(s,a))}{tM}.$$

Therefore,

$$\begin{aligned}
\mathcal{H}_{\pi_k^*}\frac{1}{T}\sum_{i=0}^{T-1}\left[P\hat{V}^i_\square - \hat{P}_i\hat{V}^i_\square\right] &\le \mathcal{H}_{\pi_k^*}\sqrt{\frac{3H^2(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)}\mathbf{1} + \mathcal{H}_{\pi_k^*}\frac{H(1+\lambda_k)\iota}{TM}\mathbf{1} + \mathcal{H}_{\pi_k^*}\sqrt{\frac{3}{TM}}\sigma(V_\square^{\pi_k^*}) \\
&\le \mathcal{H}_{\pi_k^*}\sqrt{\frac{3H^2(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)}\mathbf{1} + \mathcal{H}_{\pi_k^*}\frac{H(1+\lambda_k)\iota}{TM}\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1} \\
&\qquad \text{(By Lemma 67)} \\
&\le \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)}\mathbf{1} + \frac{H^2(1+\lambda_k)\iota}{TM}\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1}.
\end{aligned}$$

Lastly, combining the upper bounds for Term (i) and Term (iii), we have

$$\begin{aligned}
V_\square^{\pi_k^*} - \bar{V}_\square^T &\le \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)}\mathbf{1} + \frac{H^2(1+\lambda_k)\iota}{TM}\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1} + \frac{3H^2(1+\lambda_k)}{T}\mathbf{1} \\
&\le \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{16H^2\iota^2}{M}\right)}\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1} + \frac{4H^2(1+\lambda_k)}{T}\mathbf{1}.
\end{aligned}$$

□

Lemma 40. *Let π_T be the output policy, and let $\bar{V}_\diamond^T$ denote the averaged empirical value function in Algorithm 6 when run with λ_k. For any $k \in [K]$ and $T \ge 2\log(T)/\gamma$, we have*

$$\bar{V}_r^T(\rho) + \lambda_k\bar{V}_c^T(\rho) - V_r^{\pi_T}(\rho) - \lambda_k V_c^{\pi_T}(\rho) \le \frac{3H^2(1+\lambda_k)}{T} + \sqrt{\frac{3H^3}{TM}} + \sqrt{\frac{3H^6(1+\lambda_k)^2}{TM}\left(\frac{50}{T^2}+\frac{4\iota^2}{M}\right)}$$

with probability at least $1-\delta$.

Proof. Similarly as before, we have

$$\begin{aligned}
\bar{V}_\square^T - V_\square^{\pi_T} &\le \mathcal{H}_{\pi_T}P(\bar{V}_\square^{T-1} - \bar{V}_\square^T) + \mathcal{H}_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i\hat{V}^i_\square - P\hat{V}^i_\square\right] && \text{(By eq. (G.6))} \\
&\le \frac{2H^2(1+\lambda_k)}{T}\mathbf{1} + \mathcal{H}_{\pi_T}\frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i\hat{V}^i_\square - P\hat{V}^i_\square\right] && \text{(By Lemma 44)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2H^2(1+\lambda_k)}{T}\mathbf{1} + \mathcal{H}_{\pi_T}\sqrt{\frac{3H^2(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{4H^2\iota^2}{M}\right)}\mathbf{1} \\
&+ \mathcal{H}_{\pi_T}\frac{H(1+\lambda_k)\iota}{TM}\mathbf{1} + \mathcal{H}_{\pi_T}\sqrt{\frac{3}{TM}}\sigma(V_\square^{\pi_k^*}) && \text{(By Lemma 50)} \\
&\leq \frac{2H^2(1+\lambda_k)}{T}\mathbf{1} + \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{4H^2\iota^2}{M}\right)}\mathbf{1} \\
&+ \frac{H^2(1+\lambda_k)\iota}{TM}\mathbf{1} + \mathcal{H}_{\pi_T}\sqrt{\frac{3}{TM}}\sigma(V_\square^{\pi_k^*}) \\
&\leq \frac{3H^2(1+\lambda_k)}{T}\mathbf{1} + \sqrt{\frac{3H^4(1+\lambda_k)^2}{TM}\left(\frac{4}{T^2}+\frac{4H^2\iota^2}{M}\right)}\mathbf{1} + \mathcal{H}_{\pi_T}\sqrt{\frac{3}{TM}}\sigma(V_\square^{\pi_k^*}).
\end{aligned}
$$

Now, it remains to bound the last term. We first observe that

$$
\begin{aligned}
\sigma(V_\square^{\pi_k^*}) &\leq \left(V_\square^{\pi_k^*} - V_\square^{\pi_T}\right) + \sigma\left(V_\square^{\pi_T}\right) && \text{(By Lemma 66)} \\
&\leq |V_\square^{\pi_k^*} - V_\square^{\pi_T}| + \sigma\left(V_\square^{\pi_T}\right) && \text{(By Lemma 65)} \\
&\leq \frac{5H^2(1+\lambda_k)}{T}\mathbf{1} + 4H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}\mathbf{1} + \sigma\left(V_\square^{\pi_T}\right) \\
& && \text{(By combining Lemma 37 and Lemma 38)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathcal{H}_{\pi_T}\sqrt{\frac{3}{TM}}\sigma(V_\square^{\pi_k^*}) &\leq \mathcal{H}_{\pi_T}\sqrt{\frac{3}{TM}}\left(\frac{5H^2(1+\lambda_k)}{T} + 4H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}\right)\mathbf{1} + \sqrt{\frac{3}{TM}}\mathcal{H}_{\pi_T}\sigma\left(V_\square^{\pi_T}\right) \\
&\leq \sqrt{\frac{3H^2}{TM}}\left(\frac{5H^2(1+\lambda_k)}{T} + 4H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}\right)\mathbf{1} + \sqrt{\frac{3}{TM}}\mathcal{H}_{\pi_T}\sigma\left(V_\square^{\pi_T}\right) \\
&\leq \sqrt{\frac{3H^2}{TM}}\left(\frac{5H^2(1+\lambda_k)}{T} + 4H^2(1+\lambda_k)\sqrt{\frac{\iota}{TM}}\right)\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1} \\
& && \text{(By Lemma 67)} \\
&= \sqrt{\frac{3H^6(1+\lambda_k)^2}{TM}}\left(\frac{5}{T} + \sqrt{\frac{16\iota}{TM}}\right)\mathbf{1} + \sqrt{\frac{3H^3}{TM}}\mathbf{1}.
\end{aligned}
$$

By combining the above results and consolidating like terms, we conclude the proof. □

G.1.3 Auxiliary Lemmas

Lemma 41. *Denote $\square = r + \lambda_k c$. For any $k \in [K]$ and any $t \in [T]$, $\hat{Q}_\square^t(s,a)$ and $\hat{V}_\square^t(s)$ are bounded by $(1+\lambda_k)H$.*

Proof. We prove it by induction. By initialization, $\hat{Q}_\square^1(s,a) = r(s,a) + \lambda_k c(s,a) \leq 1 + \lambda_k$ and $\hat{V}_\square^1(s) = \hat{Q}_\square^1(s, \pi_1(\cdot|s)) \leq 1 + \lambda_k \leq (1+\lambda_k)H$. Now, suppose $\hat{V}_\square^{t-1}(s)$ is bounded by

$(1+\lambda_k)H$ for some $t \geq 1$. We have

$$\begin{aligned}
\hat{V}^t_\square &= \sum_{i=0}^{t}(\pi_t \hat{Q}^i_\square) - \sum_{i=0}^{t-1}(\pi_{t-1}\hat{Q}^i_\square) && \text{(From the last line in Algorithm 6)} \\
&\leq \sum_{i=0}^{t}(\pi_t \hat{Q}^i_\square) - \sum_{i=0}^{t-1}(\pi_t\hat{Q}^i_\square) && \text{(By the greediness of } \pi_{t-1}\text{)} \\
&= (\pi_t \hat{Q}^{t-1}_\square) \\
&\leq (\pi_t \square) + \gamma(\hat{P}^{t-1}_{\pi_t}\hat{V}^{t-1}_\square) && \text{(From the second last line in Algorithm 6)} \\
&\leq (1+\lambda_k+\gamma(1+\lambda_k)H)\mathbf{1} && \text{(Induction hypothesis)} \\
&= (1+\lambda_k)H\mathbf{1}. && (1+\tfrac{\gamma}{1-\gamma} = \tfrac{1}{1-\gamma})
\end{aligned}$$

Therefore, $\hat{V}^t_\square(s)$ is bounded by $(1+\lambda_k)H$. As a consequence, $\hat{Q}^t_\square(s,a)$ is also bounded by $(1+\lambda_k)H$. □

Lemma 42. *For any $k \in [K]$ and $t \in [T]$, we have*

$$\bar{V}^t_\diamond := \frac{1}{t}\sum_{i=1}^{t}\hat{V}^i_\diamond = \frac{1}{t}\sum_{i=0}^{t-1}(\pi_t\hat{Q}^i_\diamond).$$

Proof.

$$\begin{aligned}
\bar{V}^t_\diamond &:= \frac{1}{t}\sum_{i=1}^{t}\hat{V}^i_\diamond \\
&= \frac{1}{t}\sum_{i=0}^{t-1}\left(\sum_{j=0}^{i}(\pi_{i+1}\hat{Q}^j_\diamond) - \sum_{j=0}^{i-1}(\pi_i\hat{Q}^j_\diamond)\right) && \text{(From the last line in Algorithm 6)} \\
&= \frac{1}{t}\sum_{i=0}^{t-1}(\pi_t\hat{Q}^i_\diamond). && \text{(Due to telescoping sum)}
\end{aligned}$$

