KL-Entropy-Regularized RL with a Generative Model is Minimax Optimal

Tadashi Kozuno*[,1], Wenhao Yang[2], Nino Vieillard[3,4], Toshinori Kitamura[5], Yunhao Tang[6], Jincheng Mei[3], Pierre Ménard[7], Mohammad Gheshlaghi Azar[6], Michal Valko[6], Rémi Munos[6], Olivier Pietquin[3], Matthieu Geist[3], Csaba Szepesvári[1,6]

Abstract

In this work, we consider and analyze the sample complexity of model-free reinforcement learning with a generative model. Particularly, we analyze mirror descent value iteration (MDVI) by Geist et al. (2019) and Vieillard et al. (2020a), which uses the Kullback-Leibler divergence and entropy regularization in its value and policy updates. Our analysis shows that it is nearly minimax-optimal for finding an ε-optimal policy when ε is sufficiently small. This is the first theoretical result that demonstrates that a simple model-free algorithm without variance-reduction can be nearly minimax-optimal under the considered setting.

1 Introduction

In the generative model setting, the agent has access to a simulator of a Markov decision process (MDP), to which the agent can query next states of arbitrary state-action pairs (Azar et al., 2013). The agent seeks a near-optimal policy using as small number of queries as possible.

While the generative model setting is simpler than the online reinforcement learning (RL) setting, proof techniques developed under this setting often generalize to more complex settings. For example, the total-variance technique developed by Azar et al. (2013) and Lattimore & Hutter (2012) is now an indispensable tool for a sharp analysis of RL algorithms in the online RL setting for tabular MDP (Azar et al., 2017; Jin et al., 2018) and linear function approximation (Zhou et al., 2021).

In this paper, we consider a model-free approach for the generative model setting with tabular MDP. Particularly, we analyze mirror descent value iteration (MDVI) by Geist et al. (2019) and Vieillard et al. (2020a), which uses Kullback-Leibler (KL) divergence and entropy regularization in its value and policy updates. We prove its near minimax-optimal sample complexity for finding an ε-optimal policy when ε is sufficiently small. Our result and analysis have the following consequences.

First, we demonstrate the effectiveness of KL and entropy regularization. There are some previous works that argue the benefit of regularization from a theoretical perspective in value-iteration-like algorithms (Kozuno et al., 2019; Vieillard et al., 2020a,b) and policy optimizaiton (Mei et al., 2020; Cen et al., 2021; Lan, 2022). Compared to those works, we show that simply combining value iteration with regularization achieves the near minimax-optimal sample complexity.

Second, as discussed by Vieillard et al. (2020a), MDVI encompasses various algorithms as special cases or equivalent forms. While we do not analyze each algorithm, most of them are minimax-optimal too in the generative model setting with tabular MDP.

Lastly and most importantly, MDVI uses no variance-reduction technique, in contrast to previous model-free approaches (Sidford et al., 2018; Wainwright, 2019; Khamaru et al., 2021). Consequently, our analysis is straightforward, and it would be easy to extend it to more complex settings, such as the online RL and linear function approximation. Furthermore, previous approaches need pessimism to obtain a near-optimal

*Correspondence: tadashi.kozuno@gmail.com. [1] University of Alberta, [2] Peking University, [3] Google Research, Brain team, [4] Université de Lorraine, CNRS, INRIA, IECL, [5] University of Tokyo, [6] DeepMind, [7] Otto von Guericke University Magdeburg.

policy, which prevents them from being extended to the online RL setting, where the optimism plays an important role for an efficient exploration (Azar et al., 2017; Jin et al., 2018). On the other hand, MDVI is compatible with optimism. Our analysis paves the way for the combination of online exploration techniques with minimax model-free algorithms.

2 Related work

Write γ, H, X, and A for the discount factor, effective horizon $\frac{1}{1-\gamma}$, and number of states and actions.

Learning with a generative model In the generative model setting, there are two problem settings: finding (i) an ε-optimal Q-value function with probability at least $1-\delta$, and (ii) an ε-optimal policy with probability at least $1-\delta$, where $\delta \in (0,1)$, and $\varepsilon > 0$. Both problems are known to have sample complexity lower bounds of $\Omega(XAH^3/\varepsilon^2)$ (Azar et al., 2013; Sidford et al., 2018). Note that even if an ε-optimal Q-value function is obtained, additional data and computation are necessary to find an ε-optimal policy (Sidford et al., 2018). In this paper, we consider the learning of an ε-optimal policy.

There exist minimax-optimal model-based algorithms for learning a near-optimal value function (Azar et al., 2013) and policy (Agarwal et al., 2020; Li et al., 2020). Also, there exist minimax-optimal model-free algorithms for learning a near-optimal value function (Wainwright, 2019; Khamaru et al., 2021; Li et al., 2021b) and policy (Sidford et al., 2018). While model-based algorithms are conceptually simple, they have a higher computational complexity than that of model-free algorithms. The algorithm (MDVI) we analyze in this paper is a model-free algorithm for finding a near-optimal policy, and has a low computational complexity.

Arguably, Q-learning is one of the simplest model-free algorithms (Watkins & Dayan, 1992; Even-Dar et al., 2003). Unfortunately, Li et al. (2021a) provide a tight analysis of Q-learning and show that its sample complexity is $\widetilde{\Theta}(XAH^4/\varepsilon^2)$ for finding an ε-optimal Q-value function, [1] which is one H factor away from the lower bound. To remove the extra H factor, some works (Sidford et al., 2018; Wainwright, 2019; Khamaru et al., 2021) leverage variance reduction techniques. While elegant, variance reduction techniques lead to multi-epoch algorithms with involved analyses. In contrast, MDVI requires no variance reduction and is significantly simpler.

MDVI's underlying idea that enables such simplicity is, while implicit, the averaging of value function estimates. Li et al. (2021b) shows that averaging Q-functions computed in Q-learning can find a near-optimal Q-function with a minimax-optimal sample complexity. Azar et al. (2011) also provides a simple algorithm called Speedy Q-learning (SQL), which performs the averaging of value function estimates. In fact, as argued in (Vieillard et al., 2020a), SQL is equivalent to a special case of MDVI with only KL regularization. While previous works on MDVI (Vieillard et al., 2020a) and an equivalent algorithm called CVI (Kozuno et al., 2019) provide error propagation analyses, they do not provide sample complexity. [2] This paper proves the first nearly minimax-optimal sample complexity bound for MDVI-type algorithm. We tighten previous results by (i) using the entropy regularization, which speeds up the convergence rate, (ii) improved error propagation analyses (Lemmas 1 and 9), and (iii) careful application of the total variance technique (Azar et al., 2013).

In addition to the averaging, Theorem 1 is based on the idea of using a non-stationary policy (Scherrer & Lesner, 2012). While the last policy of MDVI is near-optimal when ε is small, a non-stationary policy constructed from policies outputted by MDVI is near-optimal for a wider range of ε.

Range of Valid ε Although there are multiple minimax-optimal algorithms for the generative model setting, there ranges of valid ε differ. The model-based algorithm by Azar et al. (2013) is nearly minimax-optimal for $\varepsilon \leq \sqrt{H/X}$, which is later improved to $\sqrt{H}$ by Agarwal et al. (2020), and to H by Li et al. (2020). As for model-free approaches, the algorithm by Sidford et al. (2018) is nearly minimax-optimal for $\varepsilon \leq 1$. MDVI is nearly minimax-optimal for $\varepsilon \leq 1/\sqrt{H}$ (non-stationary policy case, Theorem 1) and $\varepsilon \leq 1/H$ (last policy

[1] $\widetilde{\mathcal{O}}$ hides terms poly-logarithmic in H, X, A, $1/\varepsilon$, and $1/\delta$.

[2] Vieillard et al. (2020a) note SQL's sample complexity of $\widetilde{\mathcal{O}}(XAH^4/\varepsilon^2)$ for finding a near-optimal policy as a corollary of their result without proof.

case, Theorem 2). Therefore, it has one of the narrowest range of valid ε (second worst) compared to other algorithms. It is unclear if this is an artifact of our analysis or the real limitation of MDVI-type algorithm. We leave this topic as a future work.

Regularization in MDPs Sometimes, regularization is added to the reward to encourage exploration in MDPs (Fox et al., 2016; Vamplew et al., 2017). In recent years, Neu et al. (2017); Geist et al. (2019); Lee et al. (2018); Yang et al. (2019) have provided a unified framework for regularized MDPs. Specifically, Geist et al. (2019) propose the Regularized Modified Policy Iteration algorithm and Mirror Descent Modified Policy Iteration to solve regularized MDPs. In the meantime, Vieillard et al. (2020a) provide theoretical guarantees of KL-regularized value iteration in the approximate setting. Particularly, they show that KL regularization results in the averaging of Q-value functions and show that the averaging leads to an improved error propagation result. We extend their improved error propagation result to a KL- and entropy- regularization case. Our results provide theoretical underpinnings to many regularized RL algorithms in Vieillard et al. (2020a, Table 1) and a high-performing deep RL algorithm called Munchausen DQN (Vieillard et al., 2020b).

3 Preliminaries

For a set $\mathbf{S}$, we denote its complement as $\mathbf{S}^c$. For a positive integer N, we let $[N] \coloneqq \{1, \dots, N\}$. Without loss of generality, every finite set is assumed to be a subset of integers. For a finite set, say $\mathbf{S}$, the set of probability distributions over $\mathbf{S}$ is denoted by $\Delta(\mathbf{S})$. For a vector $v \in \mathbf{R}^M$, its m-th element is denoted by v_m or $v(m)$. [3] We let $\mathbf{1} \coloneqq (1, \dots, 1)^\top$ and $\mathbf{0} \coloneqq (0, \dots, 0)^\top$, whose dimension will be clear from the context. For a matrix $A \in \mathbf{R}^{N \times M}$, we denote its n-th row and m-th value of the n-th row by A_n and A_n^m, respectively. The expectation and variance of a random variable X are denoted as $\mathbb{E}[X]$ and $\mathbb{V}[X]$, respectively. The empty sum is defined to be 0, e.g., $\sum_{i=j}^{k} a_i = 0$ if $j > k$.

We consider a Markov Decision Process (MDP) defined by $(\mathbf{X}, \mathbf{A}, \gamma, r, P)$, where $\mathbf{X}$ is the state space of size X, $\mathbf{A}$ the action space of size A, $\gamma \in [0, 1)$ the discount factor, $r \in [-1, 1]^{XA}$ the reward vector with $r_{x,a}$ denoting the reward when taking an action a at a state x, and $P \in \mathbf{R}^{XA \times X}$ state transition probability matrix with $P_{x,a}^y$ denoting the state transition probability to a new state y from a state x when taking an action a. We let H be the (effective) time horizon $1/(1-\gamma)$.

Note that $(Pv)(x, a) = \mathbb{E}[v(X_1) | X_0 = x, A_0 = a]$ for any $v \in \mathbf{R}^X$. Any policy π is identified as a matrix $\pi \in \mathbf{R}^{X \times XA}$ such that $(\pi q)(x) \coloneqq \sum_{a \in \mathbf{A}} \pi(a|x) q(x, a)$ for any $q \in \mathbf{R}^{XA}$. For convenience, we adopt a shorthand notation, $P^\pi \coloneqq P\pi$. With these notations, the Bellman operator T^π for a policy π is defined as an operator such that $T^\pi q \coloneqq r + \gamma P^\pi q$. The Q-value function q^π for a policy π is its unique fixed point. The state-value function v^π is defined as πq^π. An optimal policy π^* is a policy such that $v^* \coloneqq v^{\pi^*} \geq v^\pi$ for any policy π, where the inequality is point-wise.

4 Mirror Descent Value Iteration and Main Results

For any policies π and μ, let $\log \pi$ and $\log \frac{\pi}{\mu}$ be the functions $x, a \mapsto \log \pi(a|x)$ and $x, a \mapsto \log \frac{\pi(a|x)}{\mu(a|x)}$ over $\mathbf{X} \times \mathbf{A}$. We analyze (approximate) MDVI whose update is the following (Vieillard et al., 2020a):

$$q_{k+1} = r + \gamma P v_k + \varepsilon_k \,, \tag{1}$$

where $v_k \coloneqq \pi_k \left(q_k - \tau \log \frac{\pi_k}{\pi_{k-1}} - \kappa \log \pi_k \right)$,

$$\pi_k(\cdot|x) = \arg\max_{p \in \Delta(\mathbf{A})} \sum_{a \in \mathbf{A}} p(a) \left(q_k(s, a) - \tau \log \frac{p(a)}{\pi_{k-1}(a|x)} - \kappa \log p(a) \right) \tag{2}$$

[3] Unless noted otherwise, all vectors are column vectors.

Algorithm 1: MDVI(α, K, M)

Input: $\alpha \in [0,1)$, number of iterations K, and number of next-state samples per iteration M.
Let $s_0 = \mathbf{0} \in \mathbf{R}^{X A}$ and $w_0 = w_{-1} = \mathbf{0} \in \mathbf{R}^X$;
for k from 0 to $K - 1$ **do**
  Let $v_k = w_k - \alpha w_{k-1}$;
  for each state-action pair $(x,a) \in \mathbf{X} \times \mathbf{A}$ **do**
    Sample $(y_{k,m,x,a})_{m=1}^M$ from the generative model $P(\cdot|x,a)$;
    Let $q_{k+1}(x,a) = r(x,a) + \gamma M^{-1} \sum_{m=1}^M v_k(y_{k,m,x,a})$;
  end
  Let $s_{k+1} = q_{k+1} + \alpha s_k$ and $w_{k+1}(x) = \max_{a \in \mathbf{A}} s_{k+1}(x,a)$ for each $x \in \mathbf{X}$;
end
return $(\pi_k)_{k=0}^K$, *where* π_k *is greedy policy with respect to* s_k;

for all $x \in \mathbf{X}$, and $\varepsilon_k : \mathbf{X} \times \mathbf{A} \to \mathbf{R}$ is an "error" function, which abstractly represents the deviation of q_{k+1} from the update target $r + \gamma P v_k$. In other words, MDVI is value iteration with KL and entropy regularization.

Let $s_k := q_k + \alpha s_{k-1} = \sum_{j=0}^k \alpha^j q_{k-j}$. The policy (2) can be rewritten as a Boltzmann policy of s_k, i.e., $\pi_k(a|x) \propto \exp(\beta s_k(x,a))$, where $\alpha := \tau/(\tau+\kappa)$, and $\beta := 1/(\tau+\kappa)$, see Appendix B for details. Substituting π_{k-1} and π_k in v_k with this expression of the policy, we deduce that

$$v_k(x) = \frac{1}{\beta} \log \sum_{a \in \mathbf{A}} \exp(\beta s_k(x,a)) - \frac{\alpha}{\beta} \log \sum_{a \in \mathbf{A}} \exp(\beta s_{k-1}(x,a)) .$$

Thus, letting w_k be the function $x \mapsto \beta^{-1} \log \sum_{a \in \mathbf{A}} \exp(\beta s_k(x,a))$ over $\mathbf{X}$, MDVI's update rules can be equivalently written as

$$q_{k+1} = r + \gamma P(w_k - \alpha w_{k-1}) + \varepsilon_k \text{ and } \pi_k(a|x) \propto \exp(\beta s_k(x,a)) \text{ for all } (x,a) \in \mathbf{X} \times \mathbf{A} .$$

A sample-approximate version of MDVI shown in Algorithm 1 (MDVI) uses this equivalent form of MDVI. Furthermore, for simplicity of the analysis, we consider the limit of $\tau, \kappa \to 0$ while keeping $\alpha = \tau/(\tau+\kappa)$ to a constant value (which corresponds to letting $\beta \to \infty$).

Remark 1. *Even if β is finite, MDVI is nearly minimax-optimal as long as β is large enough. Indeed, $\beta^{-1} \log \sum_{a \in \mathbf{A}} \exp(q(x,a))$ satisfies (Kozuno et al., 2019, Lemma 7) that*

$$\max_{a \in \mathbf{A}} q(x,a) - \beta^{-1} \log \sum_{a \in \mathbf{A}} \exp(q(x,a)) - \max_{a \in \mathbf{A}} q(x,a) + \beta^{-1} \log A .$$

Thus, while β appears in the proofs of Theorems 1 and 2 if it is finite, it always appear as $\beta^{-1} \log A$ multiplied by H-dependent constant. Therefore, MDVI is nearly minimax-optimal as long as β is large enough.

Why KL Regularization? The weight α used in s_k updates monotonically increases as the coefficient of the KL regularization τ increases. As we see later, error terms appear in upper bounds of $\|v^* - v^{\pi_k}\|_1$ as $(1-\alpha)\sum_{j=1}^k \alpha^{k-j}\varepsilon_j$. Applying Azuma-Hoeffding inequality, it is approximately bounded by $H\sqrt{\frac{1}{1-\alpha}}$. Therefore, MDVI becomes more robust to sampling error as α increases. The KL regularization confers this benefit to the algorithm.