□

Lemma 43. *For any $k \in [K]$ and $t \in [T]$, we have*

$$\sum_{i=0}^{t-1}\hat{Q}^i_\diamond = t\diamond + \gamma P\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\diamond) + \sum_{i=0}^{t-1}\left[\gamma\hat{P}_i\hat{V}^i_\diamond - \gamma P\hat{V}^i_\diamond\right],$$

and

$$\sum_{i=0}^{t-1}\hat{Q}^i_\square = t(r+\lambda_k c) + \gamma P\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) + \sum_{i=0}^{t-1}\left[\gamma\hat{P}^i_{\pi_t}\hat{V}^i_\square - \gamma P_{\pi_t}\hat{V}^i_\square\right].$$

Proof. We prove the first equality. The second equality follows by linearity.

$$\sum_{i=0}^{t-1}\hat{Q}^i_\diamond = \sum_{i=0}^{t-1}\left[\diamond + \gamma\hat{P}_i\hat{V}^i_\diamond\right] \qquad \text{(From the second last line in Algorithm 6)}$$

$$
\begin{aligned}
&= \sum_{i=0}^{t-1} \left[\diamond + \gamma P\hat{V}_\diamond^i - \gamma P\hat{V}_\diamond^i + \gamma \hat{P}_i \hat{V}_\diamond^i\right] \\
&= t \diamond + \gamma P \sum_{i=0}^{t-1} \hat{V}_\diamond^i + \sum_{i=0}^{t-1} \left[\gamma \hat{P}_i \hat{V}_\diamond^i - \gamma P \hat{V}_\diamond^i\right] \\
&= t \diamond + \gamma P \sum_{i=0}^{t-2} (\pi_{t-1} \hat{Q}_\diamond^i) + \sum_{i=0}^{t-1} \left[\gamma \hat{P}_i \hat{V}_\diamond^i - \gamma P \hat{V}_\diamond^i\right]. && \text{(From Lemma 42)}
\end{aligned}
$$

□

Lemma 44. *For any $k \in [K]$,*

$$\|\bar{V}_\square^T - \bar{V}_\square^{T-1}\|_\infty \leq \frac{2H(1+\lambda_k)}{T}.$$

Proof. We present the proof for the case of $\bar{V}_\square^T - \bar{V}_\square^{T-1}$. The proof for the another case is similar. By the definition of $\bar{V}_\square^t$ and due to the greediness of π_{T-1}, we have

$$
\begin{aligned}
\bar{V}_\square^T - \bar{V}_\square^{T-1} &= \frac{1}{T} \sum_{i=0}^{T-1} (\pi_T \hat{Q}_\square^i) - \frac{1}{T-1} \sum_{i=0}^{T-2} (\pi_{T-1} \hat{Q}_\square^i) \\
&\leq \frac{1}{T} \sum_{i=0}^{T-1} (\pi_T \hat{Q}_\square^i) - \frac{1}{T-1} \sum_{i=0}^{T-2} (\pi_T \hat{Q}_\square^i) \\
&\leq \frac{1}{T} \sum_{i=0}^{T-1} (\pi_T \hat{Q}_\square^i) - \frac{1}{T} \sum_{i=0}^{T-2} (\pi_T \hat{Q}_\square^i) \\
&\leq \frac{1}{T} (\pi_T \hat{Q}_\square^{T-1}) \\
&\leq \frac{2H(1+\lambda_k)}{T} \mathbf{1}. && \text{(By Lemma 41)}
\end{aligned}
$$

□

G.2 Proof of Lemma 34 (Optimality Guarantees for Algorithm 7 - Tabular CMDP)

Lemma 45. *For a fixed $\varepsilon \in (0, H]$, $\delta \in (0,1)$, algorithm 7 with $M = \tilde{O}\left(\frac{H}{\varepsilon}\right)$ and $T = O\left(\frac{H^2}{\varepsilon}\right)$, the output $\bar{\mathcal{V}}_\diamond^T$ satisfies the following condition with probability $1-\delta$,*

$$|\bar{\mathcal{V}}_\diamond^T(\rho) - V_\diamond^\pi(\rho)| \leq O(\varepsilon),$$

The resulting sample complexity is $N = T\, M |\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$.

Proof. By Lemma 48, we have

$$\left\|\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi\right\|_\infty \leq \tilde{O}\left(\frac{H^2}{T} + \frac{H}{tM} + \sqrt{\frac{H^4}{tM^2}} + \sqrt{\frac{H^3}{tM}}\right)$$

with probability at least $1-\delta$. By letting $M = \tilde{O}\left(\frac{H}{\varepsilon}\right)$, $T = \tilde{O}\left(\frac{H^2}{\varepsilon}\right)$, and $\varepsilon \in (0, H]$, we have

$$\left\|\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi\right\|_\infty \leq O\left(\varepsilon + \frac{\varepsilon^2}{H^2} + \varepsilon + \varepsilon\right) \leq O\left(\varepsilon\right)$$

with total sample complexity $N = TM|\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$ and probability at least $1-\delta$. □

G.2.1 Auxiliary Lemmas

Since Algorithm 7 is equivalent to running Algorithm 6 with a fixed policy, the following lemma follows directly from Lemma 41.

Lemma 46. *For any $t \in [T]$, $\hat{\mathcal{Q}}_\diamond^t(s,a)$ and $\hat{\mathcal{V}}_\diamond^t(s)$ are bounded by H.*

Lemma 47. $\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi \leq \mathcal{H}_\pi P(\bar{\mathcal{V}}_\diamond^{T-1} - \bar{\mathcal{V}}_\diamond^T) + \mathcal{H}_\pi \frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i\hat{\mathcal{V}}_\diamond^i - P\hat{\mathcal{V}}_\diamond^i\right]$.

Proof. First, we notice that

$$\begin{aligned}
\sum_{i=0}^{t-1}\hat{\mathcal{Q}}_\diamond^i &= \sum_{i=0}^{t-1}\left[\diamond + \gamma\hat{P}_i\hat{\mathcal{V}}_\diamond^i\right] \\
&= \sum_{i=0}^{t-1}\left[\diamond + \gamma P\hat{\mathcal{V}}_\diamond^i - \gamma P\hat{\mathcal{V}}_\diamond^i + \gamma\hat{P}_i\hat{\mathcal{V}}_\diamond^i\right] \\
&= t\diamond + \gamma P\sum_{i=0}^{t-1}\hat{\mathcal{V}}_\diamond^i + \sum_{i=0}^{t-1}\left[\gamma\hat{P}_i\hat{\mathcal{V}}_\diamond^i - \gamma P\hat{\mathcal{V}}_\diamond^i\right] \\
&= t\diamond + \gamma P\sum_{i=0}^{t-2}(\pi\hat{\mathcal{Q}}_\diamond^i) + \sum_{i=0}^{t-1}\left[\gamma\hat{P}_i\hat{\mathcal{V}}_\diamond^i - \gamma P\hat{\mathcal{V}}_\diamond^i\right].
\end{aligned} \tag{G.7}$$

It is different from Lemma 43 because the policy is now fixed at each iteration. The rest of the proof follows the same set of steps as in the proof of Lemma 37. Denote $\diamond = r$ or c. Since $(I - \gamma P_\pi)V_\diamond^\pi = \pi\diamond$, we have

$$\begin{aligned}
(I-\gamma P_\pi)(\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi) &= (\bar{\mathcal{V}}_\diamond^T - \gamma P_\pi\bar{\mathcal{V}}_\diamond^T) - \pi\diamond \\
&= \bar{\mathcal{V}}_\diamond^T - (\pi\diamond) + \gamma P_\pi\bar{\mathcal{V}}_\diamond^T \\
\implies \bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi &= (I-\gamma P_\pi)^{-1}(\bar{\mathcal{V}}_\diamond^T - (\pi\diamond) + \gamma P_\pi\bar{\mathcal{V}}_\diamond^T)
\end{aligned} \tag{G.8}$$

Recall that $\bar{\mathcal{V}}_\diamond^t = \frac{1}{t}\sum_{i=1}^t\hat{\mathcal{V}}_\diamond^i = \pi\frac{1}{t}\sum_{i=0}^{t-1}\hat{\mathcal{Q}}_\diamond^i$ for all $t \in [T]$. Now, we have

$$\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi = (I-\gamma P_\pi)^{-1}(\bar{\mathcal{V}}_\diamond^T - (\pi\diamond) + \gamma P_\pi\bar{\mathcal{V}}_\diamond^T) \qquad \text{(By eq. (G.8))}$$

$$= (I-\gamma P_\pi)^{-1}\left(\frac{1}{T}\sum_{i=0}^{T-1}(\pi\hat{\mathcal{Q}}^i_\diamond) - (\pi\diamond) + \gamma P_\pi \bar{\mathcal{V}}^T_\diamond\right)$$

$$= (I-\gamma P_\pi)^{-1}\left((\pi\diamond) + \gamma P_\pi \frac{1}{T}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}^i_\diamond) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}^i_\pi \hat{\mathcal{V}}^i_\diamond - \gamma P_\pi \hat{\mathcal{V}}^i_\diamond\right] - (\pi\diamond) - \gamma P_\pi \bar{\mathcal{V}}^T_\diamond\right)$$

(By eq. (G.7))

$$= (I-\gamma P_\pi)^{-1}\left(\gamma P_\pi \frac{1}{T}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}^i_\diamond) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}^i_\pi \hat{\mathcal{V}}^i_\diamond - \gamma P_\pi \hat{\mathcal{V}}^i_\diamond\right] - \gamma P_\pi \bar{\mathcal{V}}^T_\diamond\right)$$

$$\leq (I-\gamma P_\pi)^{-1}\left(\gamma P_\pi \frac{1}{T-1}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}^i_\diamond) + \frac{1}{T}\sum_{i=0}^{T-1}\left[\gamma \hat{P}^i_\pi \hat{\mathcal{V}}^i_\diamond - \gamma P_\pi \hat{\mathcal{V}}^i_\diamond\right] - \gamma P_\pi \bar{\mathcal{V}}^T_\diamond\right).$$

We note that $\frac{1}{T-1}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}^i_\diamond) = \bar{\mathcal{V}}^{T-1}_\diamond$, as it is an equivalent result of Lemma 42 with a fixed policy. By letting $\mathcal{H}_\pi = \gamma(I-\gamma P_\pi)^{-1}\pi$, we obtain

$$\bar{\mathcal{V}}^T_\diamond - V^\pi_\diamond \leq \mathcal{H}_\pi P(\bar{\mathcal{V}}^{T-1}_\diamond - \bar{\mathcal{V}}^T_\diamond) + \mathcal{H}_\pi \frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i \hat{\mathcal{V}}^i_\diamond - P\hat{\mathcal{V}}^i_\diamond\right]. \tag{G.9}$$

□

Lemma 48. *We have*

$$\left\|\bar{\mathcal{V}}^T_\diamond - V^\pi_\diamond\right\|_\infty \leq \tilde{O}\left(\frac{H^2}{T} + \frac{H}{tM} + \sqrt{\frac{H^4}{tM^2}} + \sqrt{\frac{H^3}{tM}}\right)$$

with probability at least $1-\delta$.