Why Entropy Regularization? When there is no entropy regularization ($\alpha = 1$), the convergence rate of MDVI becomes $1/K$ while it is α^K for $\gamma - \alpha < 1$ (Vieillard et al., 2020a). In the former case, we need to set $K - H^2/\varepsilon$, whereas in the latter case, $K - 1/(1-\alpha)$ suffices. Since we will set α to either γ or $1 - (1-\gamma)^2$, $K - H$ or H^2. Thus, we can use more samples per one value update (i.e., larger M). A larger M leads to a smaller value estimation variance ($\sigma(v_k)$ in Lemma 6), which is important to improve the range of ε. Even when $\alpha = 1$, MDVI is nearly minimax-optimal (proof omitted). However, ε must be less than or equal to $1/H^2$.

Main Theoretical Results The following theorems show the near minimax-optimality of MDVI. For a sequence of policies $(\pi_k)_{k=0}^K$ outputted by MDVI, we let π_k^0 be the non-stationary policy that follows $\pi_k \; _t$ at the t-th time step until $t = k$, after which π_0 is followed. [4] Note that the value function of such a non-stationary policy is given by $v^{\pi_k^0} = \pi_k T^{\pi_k \; 1} \quad T^{\pi_1} q^{\pi_0}$.

Theorem 1. *Assume that ε 2 $(0, 1/^p\overline{H}]$. Then, there exist positive constants $c_1, c_2 \quad 1$ independent of H, X, A, ε, and δ such that when MDVI is run with the settings*

$$\alpha = \gamma\,, K = \left\lceil \frac{3}{1 \quad \alpha} \log \frac{c_1 H}{\varepsilon} + 2 \right\rceil \text{, and } M = \left\lceil \frac{c_2 H^2}{\varepsilon^2} \log \frac{16KXA}{\delta} \right\rceil ,$$

it outputs a sequence of policies $(\pi_k)_{k=0}^K$ such that $kv \quad v^{\pi_K^0} k_1 \quad \varepsilon$ with probability at least $1 \quad 3\delta/4$, using $\widetilde{O}\left(H^3 XA/\varepsilon^2\right)$ samples from the generative model.

Storing all policies requires the memory space of KXA and can be prohibitive in some cases. The next theorem shows that the last policy outputted by MDVI is near-optimal when $\varepsilon \quad 1/H$.

Theorem 2. *Assume that ε 2 $(0, 1/H]$. Then, there exist positive constants $c_3, c_4 \quad 1$ independent of H, X, A, ε, and δ such that when MDVI is run with the settings*

$$\alpha = 1 \quad (1 \quad \gamma)^2\,, K = \left\lceil \frac{5}{1 \quad \alpha} \log \frac{c_3 H}{\varepsilon} + 2 \right\rceil \text{, and } M = \left\lceil \frac{c_4 H}{\varepsilon^2} \log \frac{16KXA}{\delta} \right\rceil ,$$

it outputs a sequence of policies $(\pi_k)_{k=0}^K$ such that $kv \quad v^{\pi_K} k_1 \quad \varepsilon$ with probability at least $1 \quad \delta$, using $\widetilde{O}\left(H^3 XA/\varepsilon^2\right)$ samples from the generative model.

5 Proofs of the Main Results

Before the proof, we introduce some notations. A table of notations is provided in Appendix A.

Notation. denotes an indefinite constant that changes throughout the proof and is independent of H, X, A, ε, and δ. We let $A_{\gamma,k} := \sum_{j=0}^{k} {}^{1} \gamma^{k \; j} \alpha^j$ and $A_k := \sum_{j=0}^{k} {}^{1} \alpha^j$ for any non-negative integer k with $A_1 := 1/(1 \quad \alpha)$. $\mathbf{F}_{k,m}$ denotes the σ-algebra generated by random variables $fy_{j,n,x,a}j(j,n,x,a)$ 2 $[k \quad 2] \quad [M] \quad \mathbf{X} \quad \mathbf{A}g$ [$fy_{j,n,x,a}j(j,n,x,a)$ 2 $fk \quad 1g \quad [m \quad 1] \quad \mathbf{X} \quad \mathbf{A}g$. For any k 2 $f0g$ [$[K \quad 1]$ and v 2 $\mathbf{R}^X$, $\mathrm{Var}(v)$ and $\widehat{P}_k v$ denote the functions

$$\mathrm{Var}(v) : (x,a) \; 7! \; (Pv^2)(x,a) \quad (Pv)^2(x,a) \text{ and } \widehat{P}_k v : (x,a) \; 7! \; \sum_{m=1}^{M} v(y_{k,m,x,a})/M\,,$$

respectively. We often write $\sqrt{\mathrm{Var}(v)}$ as $\sigma(v)$. Furthermore, ε_k and E_k denote "error" functions

$$\varepsilon_k : (x,a) \; 7! \; \gamma \widehat{P}_k \; {}_1 v_k \; {}_1(x,a) \quad \gamma P v_k \; {}_1(x,a) \text{ and } E_k : (x,a) \; 7! \; \sum_{j=1}^{k} \alpha^{k \; j} \varepsilon_j(x,a)\,,$$

respectively. (Note that $\varepsilon_1 = E_1 = \mathbf{0}$ since $v_0 = \mathbf{0}$.) For a sequence of policies $(\pi_k)_{k2\mathbf{Z}}$, we let $T_j^i := T^{\pi_i} T^{\pi_i \; 1} \quad T^{\pi_{j+1}} T^{\pi_j}$ for $i \quad j$, and $T_j^i := I$ otherwise. We also let $P_j^i := P^{\pi_i} P^{\pi_i \; 1} \quad P^{\pi_{j+1}} P^{\pi_j}$ for $i \quad j$, and $P_j^i := I$ otherwise. As a special case with $\pi_k = \pi$ for all k, we let $P^i := (P^{\pi} \;)^i$. Finally, throughout the proof, ι_1 and ι_2 denotes $\log(8KXA/\delta)$ and $\log(16KXA/\delta)$, repspectively.

[4] The time step index t starts from 0.

5.1 Proof of Theorem 1 (Near-optimality of the Non-stationary Policy)

The first step of the proof is the error propagation analysis of MDVI given below. It differs from the one of Vieillard et al. (2020a) since ours upper-bounds $v - v^{\pi_k^0}$. It is proven in Appendix F.1

Lemma 1. *For any $k \in [K]$, $\mathbf{0} \le v - v^{\pi_k^0} \le \Gamma_k$, where*

$$\Gamma_k := \frac{1}{A_1} \sum_{j=0}^{k-1} \gamma^j \left(\pi_k P_k^{k-1-j} - \pi P^j \right) E_{k-1-j} + 2H \left(\alpha^k + \frac{A_{\gamma,k}}{A_1} \right) \mathbf{1} .$$

From this result, it can be seen that an upper bound for each E_k is necessary. The following lemma provides an upper bound, which readily lead to Lemma 3 when combined with Lemma 1. These lemmas are proven in Appendix F.2.

Lemma 2. *Let E_1 be the event that $\|E_k\|_1 < 3H\sqrt{A_1 \iota_1 / M}$ for all $k \in [K]$. Then, $\mathbb{P}(E_1^c) \le \delta/4$.*

Lemma 3. *Assume that $\varepsilon \in (0, 1]$. When MDVI is run with the settings α, K, and M in Theorem 1, under the event E_1, its output policies $(\pi_k)_{k=0}^K$ satisfy that $\|v - v^{\pi_k^0}\|_1 \le 2(k+H)\gamma^k + \varepsilon\sqrt{H/c_2}$ for all $k \in [K]$. Furthermore, $\|v - v^{\pi_K^0}\|_1 \le \sqrt{H}\varepsilon$ for some $c_1, c_2 \ge 1$.*

Unfortunately, Lemma 3 is insufficient to show the minimax optimality of MDVI since it only holds that $\|v - v^{\pi_K^0}\|_1 \le \sqrt{H}\varepsilon$ while $\mathbb{P}(E_1) \ge 1 - \delta$. Any other setting of α, β, K, and M does not seem to lead to $\|v - v^{\pi_K^0}\|_1 \le \varepsilon$. Nonetheless, Lemma 3 turns out to be useful later to obtain a refined result.

To show the minimax optimality, we need to remove the extra $\sqrt{H}$ factor. The standard tools for this purpose are a Bernstein-type inequality and the total variance (TV) technique (Azar et al., 2013), which leverages the fact that $\|(I - \gamma P^\pi)^{-1}\sigma(v^\pi)\|_1 \le \sqrt{2H^3}$ for any policy π. In our case, the TV technique for a non-stationary policy is required due to $\pi_k P_k^{k-1-j}$, though.

Recall the definition of ε_k and note that its standard deviation consists of $\sigma(v_{k-1})$. As we use a Bernstein inequality for martingale because of E_k, we derive an upper bound for the sum of $\sigma(v_{j-1})^2$ over $j \in [k]$ (V in Lemma 19) using the fact that $\sigma(v_{j-1}) \le \sigma(v^*)$ when $v_{j-1} \le v^*$. To this end, the following lemma, proven in Appendix F.3, is useful.

Lemma 4. *For any $k \in [K]$,*

$$2\gamma^k H\mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_k^{k-1-j} \varepsilon_{k-1-j} \le v - v_k \le \Gamma_{k-1} + 2H\gamma^k \mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_k^{k-2-j} \varepsilon_{k-1-j} .$$

Combining this lemma with Lemma 2 and the following one, we can obtain an upper-bound for $\sigma(v_{k-1})$. The proofs of both results are given in Appendix F.4.

Lemma 5. *Let E_2 be the event that $\|\varepsilon_k\|_1 < 3H\sqrt{\iota_1 / M}$ for all $k \in [K]$. Then, $\mathbb{P}(E_2^c) \le \delta/4$.*

Lemma 6. *Conditioned on the event $E_1 \setminus E_2$, it holds for any $k \in [K]$ that*

$$\sigma(v_k) \le 2H \min\left\{ 1, 2\max\{\alpha, \gamma\}^{k-1} + \frac{A_{\gamma,k-1}}{A_1} + 6H\sqrt{\frac{\iota_1}{M}} \right\} \mathbf{1} + \sigma(v^*) . \tag{3}$$

Furthermore, $\sigma(v_0) = \mathbf{0}$.

Using Lemma 6, we can prove refined bounds for E_k and ε_k, as in Appendix F.5.

Lemma 7. *Let E_3 be the event that*

$$|E_k|(x,a) < \frac{4H\iota_2}{3M} + \sqrt{2V_k(x,a)\iota_2} \quad \text{for all } (x,a,k) \in \mathbf{X} \times \mathbf{A} \times [K] ,$$

where $V_k \coloneqq 4\sum_{j=1}^{k} \alpha^{2(k-j)}\overline{\mathrm{Var}}_j/M$ *with*

$$\overline{\mathrm{Var}}_j \coloneqq \mathrm{Var}(v_{\cdot j}) + 4H^2\left(4\max\{\alpha,\gamma\}^{2j-2} + \frac{A_{\gamma,j}^2}{A_1^2} + \frac{36H^2\iota_1}{M}\right)\mathbf{1}$$

for $k \geq 2$ *and* $\overline{\mathrm{Var}}_1 \coloneqq \mathbf{0}$. *Then,* $\mathsf{P}(E_3^c \mid E_1 \setminus E_2) \geq 1-\delta/4$.

Lemma 8. *Let* E_4 *be the event that*

$$|\varepsilon_k|(x,a) < \frac{4H\iota_2}{3M} + \sqrt{2W_k(x,a)\iota_2} \text{ for all } (x,a,k) \in \mathbf{X} \times \mathbf{A} \times [K]$$

where $W_k \coloneqq 4\overline{\mathrm{Var}}_k/M$. *Then,* $\mathsf{P}(E_4^c \mid E_1 \setminus E_2) \geq 1-\delta/4$.

With these lemmas, we are ready to prove Theorem 1.

Proof of Theorem 1. We condition the proof by $E_1 \setminus E_2 \setminus E_3$. As for any events A and B, $\mathsf{P}(A \setminus B) = \mathsf{P}((A \cap B^c) \setminus B) \geq 1 - \mathsf{P}(A^c \setminus B) \geq \mathsf{P}(B^c)$, and $\mathsf{P}(A^c \setminus B) = \mathsf{P}(A^c|B)\mathsf{P}(B) \geq \mathsf{P}(A^c|B)$,

$$\begin{aligned}\mathsf{P}(E_1 \setminus E_2 \setminus E_3) &\geq 1 - \mathsf{P}(E_3^c|E_1 \setminus E_2) - \mathsf{P}((E_1 \setminus E_2)^c)\\ &\geq 1 - \mathsf{P}(E_3^c|E_1 \setminus E_2) - \mathsf{P}(E_1^c) - \mathsf{P}(E_2^c)\,.\end{aligned}$$

Therefore, from Lemmas 2, 5, and 7, we conclude that $\mathsf{P}(E_1 \setminus E_2 \setminus E_3) \geq 1 - 3\delta/4$. Accordingly, any claim proven under $E_1 \setminus E_2 \setminus E_3$ holds with probability at least $1 - 3\delta/4$.

From Lemma 1, the setting that $\alpha = \gamma$, and the monotonicity of stochastic matrices,

$$v - v^{\pi_K^0} \leq \underbrace{\frac{1}{H}\sum_{k=0}^{K-1}\gamma^k\pi\,|P^k - \hat{P}^k|E_K\,|_{\hat{k}}}_{\sim} + \underbrace{\frac{1}{H}\sum_{k=0}^{K-1}\gamma^k\pi_K|P_K^{K-k} - \hat{P}_K^{K-k}|E_K\,|_{\hat{k}}}_{|} + 2(H+K)\gamma^K\mathbf{1}\,.$$

As the last term is less than ε/c_1 from Lemma 15, it remains to upper-bound $\sim$ and $|$. We note that $A_1 = H$ and $A_{\gamma,k} = k\gamma^k$ under the considered setting of α.

From the settings of α and M,

$$2V_k\iota_2 \leq \frac{\mathrm{Var}(v_{\cdot j})\varepsilon^2}{c_2 H} + \frac{\varepsilon^2}{c_2}\Big(k\gamma^{2(k-2)} + \frac{\gamma^{2(k-2)}}{H^2}\underbrace{\sum_{j=2}^{k}(j-2)^2}_{\leq k^3 \text{ from (a)}} + \frac{\varepsilon^2}{c_2}\underbrace{\sum_{j=2}^{k}\gamma^{2(k-j)}}_{\leq H}\Big)\mathbf{1}\,,$$

where (a) follows from Lemma 16. From this result and Lemma 11, it follows that

$$\sim \leq \frac{\varepsilon^2}{c_2}\mathbf{1} + \sqrt{\frac{\varepsilon}{c_2 H}}\underbrace{\sum_{k=0}^{K-1}\gamma^k\pi\,|P^k\sigma(v_{\cdot j})|}_{\leq\sqrt{2H^3}\mathbf{1} \text{ from Lemma 22}} + \sqrt{\frac{\varepsilon}{c_2}}\Big(\gamma^{K-2}\underbrace{\sum_{k=1}^{K}\left(\sqrt{p_k} + \sqrt{\frac{k}{H}}\right)}_{\leq (K^{2.5}/H) \text{ from (a)}} + H\sqrt{H\varepsilon}\Big)\mathbf{1}\,,$$

where (a) follows from Lemma 16 and that $H \leq K$. From Lemma 15, $K^{2.5}\gamma^{K-2}/H \leq \varepsilon/c_1$. Therefore, using the inequality $\sqrt{\varepsilon} \leq 1/\sqrt{H} \leq 1$, $H^{-1} \leq \sqrt{\varepsilon}$, $\sim \leq \left(c_2^{-1} + c_2^{-0.5}\right)\sqrt{\varepsilon}\mathbf{1}$.

Although an upper bound for $|$ can be similarly derived, a care must be taken when upper-bounding $\}$ $\coloneqq \sum_{k=0}^{K-1}\gamma^k\pi_K|P_K^{K-k} - \hat{P}_K^{K-k}|\sigma(v_{\cdot j})$. From Lemma 21, for any $k \in [K]$,

$$\sigma(v_{\cdot j}) \leq \sigma(v - v^{\pi_k^0}) + \sigma(v^{\pi_k^0}) \leq 2(k+H)\gamma^k\mathbf{1} + \sqrt{H/c_2}\sqrt{\varepsilon}\mathbf{1} + \sigma(v^{\pi_k^0})\,,$$

where the second inequality follows from Lemmas 3 and 20. Accordingly,

$$\} \quad 2\gamma^K \underbrace{\sum_{k=0}^{K} \frac{1}{(k+H)}}_{K^2 \text{ from (a)}} \mathbf{1} + \quad H\sqrt{H/c_2}\varepsilon\mathbf{1} + \underbrace{\sum_{k=0}^{K} \frac{1}{\gamma^k \pi_K P_K^K \; {}_k^1 \sigma(v^{\pi_K^0} \; {}_k)}}_{\sqrt{2H^3}\mathbf{1} \text{ from Lemma 22}} \qquad H \sqrt{H}\overline{H},$$

where (a) follows from Lemma 16 and that $H \quad K$, and the second inequality follows since $\varepsilon \quad 1/\sqrt{H} \quad 1$ and $K^2\gamma^K \qquad \varepsilon/c_1$ from Lemma 15. Thus, $H \;\; {}^1 \mid \qquad (c_2{}^1 + c_2{}^{0.5})\varepsilon\mathbf{1}$.

Combining these results, we conclude that there are constants c_1 and c_2 that satisfy the claim. □

5.2 Proof of Theorem 2 (Near-optimality of the Last Policy)

We need the following error propagation result. Its proof is given in Appendix G.1.