Proof. By Lemma 47, we have

$$\begin{aligned}\bar{\mathcal{V}}^T_\diamond - V^\pi_\diamond &\leq \mathcal{H}_\pi P(\bar{\mathcal{V}}^{T-1}_\diamond - \bar{\mathcal{V}}^T_\diamond) + \mathcal{H}_\pi \frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i \hat{\mathcal{V}}^i_\diamond - P\hat{\mathcal{V}}^i_\diamond\right] \\ &\leq \tilde{O}\left(\frac{H^2}{T}\right)\mathbf{1} + \mathcal{H}_\pi \frac{1}{T}\sum_{i=0}^{T-1}\left[\hat{P}_i \hat{\mathcal{V}}^i_\diamond - P\hat{\mathcal{V}}^i_\diamond\right]. && \text{(By Lemma 49)}\end{aligned}$$

Thus, it remains to bound the second term. By Lemma 51 we have

$$\frac{1}{t}\sum_{i=0}^{t-1}\left[\hat{P}_i \hat{\mathcal{V}}^i_\diamond - P\hat{\mathcal{V}}^i_\diamond\right](s,a) \leq \frac{H\iota}{tM} + \sqrt{Z}$$

where

$$Z := \frac{1}{tM}\left(\frac{H^4}{M} + \mathrm{Var}(V^\pi_\diamond(s,a))\right)$$

with probability at least $1-\delta$. Therefore,

$$\begin{aligned}
\bar{\mathcal{V}}_\diamond^T - V_\diamond^\pi &\le \tilde{O}\left(\frac{H^2}{T}\right)\mathbf{1} + \frac{H\iota}{tM}\mathbf{1} + \sqrt{\frac{H^4}{tM^2}}\mathbf{1} + \sqrt{\frac{1}{tM}}\mathcal{H}_\pi \sigma(V_\diamond^\pi) \\
&\le \tilde{O}\left(\frac{H^2}{T}\right)\mathbf{1} + \frac{H\iota}{tM}\mathbf{1} + \sqrt{\frac{H^4}{tM^2}}\mathbf{1} + \sqrt{\frac{H^3}{tM}}\mathbf{1} && \text{(By Lemma 67)} \\
&\le \tilde{O}\left(\frac{H^2}{T} + \frac{H\iota}{tM} + \sqrt{\frac{H^4}{tM^2}} + \sqrt{\frac{H^3}{tM}}\right)\mathbf{1}
\end{aligned}$$

which completes the proof. □

Lemma 49. $\|\bar{\mathcal{V}}_\diamond^T - \bar{\mathcal{V}}_\diamond^{T-1}\|_\infty \le \frac{H}{T}$.

Proof. Similar to the proof of Lemma 44, we have

$$\begin{aligned}
\bar{\mathcal{V}}_\diamond^T - \bar{\mathcal{V}}_\diamond^{T-1} &= \frac{1}{T}\sum_{i=0}^{T-1}(\pi\hat{\mathcal{Q}}_\diamond^i) - \frac{1}{T-1}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}_\diamond^i) \\
&= \frac{1}{T}\sum_{i=0}^{T-1}(\pi\hat{\mathcal{Q}}_\diamond^i) - \frac{1}{T-1}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}_\diamond^i) \\
&\le \frac{1}{T}\sum_{i=0}^{T-1}(\pi\hat{\mathcal{Q}}_\diamond^i) - \frac{1}{T}\sum_{i=0}^{T-2}(\pi\hat{\mathcal{Q}}_\diamond^i) \\
&\le \frac{1}{T}(\pi\hat{\mathcal{Q}}_\diamond^{T-1}) \\
&\le \frac{H}{T}\mathbf{1}. && \text{(By Lemma 46)}
\end{aligned}$$

□

G.3 Proof of Lemma 50 and Lemma 51 (Concentration Error Bounds with Bernstein's Inequality - Tabular CMDP)

All the proofs presented in this section are adapted from the proofs for Lemmas 5 to 8 in Kozuno et al. [2022], with substantial modifications to suit our setting.

Lemma 50. *For any $t \ge 2\log(t)/\gamma$ and $k \in [K]$, we have*

$$\frac{1}{t}\sum_{i=1}^{t}\left[\hat{P}_i\hat{V}_\square^i - P\hat{V}_\square^i\right](s,a) \le \frac{H(1+\lambda_k)\iota}{tM} + \sqrt{Z}$$

where

$$Z := \frac{3H^2(1+\lambda_k)^2}{tM}\left(\frac{4}{t^2} + \frac{16H^2\iota^2}{M}\right) + \frac{3\mathrm{Var}(V_\square^{\pi_k^*}(s,a))}{tM}$$

with probability at least $1-\delta$.

Proof. We have

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t}\left[\hat{P}_i\hat{V}_\Box^i - P\hat{V}_\Box^i\right](s,a) &= \frac{1}{t}\sum_{i=1}^{t}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}_\Box^i(y_{i,m,s,a}) - (P\hat{V}_\Box^i)(s,a)\right] \\ &= \frac{1}{tM}\sum_{i=1}^{t}\sum_{m=1}^{M}\left[\hat{V}_\Box^i(y_{i,m,s,a}) - (P\hat{V}_\Box^i)(s,a)\right].\end{aligned}$$

The above is a sum of bounded martingale differences with respect to the filtraion $(\mathcal{F})_{i=1,m=1}^{t,M}$. Let $X_{i,m} = \frac{1}{tM}\left(\hat{V}_\Box^i(y_{i,m,s,a}) - (P\hat{V}_\Box^i)(s,a)\right)$. It can be noted that $X_{i,m} \le \frac{H(1+\lambda_k)}{tM}$ (by Lemma 41) and $\mathbb{E}[X_{i,m}] = 0$. Next, we bound Z' as defined in Lemma 64

$$\begin{aligned}Z' &= \sum_{i=1}^{t}\sum_{m=1}^{M}\mathbb{E}\left[X_{i,m}^2\right] \\ &= \sum_{i=1}^{t}\sum_{m=1}^{M}\mathbb{E}\left[\frac{1}{t^2M^2}\left(\hat{V}_\Box^i(y_{i,m,s,a}) - (P\hat{V}_\Box^i)(s,a)\right)^2\right] \\ &= \frac{1}{t^2M^2}\sum_{i=1}^{t}\sum_{m=1}^{M}\mathrm{Var}(\hat{V}_\Box^i(y_{i,m,s,a})) \\ &\le \frac{3}{t^2M^2}\sum_{i=1}^{t}\sum_{m=1}^{M}\left(\frac{4H^2(1+\lambda_k)^2}{t^2} + \frac{16H^4(1+\lambda_k)^2\iota^2}{M} + \mathrm{Var}(V_\Box^{\pi_k^*}(s,a))\right) \\ &\qquad\qquad\text{(By Lemma 55)} \\ &= \frac{3}{tM}\left(\frac{4H^2(1+\lambda_k)^2}{t^2} + \frac{16H^4(1+\lambda_k)^2\iota^2}{M} + \mathrm{Var}(V_\Box^{\pi_k^*}(s,a))\right) \\ &:= Z.\end{aligned}$$

By letting $U = H(1+\lambda_k)$ in Lemma 64, we have

$$\frac{1}{t}\sum_{i=1}^{t}\left[\hat{P}_i\hat{V}_\Box^i - P\hat{V}_\Box^i\right](s,a) \le \frac{H(1+\lambda_k)\iota}{tM} + \sqrt{Z}$$

with probability at least $1-\delta$. □

Lemma 51. *For any* $t \ge 2\log(t)/\gamma$, *we have*

$$\frac{1}{t}\sum_{i=0}^{t-1}\left[\hat{P}_i\hat{\mathcal{V}}_\diamond^i - P\hat{\mathcal{V}}_\diamond^i\right](s,a) \le \frac{H\iota}{tM} + \sqrt{Z}$$

where

$$Z := \frac{1}{tM}\left(\frac{H^4}{M} + \mathrm{Var}(V_\diamond^\pi(s,a))\right)$$

with probability at least $1-\delta$.