Lemma 9 (Error Propagation of MDVI). *For any k 2 $[K]$,*

$$\mathbf{0} \quad v \quad v^{\pi_k} \quad 2H\left(\alpha^k + \frac{A_{\gamma,k}}{A_1}\right)\mathbf{1} + \frac{1}{A_1}\left(N^{\pi_k}\pi_k \quad N^{\pi} \;\; \pi \;\;\right)E_k$$
$$+ \frac{1}{A_1}\sum_{j=1}^{k}\gamma^j\left(N^{\pi} \;\; \pi \;\; P_{k+1}^k \;\; {}_j \quad N^{\pi_k}\pi_k P_k^k \;\; {}_j^1\right)E_{k+1}^0 \;\; {}_j,$$

where $N^{\pi} \coloneqq \sum_{t=0}^{1}(\gamma\pi P)^t$ for any policy π, and $E_{k+1}^0 \;\; {}_j \coloneqq \varepsilon_{k+1} \;\; {}_j \quad (1 \quad \alpha)E_k \;\; {}_j$.

The following lemma is an analogue of Lemma 3. It is proven in Appendix G.2.

Lemma 10. *Assume that ε 2 $(0,1]$. When MDVI is run with the settings α, K, and M in Theorem 2, under the event $E_1 \setminus E_2$, its output policies $(\pi_k)_{k=0}^K$ satisfy that $kv \quad v^{\pi_k^0}k_1 \qquad H\alpha^k + \quad \varepsilon\sqrt{H/c_4}$ and $kv \quad v^{\pi_k}k_1 \qquad H\alpha^k + \quad \varepsilon\sqrt{H/c_4}$ for all k 2 $[K]$.*

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. We condition the proof by $E_1 \setminus E_2 \setminus E_3 \setminus E_4$. Since for any events A and B, $\mathsf{P}(A \setminus B) = \mathsf{P}((A \;[\; B^c) \setminus B) \quad 1 \quad \mathsf{P}(A^c \setminus B) \quad \mathsf{P}(B^c)$, and $\mathsf{P}(A^c \setminus B) = \mathsf{P}(A^c|B)\mathsf{P}(B) \quad \mathsf{P}(A^c|B)$,

$$\begin{aligned}\mathsf{P}(E_1 \setminus E_2 \setminus E_3 \setminus E_4) \quad & 1 \quad \mathsf{P}((E_3 \setminus E_4)^c|E_1 \setminus E_2) \quad \mathsf{P}((E_1 \setminus E_2)^c) \\ & 1 \quad \mathsf{P}(E_3^c \;[\; E_4^c|E_1 \setminus E_2) \quad \mathsf{P}(E_1^c) \quad \mathsf{P}(E_2^c) \\ & 1 \quad \mathsf{P}(E_3^c|E_1 \setminus E_2) \quad \mathsf{P}(E_4^c|E_1 \setminus E_2) \quad \mathsf{P}(E_1^c) \quad \mathsf{P}(E_2^c).\end{aligned}$$

Therefore, from Lemmas 2, 5, 7, and 8, we conclude that $\mathsf{P}(E_1 \setminus E_2 \setminus E_3 \setminus E_4) \quad 1 \quad \delta$. Accordingly, any claim proven under $E_1 \setminus E_2 \setminus E_3 \setminus E_4$ holds with probability at least $1 \quad \delta$.

From Lemma 9, the setting that $\alpha = 1 \quad (1 \quad \gamma)^2$, and the monotonicity of stochastic matrices,

$$v \quad v^{\pi_K} \quad 2H\left(\alpha^K + \frac{2A_{\gamma,K}}{H}\right)\mathbf{1} + \frac{1}{H^2}\underbrace{\left(N^{\pi_K}\pi_K + N^{\pi} \;\; \pi \;\;\right)|E_K|}_{\coloneqq \sim}$$
$$+ \underbrace{\frac{1}{H^2}\sum_{k=1}^{K}\gamma^k\left(N^{\pi} \;\; \pi \;\; P_{K+1}^K \;\; {}_k + N^{\pi_K}\pi_K P_K^K \;\; {}_k^1\right)\left(|\varepsilon_{K+1} \;\; {}_k| + \frac{1}{H^2}|E_K \;\; {}_k|\right)}_{\coloneqq |},$$

where $E_0 \coloneqq \mathbf{0}$. The first term can be bounded by $\quad \alpha^K H \qquad \varepsilon/c_3$ from Lemmas 13 and 15. In the sequel, we derive upper bounds for $\sim$ and $|$. We note that $A_1 = H^2$ and $A_{\gamma,k} \quad \alpha^k H$.

Next, we derive an upper bound for $\sim$. From the settings of $\alpha(\quad\gamma)$ and M,

$$2V_k\iota_2 \quad \frac{H\mathrm{Var}(v\ )\varepsilon^2}{c_4} + \frac{H\varepsilon^2}{c_4}\left(k\alpha^{2(k\ \ 2)} + \frac{H^3\varepsilon^2}{c_4}\right)\mathbf{1}\,.$$

From this result and Lemma 11, it follows that

$$\frac{\sim}{H^2} \quad \frac{\varepsilon^2}{c_4 H} + \frac{\varepsilon}{H\not{p}\overline{c_4 H}}\left(N^{\pi_K}\pi_K + N^{\pi}\ \pi\ \right)\sigma(v\ ) + \not{p}\frac{\varepsilon}{\overline{c_4}}\left(\underbrace{\sqrt{K/H}\alpha^{K\ \ 2}}_{\varepsilon/c_3 \text{ from (a)}} + \underbrace{H\varepsilon/\not{p}\overline{c_4}}_{1/\not{p}\overline{c_4} \text{ from (b)}}\right)\mathbf{1}\,,$$

where (a) follows from Lemma 15, and (b) follows by the assumption that $\varepsilon \quad 1/H$. By Lemma 22, $N^{\pi}\ \pi\ \sigma(v\ ) \quad \not{p}\overline{H^3}$. Furthermore, from Lemmas 10 and 20,

$$N^{\pi_K}\pi_K\sigma(v\ ) \quad \underbrace{H^2\alpha^K}_{H\not{p}\overline{H} \text{ from (a)}} + \varepsilon H\sqrt{H/c_4} + \underbrace{N^{\pi_K}\pi_K\sigma(v^{\pi_K})}_{H\not{p}\overline{H} \text{ from Lemma 22}} \quad H\not{p}\overline{H}\mathbf{1}\,,$$

where (a) follows from Lemma 15, and the last inequality follows since $\varepsilon \quad 1$. Consequently, $H^{\ 2}\sim (1/c_4 + 1/\not{p}\overline{c_4})\,\varepsilon\mathbf{1}$.

As for an upper bound for $|$, we derive upper bounds for the following two components:

$$\} := \frac{1}{H^2}\sum_{k=1}^{K-1}\gamma^k N^{\pi}\ \pi\ P^K_{K+1\ \ k}jE_{K\ \ k}j \text{ and } \bullet := \sum_{k=1}^{K}\gamma^k N^{\pi}\ \pi\ P^K_{K+1\ \ k}j\varepsilon_{K+1\ \ k}j\,.$$

Upper bounds for $H^{\ 2}\sum_{k=1}^{K}\gamma^k N^{\pi_K}\pi_K P^K_{K\ \ k}jE_{K\ \ k}j$ and $\sum_{k=1}^{K}\gamma^k N^{\pi_K}\pi_K P^K_{K\ \ k}j\varepsilon_{K+1\ \ k}j$ can be similarly derived.

From Lemma 2, $\} \quad \max_{k2[K]}kE_jk_1\mathbf{1} \quad \varepsilon\sqrt{H^3/c_4}$, and thus, $H^{\ 2}\} \quad \varepsilon/\not{p}\overline{c_4}$. On the other hand, from the assumption that $\gamma \quad \alpha$,

$$2W_k\iota_2 \quad \frac{\varepsilon^2}{c_4 H}\mathrm{Var}(v\ ) + \frac{H\varepsilon^2}{c_4}\left(\alpha^{2(k\ \ 2)} + \frac{\varepsilon^2 H}{c_4}\right)\mathbf{1}$$

for $k > 1$. Using Lemmas 8 and 11 as well as $\gamma \quad \alpha$,

$$\begin{aligned}
\bullet \quad & \varepsilon N^{\pi}\ \pi\ \sum_{k=1}^{K}\gamma^k P^K_{K+1\ \ k}\left(\frac{\varepsilon}{c_4}\mathbf{1} + \not{p}\frac{\sigma(v\ )}{\overline{c_4 H}} + \sqrt{\frac{H}{c_4}}\left(\alpha^{K\ \ k\ \ 2} + \varepsilon\sqrt{\frac{H}{c_4}}\right)\mathbf{1}\right) \\
& \varepsilon\left(\frac{H^2\varepsilon}{c_4}\mathbf{1} + N^{\pi}\ \pi\ \sum_{k=1}^{K}\gamma^k P^K_{K+1\ \ k}\not{p}\frac{\sigma(v\ )}{\overline{c_4 H}} + \sqrt{\frac{H^3}{c_4}}\left(\underbrace{K\alpha^{K\ \ 2}}_{\varepsilon/c_3 \text{ from (a)}} + H\varepsilon\sqrt{\frac{H}{c_4}}\right)\mathbf{1}\right) \\
& \underbrace{\varepsilon\left(\frac{H^2\varepsilon}{c_4}\mathbf{1} + \sqrt{\frac{H^3}{c_4}}\left(\frac{\varepsilon}{c_3} + H\varepsilon\sqrt{\frac{H}{c_4}}\right)\mathbf{1}\right)}_{H^2/\not{p}\overline{c_4} \text{ as } \varepsilon\ \ 1/H} + \not{p}\frac{\varepsilon}{\overline{c_4 H}}N^{\pi}\ \pi\ \sum_{k=1}^{K}\gamma^k P^K_{K+1\ \ k}\sigma(v\ )\,.
\end{aligned}$$

Now, it remains to upper-bound $\sum_{k=1}^{K}\gamma^k P^K_{K+1\ \ k}\sigma(v\ )$. From Lemma 21,

$$\sigma(v\ ) \quad \sigma(v \quad v^{\pi^0_k}) + \sigma(v^{\pi^0_k}) \quad \alpha^k H\mathbf{1} + \varepsilon\sqrt{H/c_4}\mathbf{1} + \sigma(v^{\pi^0_k})$$

for any $k \in [K]$, where Lemmas 10 and 20 are used. Consequently,

$$\sum_{k=1}^{K} \gamma^k P_{K+1}^K \; {}_k\sigma(v \;) \qquad \sum_{k=1}^{K} \gamma^k P_{K+1}^K \; {}_k \left(H\alpha^{K+1} \; {}^k\mathbf{1} + \varepsilon\sqrt{H/c_4}\mathbf{1} + \sigma(v^{\pi^0_{K+1}} \; {}^k) \right)$$
$$\left(\underbrace{HK\alpha^{K+1}}_{\varepsilon/c_3}\mathbf{1} + \varepsilon\sqrt{H^3/c_4}\mathbf{1} + \underbrace{\sum_{k=1}^{K} \gamma^k P_{K+1}^K \; {}_k\sigma(v^{\pi^0_{K+1}} \; {}^k)}_{\sqrt{H^3}\mathbf{1}} \right),$$

where the second inequality follows since $\gamma \quad \alpha$. Consequently, $H \; {}^2 \bullet \quad \varepsilon / \sqrt{c_4}$.

Combining these inequalities, we deduce that $v \quad v^{\pi_K} \quad \varepsilon \left(c_3 \; {}^1 + c_4 \; {}^{0.5} \right) \mathbf{1}$. □

6 Empirical illustration

We compare MDVI to a synchronous version of Q-learning (e.g., Even-Dar et al. (2003)) in a simple setting on a class of random MDPs called Garnets (Archibald et al., 1995), with $\gamma = 0.9$. Figure 1 shows the sample complexity of MDVI as a function of ε. We run MDVI on 100 random MDPs, and, given ε, we report the number of samples KM MDVI uses to find ε-optimal policy. We compare this empirical sample complexity with the one of Q-LEARNING, which has a tight quadratic dependency to the horizon (Li et al., 2021a) – compared to the cubic one of MDVI (Theorem 2). Figure 1 shows the difference in sample complexity between the two methods: especially for low ε, MDVI reaches an ε-optimal policy with much fewer samples, up to $H = 10$ times less samples for $\varepsilon = 10 \; {}^3$. Complete details, pseudocodes, and results with other α are provided in Appendix H.

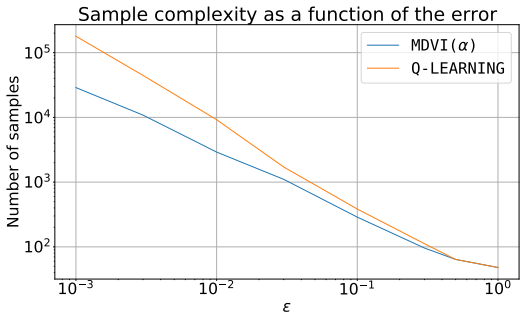

Figure 1: Sample complexities of MDVI with $\alpha = 1$ and Q-LEARNING (synchronous version of Q-learning) on Garnets. MDVI is run in the stationary policy setting. Both algorithms use $M = 1$. As noted in Section 5, MDVI with $\alpha = 1$ is also nearly minimax-optimal.

7 Conclusion

In this work, we considered and analyzed the sample complexity of a model-free algorithm called MDVI (Geist et al., 2019; Vieillard et al., 2020a) under the generative model setting. We showed that it is nearly minimax-optimal for finding an ε-optimal policy despite its simplicity compared to previous model-free algorithms (Sidford et al., 2018; Wainwright, 2019; Khamaru et al., 2021). We believe that our results are significant for the following three reasons.

First, we demonstrate the effectiveness of KL and entropy regularization. Second, as discussed by Vieillard et al. (2020a), MDVI encompasses various algorithms as special cases or equivalent forms, and our results provide theoretical guarantees for most of them at once. Third, MDVI uses no variance-reduction technique, which leads to multi-epoch algorithms and involved analyses (Sidford et al., 2018; Wainwright, 2019; Khamaru et al., 2021). As such, our analysis is straightforward, and it would be easy to extend it to more complex settings.

A disadvantage of MDVI is that its range of valid ε is limited compared to previous algorithms (Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020). It is unclear if this is an artifact of our analysis or the real limitation of MDVI-type algorithms. We leave this topic as a future work.

References

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal. In *Conference on Learning Theory*, 2020.

TW Archibald, KIM McKinnon, and LC Thomas. On the Generation of Markov Decision Processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Mohammad Azar, Mohammad Ghavamzadeh, Hilbert Kappen, and Rémi Munos. Speedy Q-Learning. In *Advances in Neural Information Processing Systems*, 2011.

Mohammad 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, Jun 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19 (3):357 – 367, 1967.

Sergei Natanovich Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, 1946.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning Rates for Q-learning. *Journal of Machine Learning Research*, 5(1), 2003.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the Noise in Reinforcement Learning via Soft Updates. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes. In *International Conference on Machine Learning*, 2019.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-Learning Provably Efficient? In *Advances in Neural Information Processing Systems*, 2018.

Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

Tadashi Kozuno, Eiji Uchibe, and Kenji Doya. Theoretical Analysis of Efficiency and Robustness of Softmax and Gap-Increasing Operators in Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pp. 1–48, 2022.

Tor Lattimore and Marcus Hutter. PAC Bounds for Discounted MDPs. In *International Conference on Algorithmic Learning Theory*, 2012.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 1st edition, 2020.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse Markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

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. In *Advances in neural information processing systems*, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is Q-Learning Minimax Optimal? A Tight Sample Complexity Analysis. *arXiv preprint arXiv:2102.06548*, 2021a.

Xiang Li, Wenhao Yang, Zhihua Zhang, and Michael I Jordan. Polyak-Ruppert Averaged Q-Leaning is Statistically Efficient. *arXiv preprint arXiv:2112.14582*, 2021b.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *International Conference on Machine Learning*, 2020.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Bruno Scherrer and Boris Lesner. On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 2012.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Advances in Neural Information Processing Systems*, 2018.

Peter Vamplew, Richard Dazeley, and Cameron Foale. Softmax exploration strategies for multiobjective reinforcement learning. *Neurocomputing*, 263:74–86, 2017.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, and Matthieu Geist. Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020a.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020b.

Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

Christopher J. C. H. Watkins and Peter Dayan. Q-Learning. *Machine Learning*, 8(3):279–292, 1992.

Wenhao Yang, Xiang Li, and Zhihua Zhang. A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *Conference on Learning Theory*, 2021.