Proof. We have

$$\frac{1}{t}\sum_{i=0}^{t-1}\left[\hat{P}_i\hat{\mathcal{V}}_\diamond^i - P\hat{\mathcal{V}}_\diamond^i\right](s,a) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{\mathcal{V}}_\diamond^i(y_{i,m,s,a}) - (P\hat{\mathcal{V}}_\diamond^i)(s,a)\right]$$
$$= \frac{1}{tM}\sum_{i=1}^{t}\sum_{m=1}^{M}\left[\hat{\mathcal{V}}_\diamond^i(y_{i,m,s,a}) - (P\hat{\mathcal{V}}_\diamond^i)(s,a)\right].$$

The above is a sum of bounded martingale differences with respect to the filtraion $(\mathcal{F})_{i=1,m=1}^{t,M}$. Let $X_{i,m} = \frac{1}{tM}\left(\hat{\mathcal{V}}_\diamond^i(y_{i,m,s,a}) - (P\hat{\mathcal{V}}_\diamond^i)(s,a)\right)$. It can be noted that $X_{i,m} \le \frac{H}{tM}$ and $\mathbb{E}[X_{i,m}] = 0$. Next, we bound Z' as defined in Lemma 64

$$\begin{aligned}
Z' &= \sum_{i=1}^{t}\sum_{m=1}^{M}\mathbb{E}\left[X_{i,m}^2\right] \\
&= \sum_{i=1}^{t}\sum_{m=1}^{M}\mathbb{E}\left[\frac{1}{t^2M^2}\left(\hat{\mathcal{V}}_\diamond^i(y_{i,m,s,a}) - (P\hat{\mathcal{V}}_\diamond^i)(s,a)\right)^2\right] \\
&= \frac{1}{t^2M^2}\sum_{i=1}^{t}\sum_{m=1}^{M}\mathrm{Var}(\hat{\mathcal{V}}_\diamond^i(y_{i,m,s,a})) \\
&\le \frac{1}{t^2M^2}\sum_{i=1}^{t}\sum_{m=1}^{M}\left(\frac{H^4}{M} + \mathrm{Var}(V_\diamond^\pi(s,a))\right) && \text{(By Lemma 58)} \\
&= \frac{1}{tM}\left(\frac{H^4}{M} + \mathrm{Var}(V_\diamond^\pi(s,a))\right) \\
&:= Z
\end{aligned}$$

with probability at least $1-\delta$. Taking the union bound over $(s,a,i) \in \mathcal{S}\times\mathcal{A}\times[t]$ and by Lemma 64, we have

$$\frac{1}{t}\sum_{i=0}^{t-1}\left[\hat{P}_i\hat{\mathcal{V}}_\diamond^i - P\hat{\mathcal{V}}_\diamond^i\right](s,a) \le \frac{H\iota}{tM} + \sqrt{Z}$$

with probability at least $1-\delta$. □

G.3.1 Auxiliary Lemmas for Lemma 50

Lemma 52. *For any $t \in [T]$ and $k \in [K]$,*

$$\mathbf{0} \le V_\square^{\pi_k^*} - V_\square^{\pi_t'} \le \sum_{i=1}^{t}\left(\prod_{j=1}^{t-i}[\gamma P_{\pi_{t-j}}]\pi_i - (\gamma P_{\pi_k^*})^{t-i}\pi_k^*\right)\frac{1}{i}\sum_{j=0}^{i-1}\left[\gamma\hat{P}_j\hat{V}_\square^j - \gamma P\hat{V}_\square^j\right] + \frac{H(1+\lambda_k)}{t(t-1)}\mathbf{1}.$$

Proof. The first inequality is due to the definition of π^*_k. For the second inequality, since we have $V_\square^{\pi^*_k} - V_\square^{\pi'_t} = \underbrace{V_\square^{\pi^*_k} - \frac{1}{t}\sum_{i=1}^{t}(\pi_t \hat{Q}^i_\square)}_{\text{Term (i)}} + \underbrace{\frac{1}{t}\sum_{i=1}^{t}(\pi_t \hat{Q}^i_\square) - V_\square^{\pi'_t}}_{\text{Term (ii)}}$, we first bound term (i)

$$
\begin{aligned}
V_\square^{\pi^*_k} - \pi_t \frac{1}{t}\sum_{i=1}^{t}\hat{Q}^i_\square &\leq (\pi^*_k Q_\square^{\pi^*_k}) - \frac{1}{t}\sum_{i=1}^{t}(\pi^*_k \hat{Q}^i_\square) && \text{(By the greediness of } \pi_t\text{)} \\
&= (\pi^*_k \square) + \gamma P_{\pi^*_k} V_\square^{\pi^*_k} - \frac{1}{t}\sum_{i=1}^{t}(\pi^*_k \hat{Q}^i_\square) \\
&= (\pi^*_k \square) + \gamma P_{\pi^*_k} V_\square^{\pi^*_k} - (\pi^*_k \square) - \gamma P_{\pi^*_k}\frac{1}{t}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) - \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi^*_k}\hat{V}^i_\square - \gamma P_{\pi^*_k}\hat{V}^i_\square\right] \\
& && \text{(By Lemma 43)} \\
&= \gamma P_{\pi^*_k}\left(V_\square^{\pi^*_k} - \frac{1}{t}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square)\right) - \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi^*_k}\hat{V}^i_\square - \gamma P_{\pi^*_k}\hat{V}^i_\square\right] \\
&= \gamma P_{\pi^*_k}\left(V_\square^{\pi^*_k} - \frac{1}{t}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) + \frac{1}{t-1}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) - \frac{1}{t-1}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square)\right) \\
&\quad - \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi^*_k}\hat{V}^i_\square - \gamma P_{\pi^*_k}\hat{V}^i_\square\right] \\
&\leq \gamma P_{\pi^*_k}\left(V_\square^{\pi^*_k} - \frac{1}{t-1}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square)\right) - \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi^*_k}\hat{V}^i_\square - \gamma P_{\pi^*_k}\hat{V}^i_\square\right] + \frac{H(1+\lambda_k)}{t(t-1)}\mathbf{1} \\
& && (\|\|\gamma \pi^*_k P\|_1 \|\pi_{t-1}\hat{Q}^i_\square\|_\infty \leq H(1+\lambda_k) \text{ for all } k \text{ and } i) \\
&\leq -\sum_{i=1}^{t}(\gamma P_{\pi^*_k})^{t-i}\frac{1}{i}\sum_{j=0}^{i-1}\left[\gamma \hat{P}^j_{\pi^*_k}\hat{V}^j_\square - \gamma P_{\pi^*_k}\hat{V}^j_\square\right] + \frac{H(1+\lambda_k)}{t(t-1)}\mathbf{1}. \\
& && \text{(By induction (Lemma 56) and } \hat{V}^0_\square = \mathbf{0}\text{)}
\end{aligned}
$$

Next, we bound term (ii). We define Q^π the Q-value function for a policy π being its unique fixed point.

$$
\begin{aligned}
\frac{1}{t}\sum_{i=1}^{t}(\pi_t \hat{Q}^i_\square) - V_\square^{\pi'_t} &\leq \frac{1}{t}\sum_{i=1}^{t}(\pi_t \hat{Q}^i_\square) - \pi_t \prod_{i=1}^{t-1}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} && \text{(From the definition of } \pi_t\text{)} \\
&= \frac{1}{t}\sum_{i=1}^{t}(\pi_t \hat{Q}^i_\square) - (\pi_t \square) - \gamma P_{\pi_t}\pi_{t-1}\prod_{i=1}^{t-2}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} \\
&= (\pi_t \square) + \gamma P_{\pi_t}\frac{1}{t}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) + \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi_t}\hat{V}^i_\square - \gamma P_{\pi_t}\hat{V}^i_\square\right] \\
&\quad - (\pi_t \square) - \gamma P_{\pi_t}\pi_{t-1}\prod_{i=1}^{t-2}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} && \text{(By Lemma 43)} \\
&= \gamma P_{\pi_t}\left(\frac{1}{t}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}^i_\square) - V_\square^{\pi'_{t-1}}\right) + \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}^i_{\pi_t}\hat{V}^i_\square - \gamma P_{\pi_t}\hat{V}^i_\square\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \gamma P_{\pi_t}\left(\frac{1}{t-1}\sum_{i=0}^{t-2}(\pi_{t-1}\hat{Q}_\square^i) - V_\square^{\pi'_{t-1}}\right) + \frac{1}{t}\sum_{i=0}^{t-1}\left[\gamma \hat{P}_{\pi_t}^i \hat{V}_\square^i - \gamma P_{\pi_t}\hat{V}_\square^i\right] \\
&\leq \sum_{i=1}^{t}\prod_{j=1}^{t-i}[\gamma P_{\pi_{t-j}}]\frac{1}{i}\sum_{j=0}^{i-1}\left[\gamma \hat{P}_{\pi_i}^j \hat{V}_\square^j - \gamma P_{\pi_i}\hat{V}_\square^j\right].
\end{aligned}
$$

(By induction (Lemma 56) and $\hat{V}_\square^0 = \mathbf{0}$)

Thus, we obtain the second inequality. □

Lemma 53. *For any $t \in [T-1]$ and $k \in [K]$,*

$$
\hat{V}_\square^{t+1} \leq V_\square^{\pi'_{t+1}} + \gamma^{t+1} tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}\gamma(\hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_\square^{t-i} - P_{\pi_{t-i}}\hat{V}_\square^{t-i})
$$

and

$$
\hat{V}_\square^{t+1} \geq V_\square^{\pi'_t} - \gamma^{t+1} tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}\gamma(\hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_\square^{t-i} - P_{\pi_{t-i}}\hat{V}_\square^{t-i}).
$$

Proof. We first note that

$$
\begin{aligned}
\hat{V}_\square^{t+1} &= \sum_{i=0}^{t}(\pi_{t+1}\hat{Q}_\square^i) - \sum_{i=0}^{t-1}(\pi_t\hat{Q}_\square^i) && \text{(From the last line in Algorithm 6)} \\
&\leq \sum_{i=0}^{t}(\pi_{t+1}\hat{Q}_\square^i) - \sum_{i=0}^{t-1}(\pi_{t+1}\hat{Q}_\square^i) && \text{(By the greediness of } \pi_t\text{)} \\
&= (\pi_{t+1}\hat{Q}_\square^t) \\
&= (\pi_{t+1}\square) + \gamma \hat{P}_{\pi_{t+1}}^t \hat{V}_\square^t \\
&= (\pi_{t+1}\square) + \gamma P_{\pi_{t+1}}\hat{V}_\square^t + \gamma(\hat{P}_{\pi_{t+1}}^t \hat{V}_\square^t - P_{\pi_{t+1}}\hat{V}_\square^t) \\
&\leq \sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}(\square + \gamma(\hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_\square^{t-i} - P_{\pi_{t-i}}\hat{V}_\square^{t-i})). && \text{(By induction on } t\text{)}
\end{aligned}
$$

Let $\mathcal{T}^\pi$ denote the Bellman operator with policy π, we have

$$
\begin{aligned}
\pi_{t+1}\prod_{i=0}^{t}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} &= \sum_{i=1}^{t}\gamma^i \prod_{j=t-i-1}^{t} P_{\pi_{t-j}}(\pi_i\square) + \gamma^{t+1}\prod_{j=0}^{t}P_{\pi_{t-j}}(\pi_0 Q_\square^{\pi_0}) \\
\Longrightarrow \sum_{i=1}^{t}\gamma^i \prod_{j=t-i-1}^{t} P_{\pi_{t-j}}(\pi_i\square) &\leq \pi_{t+1}\prod_{i=0}^{t}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} + \gamma^{t+1}tH(1+\lambda_k)\mathbf{1}.
\end{aligned}
$$

(Since $\prod_{j=0}^{t} P\pi_j Q_\square^{\pi_0} \leq tH(1+\lambda_k)\mathbf{1}$)

Combining all above, we obtain

$$
\hat{V}_\square^{t+1} \leq \pi_{t+1}\prod_{i=0}^{t}\mathcal{T}^{\pi_i}Q_\square^{\pi_0} + \gamma^{t+1}tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}\gamma(\hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_\square^{t-i} - P_{\pi_{t-i}}\hat{V}_\square^{t-i})
$$

Denoting $\pi'_{k,t}$ a non-stationary policy that follows $\pi_{t+1}, \pi_t, \pi_{t-1}, \dots$ sequentially, we simplify the above inequality as

$$\hat{V}_{\square}^{t+1} \leq V_{\square}^{\pi'_{t+1}} + \gamma^{t+1} tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma(\hat{P}_{\pi_{t-i}}^{t-i} \hat{V}_{\square}^{t-i} - P_{\pi_{t-i}} \hat{V}_{\square}^{t-i}).$$

Similarly,

$$\begin{aligned}
\hat{V}_{\square}^{t+1} &= \sum_{i=1}^{t} (\pi_{t+1} \hat{Q}_{\square}^{i}) - \sum_{i=1}^{t-1} (\pi_t \hat{Q}_{\square}^{i}) \\
&\geq \sum_{i=1}^{t} (\pi_t \hat{Q}_{\square}^{i}) - \sum_{i=1}^{t-1} (\pi_t \hat{Q}_{\square}^{i}) && \text{(By the greediness of } \pi_{t+1}\text{)} \\
&= (\pi_t \hat{Q}_{\square}^{t}) \\
&= (\pi_t \square) + \gamma \hat{P}_{\pi_t}^{t} \hat{V}_{\square}^{t} \\
&= (\pi_t \square) + \gamma P_{\pi_t} \hat{V}_{\square}^{t} + \gamma(\hat{P}_{\pi_t}^{t} \hat{V}_{\square}^{t} - P_{\pi_t} \hat{V}_{\square}^{t}) \\
&\geq \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} (\square + \gamma(\hat{P}_{\pi_{t-i}}^{t-i} \hat{V}_{\square}^{t-i} - P_{\pi_{t-i}} \hat{V}_{\square}^{t-i})) && \text{(By induction on } t\text{)}
\end{aligned}$$

and

$$\begin{aligned}
\pi_{t+1} \prod_{i=1}^{t+1} \mathcal{T}^{\pi_{i-1}} Q_{\square}^{\pi_0} &= \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} (\pi_i \square) + \gamma^{t+1} \prod_{j=1}^{t} P_{\pi_{t-j+1}} (\pi_0 Q_{\square}^{\pi_0}) \\
\implies \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} (\pi_i \square) &\geq \pi_{t+1} \prod_{i=1}^{t+1} \mathcal{T}^{\pi_{i-1}} Q_{\square}^{\pi_0} - \gamma^{t+1} tH(1+\lambda_k)\mathbf{1}. \\
&\qquad \text{(Since } \textstyle\prod_{j=1}^{t} P\pi_{j-1} Q_{\square}^{\pi_0} \leq \gamma^{t+1} tH(1+\lambda_k)\mathbf{1}\text{)}
\end{aligned}$$

Combining the above, we obtain

$$\begin{aligned}
\hat{V}_{\square}^{t+1} &\geq \pi_t \prod_{i=1}^{t+1} \mathcal{T}^{\pi_{i-1}} Q_{\square}^{\pi_0} - \gamma^{t+1} tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma(\hat{P}_{\pi_{t-i}}^{t-i} \hat{V}_{\square}^{t-i} - P_{\pi_{t-i}} \hat{V}_{\square}^{t-i}) \\
&= V_{\square}^{\pi'_t} - \gamma^{t+1} tH(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma(\hat{P}_{\pi_{t-i}}^{t-i} \hat{V}_{\square}^{t-i} - P_{\pi_{t-i}} \hat{V}_{\square}^{t-i})
\end{aligned}$$

□

Lemma 54. *For any $t \in [T]$ and $k \in [K]$,*

$$\begin{aligned}
V_{\square}^{\pi_k^*} - \hat{V}_{\square}^{t} \leq \left(\gamma^t t + \frac{1+\lambda_k}{t(t-1)}\right) H\mathbf{1} - \sum_{i=1}^{t} \gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma(P_{\pi_{t-i}} \hat{V}_{\square}^{t-i} - \hat{P}_{\pi_{t-i}}^{t-i} \hat{V}_{\square}^{t-i}) \\
+ \sum_{i=1}^{t} \left((\gamma P_{\pi_k^*})^{t-i} \pi_k^* - \prod_{j=1}^{t-i} [\gamma P_{\pi_{t-j}}] \pi_i \right) \frac{1}{i} \sum_{j=0}^{i-1} \left[\gamma \hat{P}_j \hat{V}_{\square}^{j} - \gamma P \hat{V}_{\square}^{j} \right]
\end{aligned}$$

and

$$V_{\square}^{\pi_k^*} - \hat{V}_{\square}^{t} \geq -\gamma^t t H(1+\lambda_k)\mathbf{1} - \sum_{i=1}^{t-1}\gamma^i \prod_{j=t-i}^{t-1} P_{\pi_{t-j}}\gamma(P_{\pi_{t-i-1}}\hat{V}_{\square}^{t-i-1} - \hat{P}_{\pi_{t-i-1}}^{t-i-1}\hat{V}_{\square}^{t-i-1}).$$

Proof. From Lemma 53, we know

$$\begin{aligned}\hat{V}_{\square}^{t} &\leq V_{\square}^{\pi'_t} + \gamma^t t H(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t-1}\gamma^i \prod_{j=t-i}^{t-1} P_{\pi_{t-j}}\gamma(P_{\pi_{t-i-1}}\hat{V}_{\square}^{t-i-1} - \hat{P}_{\pi_{t-i-1}}^{t-i-1}\hat{V}_{\square}^{t-i-1}) \\ &\leq V_{\square}^{\pi_k^*} + \gamma^t t H(1+\lambda_k)\mathbf{1} + \sum_{i=1}^{t-1}\gamma^i \prod_{j=t-i}^{t-1} P_{\pi_{t-j}}\gamma(P_{\pi_{t-i-1}}\hat{V}_{\square}^{t-i-1} - \hat{P}_{\pi_{t-i-1}}^{t-i-1}\hat{V}_{\square}^{t-i-1})\end{aligned}$$

which gives us the second inequality. From Lemma 53 and Lemma 52 we have,

$$V_{\square}^{\pi'_t} - \hat{V}_{\square}^{t+1} \leq \gamma^t t H(1+\lambda_k)\mathbf{1} - \sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}\gamma(P_{\pi_{t-i}}\hat{V}_{\square}^{t-i} - \hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_{\square}^{t-i})$$

and

$$V_{\square}^{\pi_k^*} - V_{\square}^{\pi'_t} \leq \sum_{i=1}^{t}\left(\prod_{j=1}^{t-i}[\gamma P_{\pi_{t-j}}]\pi_i - (\gamma P_{\pi_k^*})^{t-i}\pi_k^*\right)\frac{1}{i}\sum_{j=0}^{i-1}\left[\gamma\hat{P}_j\hat{V}_{\square}^{j} - \gamma P\hat{V}_{\square}^{j}\right] + \frac{H(1+\lambda_k)}{t(t-1)}\mathbf{1}.$$

Combining them gives us the upper bound. □

Lemma 55. *For any $t \geq 2\log(t)/\gamma$ and $k \in [K]$,*

$$\sigma\left(\hat{V}_{\square}^{t}\right) \leq \left(\frac{2}{t} + 4H\sqrt{\frac{\iota}{M}}\right)H(1+\lambda_k)\mathbf{1} + \sigma(V_{\square}^{\pi_k^*})$$

with probability at least $1-\delta$.