Appendix

Table of Contents

A Notations

Table 1: Table of Notations

Notation	Meaning
$\mathbf{A}$	action space of size A
H	effective horizon $H \coloneqq 1/(1 \quad \gamma)$
P	transition matrix
$\mathbf{X}$	state space of size X
r	reward vector bounded by 1
γ	discount factor in $[0, 1)$
ε	admissible suboptimality
δ	admissible failure probability
E_k	$E_k : (x, a) \mapsto \sum_{j=1}^{k} \alpha^{k \quad j} \varepsilon_j(x, a)$
ε_k	$\varepsilon_k : (x, a) \mapsto \gamma \widehat{P}_{k \quad 1} v_{k \quad 1}(x, a) \quad \gamma P v_{k \quad 1}(x, a)$
$A_k, A_{1}, A_{\gamma,k}$	$\sum_{j=0}^{k \quad 1} \alpha^j$, $\sum_{j=0}^{1} \alpha^j$, $\sum_{j=0}^{k \quad 1} \alpha^j \gamma^{k \quad j}$
E_1	event of small E_k for all k (not variance-aware)
E_2	event of small ε_k for all k (not variance-aware)
E_3	event of small E_k for all k (variance-aware)
E_4	event of small ε_k for all k (variance-aware)
$\mathbf{F}_{k,m}$	σ-algebra in the filtration (cf. Section 5)
K	number of value updates
M	number of samples per each value update
P^π	$P^\pi \coloneqq P\pi$
P_j^i, P^i	$P_j^i \coloneqq P^{\pi_i} P^{\pi_{i \quad 1}} \quad P^{\pi_{j+1}} P^{\pi_j}$, $P^i \coloneqq (P^\pi)^i$
T^π, T_j^i	Bellman operator for a policy π, $T_j^i \coloneqq T^{\pi_i} T^{\pi_{i \quad 1}} \quad T^{\pi_{j+1}} T^{\pi_j}$
V_k	an upper bound for E_k's predictive quadratic variance (cf. Lemma 7)
W_k	an upper bound for ε_k's predictive quadratic variance (cf. Lemma 8)
s_k	$s_k \coloneqq q_k + \alpha s_{k \quad 1}$ (cf. MDVI)
v_k	$v_k \coloneqq w_k \quad \alpha w_{k \quad 1}$ (cf. MDVI)
w_k	$w_k(x) \coloneqq \max_{a \in \mathbf{A}} s_k(x, a)$ (cf. MDVI)
α	$\alpha \coloneqq \tau/(\tau + \kappa)$, weight for s_k updates (cf. MDVI and Appendix B)
β	$\beta \coloneqq 1/(\tau + \kappa)$, inverse temperature for π_k (cf. Section 4 and Appendix B)
ι_1, ι_2	$\iota_1 \coloneqq \log(8KXA/\delta)$, $\iota_2 \coloneqq \log(16KXA/\delta)$
π_k^0	a non-stationary policy that follows $\pi_k, \pi_{k \quad 1}, \ldots$ sequentially (cf. Section 5)
	an indefinite constant independent of H, X, A, ε, and δ

B Equivalence of MDVI Update Rules

We show the equivalence of MDVI's updates (1) and (2) to those used in MDVI. We first recall MDVI's updates (1) and (2):

$$q_{k+1} = r + \gamma P^{\pi_k} \left(q_k \quad \tau \log \frac{\pi_k}{\pi_{k \quad 1}} \quad \kappa \log \pi_k \right) + \varepsilon_k \,,$$

$$\text{where } \pi_k(\cdot|x) = \arg\max_{p \in \Delta(\mathbf{A})} \sum_{a \in \mathbf{A}} p(a) \left(q_k(s, a) \quad \tau \log \frac{p(a)}{\pi_{k \quad 1}(a|x)} \quad \kappa \log p(a) \right) \text{ for all } x \in \mathbf{X} \,,$$

The policy update (2) can be rewritten as follows (e.g., Equation (5) of Kozuno et al. (2019)):

$$\pi_k(a|x) = \frac{\pi_{k-1}(a|x)^\alpha \exp(\beta q_k(x,a))}{\sum_{b \in \mathbf{A}} \pi_{k-1}(b|x)^\alpha \exp(\beta q_k(x,b))},$$

where $\alpha := \tau/(\tau+\kappa)$, and $\beta := 1/(\tau+\kappa)$. It can be further rewritten as, defining $s_k = q_k + \alpha s_{k-1}$

$$\pi_k(a|x) = \frac{\exp(\beta s_k(x,a))}{\sum_{b \in \mathbf{A}} \exp(\beta s_k(x,b))}.$$

Plugging in this policy expression to v_k, we deduce that

$$\begin{aligned} v_k(x) &= \frac{1}{\beta} \log \sum_{a \in \mathbf{A}} \exp(\beta q_k(x,a) + \alpha \log \pi_{k-1}(a|x)) \\ &= \frac{1}{\beta} \log \sum_{a \in \mathbf{A}} \exp(\beta s_k(x,a)) \quad \frac{\alpha}{\beta} \log \sum_{a \in \mathbf{A}} \exp(\beta s_{k-1}(x,a)) \,. \end{aligned}$$

Kozuno et al. (2019, Appendix B) show that when $\beta \to 1$, $v_k(x) = w_k(x) \quad \alpha w_{k-1}(x)$. Furthermore, the Boltzmann policy becomes a greedy policy. Accordingly, the update rules used in MDVI is a limit case of the original MDVI updates.

C Auxiliary Lemmas

In this appendix, we prove some auxiliary lemmas used in the proof.

Lemma 11. *For any positive real values a and b, $\sqrt{a+b} \quad \sqrt{a} + \sqrt{b}$.*

Proof. Indeed, $a+b \quad a + 2\sqrt{ab} + b = (\sqrt{a} + \sqrt{b})^2$. □

Lemma 12. *For any real values $(a_n)_{n=1}^N$, $(\sum_{n=1}^N a_n)^2 \quad N \sum_{n=1}^N a_n^2$.*

Proof. Indeed, from the Cauchy–Schwarz inequality,

$$\left(\sum_{n=1}^N a_n \cdot 1\right)^2 \quad \left(\sum_{n=1}^N 1\right)\left(\sum_{n=1}^N a_n^2\right) = N \sum_{n=1}^N a_n^2,$$

which is the desired result. □

Lemma 13. *For any $k \in [K]$,*

$$A_{\gamma,k} = \begin{cases} \gamma \dfrac{\alpha^k \quad \gamma^k}{\alpha \quad \gamma} & \text{if } \alpha \neq \gamma \\ k\gamma^k & \text{otherwise} \end{cases}.$$

Proof. Indeed, if $\alpha \neq \gamma$

$$A_{\gamma,k} = \sum_{j=0}^{k-1} \alpha^j \gamma^{k-j} = \gamma^k \frac{(\alpha/\gamma)^k \quad 1}{(\alpha/\gamma) \quad 1} = \gamma \frac{\alpha^k \quad \gamma^k}{\alpha \quad \gamma}.$$

If $\alpha = \gamma$, $A_{\gamma,k} = k\gamma^k$ by definition. □

Lemma 14. *For any real value $x \in (0,1]$, $1 \quad x \quad \log(1/x)$.*

Proof. Since $\log(1/x)$ is convex and differentiable, $\log(1/x)$ $\log(1/y)$ $(x$ $y)/y$. Choosing $y = 1$, we concludes the proof. □

Lemma 15. *Suppose* α, γ *2* $[0, 1)$, ε *2* $(0, 1]$, c *2* $[1,$ *1* $)$, m *2* $\mathbf{N}$, *and* n *2* $[0,$ *1* $)$. *Let* $K \coloneqq \frac{m}{1 \quad \alpha} \log \frac{cH}{\varepsilon}$. *Then,*

$$K^n \alpha^K \quad \left(\frac{mn}{(1 \quad \alpha)e}\right)^n \left(\frac{\varepsilon}{cH}\right)^{m \ 1}.$$

Proof. Using Lemma 14,

$$K = \frac{m}{1 \quad \alpha} \log \frac{cH}{\varepsilon} \quad \log_\alpha \left(\frac{\varepsilon}{cH}\right)^m.$$

Therefore,

$$K^n \alpha^K \quad \left(\frac{m}{1 \quad \alpha} \log \frac{cH}{\varepsilon}\right)^n \left(\frac{\varepsilon}{cH}\right)^m = \frac{m^n}{(1 \quad \alpha)^n} \left(\frac{\varepsilon}{cH}\right)^m \left(\log \frac{cH}{\varepsilon}\right)^n.$$

Since $x \left(\log \frac{1}{x}\right)^n \quad \left(\frac{n}{e}\right)^n$ for any x *2* $(0, 1]$ as shown later,

$$K^n \alpha^K \quad \left(\frac{mn}{(1 \quad \alpha)e}\right)^n \left(\frac{\varepsilon}{cH}\right)^{m \ 1}.$$

Now it remains to show $f(x) \coloneqq x \left(\log \frac{1}{x}\right)^n \quad \left(\frac{n}{e}\right)^n$ for $x < 1$. We have that

$$f^0(x) = (\quad \log x)^n \quad n(\quad \log x)^{n \ 1} =) \quad f^0(x) = 0 \text{ at } x = e^{\ n}.$$

Therefore, f takes its maximum $\left(\frac{n}{e}\right)^n$ at $e^{\ n}$ when x *2* $(0, 1)$. □

The following lemma is a special case of a well-known inequality that for any increasing function f

$$\sum_{k=1}^{K} f(k) \quad \int_1^{K+1} f(x)dx.$$

Lemma 16. *For any* K *2* $\mathbf{N}$ *and* n *2* $[0,$ *1* $)$, $\sum_{k=1}^{K} k^n \quad \frac{1}{n+1}(K+1)^{n+1}$.