Proof. We denote $\iota = \log(2|\mathcal{S}||\mathcal{A}|/\delta)$ throughout the proof. By Lemma 54,

$$\begin{aligned}V_{\square}^{\pi_k^*} - \hat{V}_{\square}^{t} \leq &\underbrace{\left(\gamma^t t + \frac{1}{t(t-1)}\right)H(1+\lambda_k)\mathbf{1}}_{\text{Term (i)}} - \underbrace{\sum_{i=1}^{t}\gamma^i \prod_{j=t-i+1}^{t} P_{\pi_{t-j}}\gamma(\hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_{\square}^{t-i} - P_{\pi_{t-i}}\hat{V}_{\square}^{t-i})}_{\text{Term (ii)}} \\ &+ \underbrace{\sum_{i=1}^{t}\left((\gamma P_{\pi_k^*})^{t-i}\pi_k^* - \prod_{j=1}^{t-i}[\gamma P_{\pi_{t-j}}]\pi_i\right)\frac{1}{i}\sum_{j=0}^{i-1}\left[\gamma\hat{P}_j\hat{V}_{\square}^{j} - \gamma P\hat{V}_{\square}^{j}\right]}_{\text{Term (iii)}} \qquad \text{(G.10)}\end{aligned}$$

We first bound Term (ii) and Term (iii). By Azuma-Hoeffding's inequality (Lemma 62), we have

$$\left\|P_{\pi_{t-i}}\hat{V}_{\square}^{t-i} - \hat{P}_{\pi_{t-i}}^{t-i}\hat{V}_{\square}^{t-i}\right\|_{\infty} \leq 2H(1+\lambda_k)\sqrt{\frac{\iota}{M}},$$

and by Lemma 61 with $t = i$, we have

$$\left\| \frac{1}{i} \sum_{j=0}^{i-1} \left[\gamma \hat{P}^{j}_{\pi_i} \hat{V}^{j}_{\square} - \gamma P_{\pi_i} \hat{V}^{j}_{\square} \right] \right\|_{\infty} \leq 2H(1+\lambda_k) \sqrt{\frac{\iota}{iM}}$$

each with probability at least $1-\delta$. Thus, to bound Term (ii), we have

$$\begin{align}
&\left\| \sum_{i=1}^{t} \gamma^{i} \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma (\hat{P}^{t-i}_{\pi_{t-i}} \hat{V}^{t-i}_{\square} - P_{\pi_{t-i}} \hat{V}^{t-i}_{\square}) \right\|_{\infty} \tag{G.11} \\
&\leq \sum_{i=1}^{t} \gamma^{i} \left\| \prod_{j=t-i+1}^{t} P_{\pi_{t-j}} \gamma \right\|_{1} \left\| \hat{P}^{t-i}_{\pi_{t-j}} \hat{V}^{t-i}_{\square} - P_{\pi_{t-i}} \hat{V}^{t-i}_{\square} \right\|_{\infty} \notag \\
&\leq \sum_{i=1}^{t} \gamma^{i} \left\| \hat{P}^{t-i}_{\pi_{t-j}} \hat{V}^{t-i}_{\square} - P_{\pi_{t-i}} \hat{V}^{t-i}_{\square} \right\|_{\infty} \notag \\
&\leq 2H^{2}(1+\lambda_k) \sqrt{\frac{\iota}{M}} \tag{G.12}
\end{align}$$

and to bound Term (iii) we have

$$\begin{align}
&\left\| \sum_{i=1}^{t} \left((\gamma P_{\pi^*_k})^{t-i} \pi^*_k - \prod_{j=1}^{t-i} [\gamma P_{\pi_{t-j}}] \pi_i \right) \frac{1}{i} \sum_{j=0}^{i-1} \left[\gamma \hat{P}_j \hat{V}^{j}_{\square} - \gamma P \hat{V}^{j}_{\square} \right] \right\|_{\infty} \notag \\
&\leq \sum_{i=1}^{t} \gamma^{t-i} \left\| \prod_{j=1}^{t-i} [P_{\pi_{t-j}}] \pi_i - (P_{\pi^*_k})^{t-i} \pi^*_k \right\|_{1} \left\| \frac{1}{i} \sum_{j=0}^{i-1} \left[\gamma \hat{P}_j \hat{V}^{j}_{\square} - \gamma P \hat{V}^{j}_{\square} \right] \right\|_{\infty} \notag \\
&\leq \sum_{i=1}^{t} \gamma^{t-i} \left\| \frac{1}{i} \sum_{j=0}^{i-1} \left[\gamma \hat{P}_j \hat{V}^{j}_{\square} - \gamma P \hat{V}^{j}_{\square} \right] \right\|_{\infty} \notag \\
&\leq \sum_{i=1}^{t} \gamma^{t-i} \sqrt{\frac{4H^{2}(1+\lambda_k)^{2}\iota}{iM}} \notag \\
&\leq \sum_{i=1}^{t} \gamma^{t-i} \sqrt{\frac{4H^{2}(1+\lambda_k)^{2}\iota}{M}} \notag \\
&\leq 2H^{2}(1+\lambda_k) \sqrt{\frac{\iota}{M}} \tag{G.13}
\end{align}$$

each with probability at least $1-\delta$. Lastly we bound Term (i). For $t \geq 2\log(t)/\gamma$, we have

$$\gamma^{t} \leq \frac{1}{t^{2}}$$

and thus

$$\gamma^{t} t + \frac{1}{t(t-1)} \leq \frac{1}{t} + \frac{1}{t} = \frac{2}{t}. \tag{G.14}$$

Combining eqs. (G.10) and (G.12) to (G.14), we have

$$|V_{\square}^{\pi_k^*} - \hat{V}_{\square}^t| \leq \left(\frac{2}{t} + 4H\sqrt{\frac{\iota}{M}}\right)(1+\lambda_k)H\mathbf{1}. \tag{G.15}$$

Finally, we have

$$\begin{aligned}\sigma\left(\hat{V}_{\square}^t\right) &\leq \sigma\left(V_{\square}^{\pi_k^*} - \hat{V}_{\square}^t\right) + \sigma\left(V_{\square}^{\pi_k^*}\right) && \text{(By Lemma 66)}\\ &\leq |V_{\square}^{\pi_k^*} - \hat{V}_{\square}^t| + \sigma\left(V_{\square}^{\pi_k^*}\right) && \text{(By Lemma 65)}\\ &\leq \left(\frac{2}{t} + 4H\sqrt{\frac{\iota}{M}}\right)(1+\lambda_k)H\mathbf{1} + \sigma\left(V_{\square}^{\pi_k^*}\right)\end{aligned}$$

with probability at least $1-\delta$, which completes the proof. □

Lemma 56 (Induction Lemma)**.** *Assume $X_k, A_k, B_k \geq 0$, $k = 1, \ldots,$ and $X_{k+1} \leq A_k X_k + B_k$, then we have $X_{k+1} \leq \prod_{i=1}^k A_i X_1 + \sum_{i=1}^k \prod_{j=i+1}^k A_j B_i$.*

G.3.2 Auxiliary Lemmas for Lemma 51

Lemma 57. *For any $t \in [T]$,*

$$\left\|\hat{\mathcal{V}}_{\diamond}^t - V_{\diamond}^{\pi}\right\|_{\infty} \leq \tilde{O}\left(\frac{H^2}{\sqrt{M}} + \gamma^t H\right)$$

with probability at least $1-\delta$.

Proof.

$$\begin{aligned}\hat{\mathcal{V}}_{\diamond}^t &= (\pi\hat{\mathcal{Q}}_{\diamond}^{t-1})\\ &= (\pi\diamond) + \gamma\hat{P}_{\pi}^{t-1}\hat{\mathcal{V}}_{\diamond}^{t-1}\\ &= (\pi\diamond) + \gamma P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-1} + \gamma(\hat{P}_{\pi}^{t-1}\hat{\mathcal{V}}_{\diamond}^{t-1} - P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-1})\\ &= \sum_{i=0}^{t}\gamma^i(P_{\pi})^i((\pi\diamond) + \gamma(\hat{P}_{\pi}^{t-i-1}\hat{\mathcal{V}}_{\diamond}^{t-i-1} - P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-i-1})) && \text{(By induction on } t\text{)}\\ &= \sum_{i=0}^{t}\gamma^i(P_{\pi})^i(\pi\diamond)\sum_{i=0}^{t}\gamma^{i+1}(P_{\pi})^i(\hat{P}_{\pi}^{t-i-1}\hat{\mathcal{V}}_{\diamond}^{t-i-1} - P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-i-1})\\ &= V_{\diamond}^{\pi} - \sum_{i=t+1}^{\infty}\gamma^i(P_{\pi})^i(\pi\diamond) + \sum_{i=0}^{t}\gamma^{i+1}(P_{\pi})^i(\hat{P}_{\pi}^{t-i-1}\hat{\mathcal{V}}_{\diamond}^{t-i-1} - P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-i-1}).\end{aligned}$$

Note that with probability at least $1-\delta$

$$\left\|\sum_{i=0}^{t}\gamma^{i+1}(P_{\pi})^i(\hat{P}_{\pi}^{t-i-1}\hat{\mathcal{V}}_{\diamond}^{t-i-1} - P_{\pi}\hat{\mathcal{V}}_{\diamond}^{t-i-1})\right\|_{\infty} \leq 2H^2\sqrt{\frac{\iota}{M}}$$

by a similar argument as in eq. (G.12), and

$$\left\|\sum_{i=t+1}^{\infty}\gamma^i(P_\pi)^i\pi\diamond\right\|_\infty \leq \gamma^t H.$$

We conclude that

$$\left\|\hat{\mathcal{V}}_\diamond^t - V_\diamond^\pi\right\|_\infty \leq \tilde{O}\left(\frac{H^2}{\sqrt{M}}+\gamma^t H\right)$$

with probability at least $1-\delta$. □

Lemma 58. *For any $i \in [T]$, we have*

$$\sigma(\hat{\mathcal{V}}_\diamond^i) \leq \tilde{O}\left(\frac{H^2}{\sqrt{M}}+\gamma^t H\right)\mathbf{1}+\sigma(V_\diamond^\pi)$$

with probability at least $1-\delta$.

Proof. We have

$$\begin{aligned}
\sigma(\hat{\mathcal{V}}_\diamond^i) &\leq \sigma(V_\diamond^\pi - \hat{\mathcal{V}}_\diamond^i)+\sigma(V_\diamond^\pi) && \text{(By Lemma 66)}\\
&\leq |V_\diamond^\pi - \hat{\mathcal{V}}_\diamond^i|+\sigma(V_\diamond^\pi) && \text{(By Lemma 65)}\\
&\leq \tilde{O}\left(\frac{H^2}{\sqrt{M}}+\gamma^t H\right)\mathbf{1}+\sigma(V_\diamond^\pi) && \text{(By Lemma 57)}
\end{aligned}$$

with probability at least $1-\delta$. □

G.4 Proof of Corollary 35

Corollary 59. *Let algorithm 6 and algorithm 7 be the instantiations of the `MDP-Solver` and `PolicyEvaluation` in algorithm 1. For a fixed $\varepsilon \in (0, 1/H^2]$, $\delta \in (0,1)$, algorithm 1 with $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$ samples, $U = O\left(\frac{1}{\zeta(1-\gamma)}\right)$, $\eta = \frac{U(1-\gamma)}{\sqrt{K}}$, $K = O\left(\frac{1}{\varepsilon^2\,(1-\gamma)^2}\right)$, and $b' = b - O(\varepsilon)$, returns a policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,*

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b - O(\varepsilon).$$

Under the same conditions, but with $b' = b + O(\varepsilon)$ and $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^5}{\zeta^2\varepsilon^2}\right)$ samples, algorithm 1 returns a policy $\bar{\pi}$ satisfying the following condition with probability $1-\delta$,

$$V_r^{\bar{\pi}}(\rho) \geq V_r^{\pi^*}(\rho) - O(\varepsilon), \quad \textit{and} \quad V_c^{\bar{\pi}}(\rho) \geq b.$$

Proof. By Lemma 33 and Lemma 34, the sample complexity required to ensure $f(\mathcal{B}) \leq O(\varepsilon)$ is $TM|\mathcal{C}| = \tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}\right)$. Therefore, the guarantee for the relaxed feasibility setting follows directly from our meta-theorem (Theorem 7). For the strict feasibility setting, we rescale ε by a factor of $O(\zeta(1-\gamma))$. Since $\varepsilon \le 1$ and $1-\gamma \le 1$, the condition of $f(\mathcal{B}) \le \zeta/6$ in Theorem 7 can be satisfied. The rescaling increases the sample complexity by a multiplicative factor of $\frac{1}{\zeta^2(1-\gamma)^2}$, thereby completing the proof. □

G.5 Instantiating the MDP-Solver: Model-based algorithm [Li et al., 2020]

Instead of using `MDVI-Tabular`, the tabular `MDP-Solver` subroutine in Algorithm 1 can be instantiated with any model-based method that computes an optimal policy with respect to the estimated model. In this section, we adapt the framework analyzed in [Li et al., 2020] to show that, when combined with our overall framework, certain model-based `MDP-Solver` algorithms can recover the near-optimal sample complexity for solving tabular constrained MDPs.

Since we are using model-based methods, we denote $\hat{P}$ as the probability transition kernel form by

$$\forall s' \in \mathcal{S}, \quad \widehat{P}(s' \mid s,a) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s^i_{s,a} = s'\}$$

where $(s^i_{s,a})_{i=1}^N$ are the next-state samples from $\mathcal{B} = \texttt{DataCollection}(\texttt{Gen}, \mathcal{S}\times\mathcal{A}, N)$. Denote the perturbed reward by

$$r_{\mathrm{p}}(s,a) = r(s,a) + \zeta(s,a), \quad \zeta(s,a) \sim \mathrm{Unif}(0,\xi)$$

where $\mathrm{Unif}(0,\xi)$ denotes the uniform distribution. For any policy π, denote $\hat{V}^{\pi}_{\mathrm{p}}$ the corresponding value function of the perturbed empirical MDP $\widehat{\mathcal{M}}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \hat{P}, r_{\mathrm{p}}, \gamma)$. Denote $\hat{\pi}^*_p$ the optimal policy w.r.t. $\widehat{\mathcal{M}}_{\mathrm{p}}$ (i.e. $\hat{\pi}^*_p = \arg\max_\pi \hat{V}^{\pi}_{\mathrm{p}}$). Their main result is stated as follows.

Theorem 60 (Theorem 1 in [Li et al., 2020])**.** *There exist some universal constants $c_0, c_1 > 0$ such that: for any $\delta > 0$ and any $0 < \varepsilon \le \frac{1}{1-\gamma}$, the policy $\hat{\pi}^*_p$ defined in (9) obeys*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad V^{\hat{\pi}^*_p}(s) \ge V^*(s) - \varepsilon \quad \textit{and} \quad Q^{\hat{\pi}^*_p}(s,a) \ge Q^*(s,a) - \gamma\varepsilon, \tag{11}$$

with probability at least $1-\delta$, provided that the perturbation size is $\xi = \frac{c_1(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ and that the sample size per state-action pair exceeds

$$N \ge \frac{c_0 \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta}\right)}{(1-\gamma)^3\varepsilon^2}. \tag{12}$$

In addition, both the empirical QVI and PI algorithms w.r.t. $\widehat{\mathcal{M}}_p$ (cf. [Azar et al., 2013], Algorithms 1-2) are able to recover $\hat{\pi}^*_p$ perfectly within $\mathcal{O}\left(\frac{1}{1-\gamma}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon\delta}\right)\right)$ iterations.

Therefore, let $\mathcal{B} = \texttt{DataCollection}(\texttt{Gen}, \mathcal{S}\times\mathcal{A}, N)$. Then, by instantiating $\texttt{MDP-Solver}(r + \lambda_k c, \mathcal{B}, \phi)$ with any model-based algorithm that returns an optimal policy with respect to the perturbed empirical MDP constructed from $\mathcal{B}$, Assumption 4 can be satisfied with

$f_{\mathrm{mdp}}(\mathcal{B}) = O(\varepsilon)$. As a consequence, we recover the near-optimal sample complexity bounds for solving tabular constrained MDPs via our meta-theorem (Theorem 7). Furthermore, the limited rage of ε (i.e. $(0, 1/H^2]$) in Corollary 35 will be improved to a full range (i.e. $(0, H]$).

Appendix H

Supporting Lemmas

H.1 Concentration Inequalities

The following lemma is used throughout the paper. In the linear setting, we take $\mathcal{C}$ to be the core set and set $R = (1 + \lambda_k)H$. In the tabular setting, we let $\mathcal{C} = \mathcal{S} \times \mathcal{A}$ and set $R = H$.

Lemma 61. *Let $\hat{V}^i$ be an empirical value function with entries bounded in $[0, R]$, and let $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{A}$. Then, for any $t \in 1, \dots, T$, the following holds:*

$$\mathbb{P}\left(\exists (s,a) \in \mathcal{C} \quad s.t. \quad \frac{1}{t}\sum_{i=0}^{t-1}[(\hat{P}_i\hat{V}^i)(s,a) - (P\hat{V}^i)(s,a)] \geq 2R\sqrt{\log(2|\mathcal{C}|/\delta)/tM}\right) \leq \delta$$

Proof. Consider a fixed $t \in \{1, \cdots, T\}$ and $(s,a) \in \mathcal{C}$. Denote $y_{t,m,s,a}$ as the m'th next-state sample we collect for state-action pair (s,a) at iteration t. Since

$$\begin{aligned}\frac{1}{t}\sum_{i=0}^{t-1}\left[(\hat{P}_i\hat{V}^i)(s,a) - (P\hat{V}^i)(s,a)\right] &= \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}^i(y_{t,m,s,a}) - (P\hat{V}^i)(s,a)\right] \\ &= \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}^i(y_{t,m,s,a}) - (P\hat{V}^i)(s,a)\right]\end{aligned}$$

is a sum of bounded martingale differences with respect to the filtration $(\mathcal{F}_{i,m})_{i=0,m=1}^{t-1,M}$. Thus, using the Azuma-Hoeffding inequality (Lemma 62),

$$\mathbb{P}\left(\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}^i(y_{t,m,s,a}) - (P\hat{V}^i)(s,a)\right] \geq 2R\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{tM}}\right) \leq \frac{\delta}{|\mathcal{C}|}.$$

Taking the union bound over $(s,a) \in \mathcal{C}$

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{C}}\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}^i(y_{t,m,s,a}) - (P\hat{V}^i)(s,a)\right] \leq 2R\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{tM}}\right)$$

$$\geq 1 - \sum_{(s,a)\in\mathcal{C}} \mathbb{P}\left(\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{M}\sum_{m=1}^{M}\left[\hat{V}^i(y_{t,m,s,a}) - (P\hat{V}^i)(s,a)\right] \geq 2R\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{tM}}\right)$$
$$\geq 1-\delta,$$

which implies the desired result. □

Lemma 62 (Azuma-Hoeffding Inequality). *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$. Assume that $X_n \in [l_n, u_n]$ and $\mathbb{E}_n[X_n] = 0$ almost surely, for all n. Then,*

$$\mathbb{P}\left(\sum_{n=1}^{N} X_n \geq \sqrt{\sum_{n=1}^{N}\frac{(u_n - l_n)^2}{2}\log\frac{1}{\delta}}\right) \leq \delta$$

for any $\delta \in (0,1)$.

Lemma 63 (Bernstein's Inequality). *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$.Suppose that $X_n \leq U$ and $\mathbb{E}_n[X_n] = 0$ almost surely, for all n. Then, letting $Z' := \sum_{n=1}^N \mathbb{E}_n[X_n^2]$,*

$$\mathbb{P}\left(\sum_{n=1}^{N} X_n \geq \frac{2U}{3}\log\frac{1}{\delta} + \sqrt{2Z\log\frac{1}{\delta}} \text{ and } Z' \leq Z\right) \leq \delta$$

for any $Z \in [0,\infty)$ and $\delta \in (0,1)$.

Lemma 64 (Conditional Bernstein's Inequality). *Consider the same notations and assumptions in Lemma 63. Furthermore, let $\mathcal{E}$ be an event that implies $Z' \leq Z$ for some $Z \in [0,\infty)$ with $\mathbb{P}(\mathcal{E}) \geq 1-\delta'$ for some $\delta' \in (0,1)$. Then,*

$$\mathbb{P}\left(\sum_{n=1}^{N} X_n \geq \frac{2U}{3}\log\frac{1}{\delta(1-\delta')} + \sqrt{2Z\log\frac{1}{\delta(1-\delta')}}\,\middle|\, \mathcal{E}\right) \leq \delta$$

for any $\delta \in (0,1)$.

H.2 Lemmas for Variances

Lemma 65 (Popoviciu's Inequality for Variances). *The variance of any random variable bounded by x is bounded by x^2.*

Lemma 66 (Azar et al. [2013]). *Suppose two real-valued random variables X, Y whose variances, $\mathbb{V}X$ and $\mathbb{V}Y$, exist and are finite. Then, $\sqrt{\mathbb{V}X} \leq \sqrt{\mathbb{V}[X-Y]} + \sqrt{\mathbb{V}Y}$.*

Lemma 67 (Total variance lemma Azar et al. [2013]). *For any policy π, $\|(I-P_\pi)^{-1}\sigma(V^\pi)\|_\infty \leq \sqrt{2H^3}$.*

H.3 Lemmas for Constrained MDPs

Lemma 68 (Constraint violation bound, Lemma B.2 in Jain et al. [2022]). *For any $C \geq \lambda^*$ and any π s.t. $V_r^*(\rho) - V_r^\pi(\rho) + C[b - V_c^\pi(\rho)]_+ \leq \beta$, we have $[b - V_c^\pi(\rho)]_+ \leq \frac{\beta}{C-\lambda^*}$.*

Lemma 69 (Bounding the dual variable, Lemma 4.1 in Jain et al. [2022]). *The objective eq.* (2.1) *satisfies strong duality, and the optimal dual variables are bounded as*

$$\lambda^* \leq \frac{1}{(1-\gamma)\zeta}, \quad \textit{where } \zeta := \max_\pi V_c^\pi(\rho) - b > 0.$$

Lemma 70 (Bounding the sensitivity error, Lemma 13 in Vaswani et al. [2022]). *If we have*

$$\hat{\pi}^* \in \arg\max_\pi V_r^\pi(\rho) \;\; \textit{s.t.} \;\; V_c^\pi(\rho) \geq b + \Delta$$

$$\tilde{\pi}^* \in \arg\max_\pi V_r^\pi(\rho) \;\; \textit{s.t.} \;\; V_c^\pi(\rho) \geq b - \Delta,$$

then the sensitivity error term can be bounded by:

$$\left| V_r^{\hat{\pi}^*}(\rho) - V_r^{\tilde{\pi}^*}(\rho) \right| \leq 2\Delta\lambda^*$$

where λ^* is the optimal Lagrange multiplier (i.e., the solution to eq. (3.1)*).*

Appendix I

Table of Notation

Notation	Meaning
$\mathcal{A}, \mathcal{S}$	action space of size $\lvert\mathcal{A}\rvert$, state space of size $\lvert\mathcal{S}\rvert$
γ, H	discount factor in $[0, 1)$, $1/(1-\gamma)$
P	transition matrix $P \in \mathbb{R}^{\lvert S\rvert\lvert A\rvert \times \lvert S\rvert}$
$P_\pi, \hat{P}^t_\pi$	$\pi P \in \mathbb{R}^{\lvert S\rvert \times \lvert S\rvert}$, $\pi \hat{P}_t \in \mathbb{R}^{\lvert S\rvert \times \lvert S\rvert}$
r, c	reward vector in [0,1] range, constraint reward vector in [0,1] range
ρ	initial distribution of states
$\diamond$	r or c
$\square$	$r + \lambda c$ where $\lambda \in \{\lambda_1, \cdots, \lambda_K\}$
b, ζ	constraint value in $[0, 1/(1-\gamma))$, Slater constant
λ, λ^*	Lagrange multiplier, the optimal Lagrange multiplier
U	projection upper bound
ϕ, d	feature map of a linear MDP and its dimension
$\tilde{\rho}, \mathcal{C}$	a design over $\mathcal{S} \times \mathcal{A}$, coreset
G	design matrix with respect to ϕ and $\tilde{\rho}$. Equal to $\sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x, b)\phi(x, b)\phi(x, b)^\top$
$W(z)$	$G^{-1} \sum_{(x,b)\in\mathcal{C}} \tilde{\rho}(x, b)\phi(x, b)z(x, b)$
ε, δ	admissible suboptimality, admissible failure probability
K, T	number of outer and inner iterations
$(\hat{P}_t \hat{V}^t_\diamond)(s, a)$	$\frac{1}{M} \sum_{m=1}^{M} \hat{V}^t_\diamond(s'_m)$ where $s'_m \in \mathcal{B}_t$
$(P\hat{V}^t_\diamond)(s, a)$	$\mathbb{E}[\hat{V}^t_\diamond(s') \mid s_0 = s, a_0 = a]$
$\mathcal{F}_{t,m}$	σ-algebra in the filtration for algorithms 2, 3, 6 and 7
$\mathcal{T}^\pi Q$	Bellman operator $r + \gamma P(\pi Q)$
Q^π	state-action value function for policy π
$\hat{Q}^t_\square$	estimated state-action value function in iteration t in algorithms 2 and 6
$\hat{\mathcal{Q}}^t_\diamond$	estimated state-action value function in iteration t in algorithms 3 and 7

$\hat{V}^t_\square(s)$ (Tabular)	$\max\limits_a \left\{\sum_{i=0}^t \hat{Q}^i_\square(s,a)\right\} - \max\limits_a \left\{\sum_{i=0}^{t-1} \hat{Q}^i_\square(s,a)\right\}$ in algorithm 6
$\hat{V}^t_\square(s)$ (Linear)	$\max\limits_a \left\{(\langle\phi, \sum_{i=0}^t \theta^i_\square\rangle)(s,a)\right\} - \max\limits_a \left\{(\langle\phi, \sum_{i=0}^{t-1} \theta^i_\square\rangle)(s,a)\right\}$ in algorithm 2
$\tilde{Q}^t_\square$ (Tabular)	$\sum_{i=0}^t \hat{Q}^i_\square$ in algorithm 6
$\tilde{Q}^t_\square$ (Linear)	$\langle\phi, \sum_{i=0}^t \theta^i_\square\rangle$ in algorithm 2
$\hat{\mathcal{V}}^t_\diamond(s)$ (Tabular)	$(\pi\hat{\mathcal{Q}}^t_\diamond)(s)$ in algorithm 7
$\hat{\mathcal{V}}^t_\diamond(s)$ (Linear)	$(\pi\,\langle\phi, \omega^t_\diamond\rangle)(s)$ in algorithm 3
$\bar{V}^t_\square(s), \bar{\mathcal{V}}^t_\diamond(s)$	$\frac{1}{t}\sum_{i=1}^t \hat{V}^i_\square(s), \frac{1}{t}\sum_{i=1}^t \hat{\mathcal{V}}^i_\diamond(s)$
$\hat{V}^k_\diamond, \bar{\mathcal{V}}^{\bar{\pi}}_\diamond$	output of the `PolicyEvaluation` oracle in line 5 in Algorithm 1, $\frac{1}{K}\sum_{k=0}^{K-1} \hat{V}^k_\diamond$
π_k	output policy of `MDP-Solver`
$\bar{\pi}$	mixture policy equal to $\frac{1}{K}\sum_{k=0}^{K-1} \pi_k$
π^*	$\text{argmax}_\pi V^\pi_r(\rho)$ s.t. $V^\pi_c(\rho) \geq b$
π^{*+}	$\text{argmax}_\pi V^\pi_r(\rho)$ s.t. $V^\pi_c(\rho) \geq b + 6f(\mathcal{B})$
π^*_k	$\text{argmax}_\pi\{V^\pi_{r+\lambda_k c}\}$
π'_t	a non-stationary policy that follows policies $\pi_t, \pi_{t-1}, \ldots$
	upto timestep t and follows π_0 thereafter
$(\pi Q)(s)$	$\sum_{a\in\mathcal{A}} \pi(a\|s)\, Q(s,a)$
$(\pi r)(s)$	$\sum_{a\in\mathcal{A}} \pi(a\|s)\, r(s,a)$
$\theta^t_\square$	least-squares value estimate in algorithm 2
$\boldsymbol{\theta}^t_\square$	parameter that satisfies $\langle\phi, \boldsymbol{\theta}^t_\square\rangle := \square + \gamma P\hat{V}^{t-1}_\square$ in the linear MDP
$\omega^t_\diamond$	least-squares value estimate in algorithm 3
$\boldsymbol{\omega}^t_\diamond$	parameter that satisfies $\langle\phi, \boldsymbol{\omega}^t_\diamond\rangle := \diamond + \gamma P\hat{\mathcal{V}}^{t-1}_\diamond$ in the linear MDP