D Tools from Probability Theory

We extensively use the following two concentration inequalities. The first one is Azuma-Hoeffding inequality (Azuma, 1967; Hoeffding, 1963; Boucheron et al., 2013), and the second one is Bernstein's inequality (Bernstein, 1946; Boucheron et al., 2013) for a martingale (Lattimore & Szepesvari, 2020, Excercises 5.14 (f)). For a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(F_n)_{n=1}^N$, we let $\mathsf{E}_n[X_n] \coloneqq \mathsf{E}[X_n | F_{n \ 1}]$ for $n \quad 1$, and $\mathsf{E}_1[X_1] \coloneqq \mathsf{E}[X_1]$.

Lemma 17 (Azuma-Hoeffding Inequality). *Consider a real-valued stochastic process* $(X_n)_{n=1}^N$ *adapted to a filtration* $(F_n)_{n=1}^N$. *Assume that* X_n *2* $[l_n, u_n]$ *and* $\mathsf{E}_n[X_n] = 0$ *almost surely, for all* n. *Then,*

$$\mathsf{P}\left(\sum_{n=1}^{N} X_n \quad \sqrt{\sum_{n=1}^{N} \frac{(u_n \quad l_n)^2}{2} \log \frac{1}{\delta}}\right) \quad \delta$$

for any δ *2* $(0, 1)$.

Lemma 18 (Bernstein's Inequality). *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(F_n)_{n=1}^N$. Suppose that $X_n \quad U$ and $\mathsf{E}_n[X_n] = 0$ almost surely, for all n. Then, letting $V^0 := \sum_{n=1}^N \mathsf{E}_n[X_n^2]$,*

$$\mathsf{P}\left(\sum_{n=1}^N X_n \quad \frac{2U}{3}\log\frac{1}{\delta} + \sqrt{2V\log\frac{1}{\delta}} \text{ and } V^0 \quad V\right) \quad \delta$$

for any V 2 $[0, 1)$ and δ 2 $(0,1)$.

In our analysis, we use the following corollary of this Bernstein's inequality.

Lemma 19 (Conditional Bernstein's Inequality). *Consider the same notations and assumptions in Lemma 18. Furthermore, let E be an event that implies $V^0 \quad V$ for some V 2 $[0, 1)$ with $\mathsf{P}(E) \quad 1 \quad \delta^0$ for some δ^0 2 $(0,1)$. Then,*

$$\mathsf{P}\left(\sum_{n=1}^N X_n \quad \frac{2U}{3}\log\frac{1}{\delta(1 \quad \delta^0)} + \sqrt{2V\log\frac{1}{\delta(1 \quad \delta^0)}}\middle| E\right) \quad \delta$$

for any δ 2 $(0,1)$.

Proof. Let A and B denote the events of

$$\sum_{n=1}^N X_n \quad \frac{2U}{3}\log\frac{1}{\delta(1 \quad \delta^0)} + \sqrt{2V\log\frac{1}{\delta(1 \quad \delta^0)}}$$

and $V^0 \quad V$, respectively. Since $E \quad B$, it follows that $A \setminus E \quad A \setminus B$, and $\mathsf{P}(A \setminus E) \quad \mathsf{P}(A \setminus B)$. Accordingly,

$$\mathsf{P}(A|E) = \frac{\mathsf{P}(A \setminus E)}{\mathsf{P}(E)} \quad \frac{\mathsf{P}(A \setminus B)}{\mathsf{P}(E)} \overset{(a)}{} \frac{\delta(1 \quad \delta^0)}{\mathsf{P}(E)} \overset{(b)}{} \delta\,,$$

where (a) follows from Lemma 18, and (b) follows from $\mathsf{P}(E) \quad 1 \quad \delta^0$. □

Lemma 20 (Popoviciu's Inequality for Variances). *The variance of any random variable bounded by x is bounded by x^2.*

E Total Variance Technique

The following lemma is due to Azar et al. (2013).

Lemma 21. *Suppose two real-valued random variables X, Y whose variances, $\mathsf{V}X$ and $\mathsf{V}Y$, exist and are finite. Then, $\sqrt{\mathsf{V}X} \quad \sqrt{\mathsf{V}[X \quad Y]} + \sqrt{\mathsf{V}Y}$.*

For completeness, we prove Lemma 21.

Proof. Indeed, from Cauchy-Schwartz inequality,

$$\begin{aligned}\mathsf{V}X &= \mathsf{V}[X \quad Y + Y] \\ &= \mathsf{V}[X \quad Y] + \mathsf{V}Y + 2\mathsf{E}\left[(X \quad Y \quad \mathsf{E}[X \quad Y])(Y \quad \mathsf{E}Y)\right] \\ &\quad \mathsf{V}[X \quad Y] + \mathsf{V}Y + 2\sqrt{\mathsf{V}[X \quad Y]\mathsf{V}Y} = \left(\sqrt{\mathsf{V}[X \quad Y]} + \sqrt{\mathsf{V}Y}\right)^2 .\end{aligned}$$

This is the desired result. □

The following lemma is an extension of Lemma 7 by Azar et al. (2013) and its refined version by Agarwal et al. (2020).

Lemma 22. *Suppose a sequence of deterministic policies* $(\pi_k)_{k=0}^K$ *and let*

$$q^{\pi_k^0} := \begin{cases} r + \gamma P v^{\pi_k^0 \ 1} & \text{for } k \ 2 \ [K] \\ q^{\pi_0} & \text{for } k = 0 \end{cases} .$$

Furthermore, let σ_k^2 *and* Σ_k^2 *be non-negative functions over* $\mathbf{X} \quad \mathbf{A}$ *defined by*

$$\sigma_k^2(x,a) := \begin{cases} P(v^{\pi_k^0 \ 1})^2(x,a) \quad (Pv^{\pi_k^0 \ 1})^2(x,a) & \text{for } k \ 2 \ [K] \\ P(v^{\pi_0})^2(x,a) \quad (Pv^{\pi_0})^2(x,a) & \text{for } k = 0 \end{cases}$$

and

$$\Sigma_k^2(x,a) := \mathsf{E}_k \left[\left(\sum_{t=0}^{1} \gamma^t r(X_t, A_t) \quad q^{\pi_k^0}(X_0, A_0) \right)^2 \middle| X_0 = x, A_0 = a \right]$$

for $k \ 2 \ f0g \ [\ [K]$*, where* E_k *is the expectation over* $(X_t, A_t)_{t=0}^{1}$ *wherein* $A_t \quad \pi_k \ _t(\ jX_t)$ *until* $t = k$*, and* $A_t \quad \pi_0(\ jX_t)$ *thereafter. Then,*

$$\sum_{j=0}^{k \ 1} \gamma^{j+1} P_k^{k \ 1}{}_j \sigma_k \ _j \qquad p \frac{}{2H^3}$$

for any $k \ 2 \ [K]$.

For its proof, we need the following lemma.

Lemma 23. *Suppose a sequence of deterministic policies* $(\pi_k)_{k=0}^K$ *and notations in Lemma 22. Then, for any* $k \ 2 \ [K]$*, we have that*

$$\Sigma_k^2 = \gamma^2 \sigma_k^2 + \gamma^2 P^{\pi_k \ 1} \Sigma_k^2 \ _1 .$$

Proof. Let $R_s^u := \sum_{t=s}^u \gamma^t \ ^s r(X_t, A_t)$ and $\mathsf{E}_k[\ jx, a] := \mathsf{E}_k[\ jX_0 = x, A_0 = a]$. We have that

$$\Sigma_k^2(x,a) = \mathsf{E}_k \left[\left(R_0^1 \quad q^{\pi_k^0}(X_0, A_0) \right)^2 \middle| x, a \right] := \mathsf{E}_k \left[(I_1 + \gamma I_2)^2 \middle| x, a \right] ,$$

where $I_1 := r(X_0, A_0) + \gamma q^{\pi_k^0 \ 1}(X_1, A_1) \quad q^{\pi_k^0}(X_0, A_0)$, and $I_2 := R_1^1 \quad q^{\pi_k^0 \ 1}(X_1, A_1)$. With these notations, we see that

$$\begin{aligned} \Sigma_k^2(x,a) &= \mathsf{E}_k \left[I_1^2 + \gamma^2 I_2^2 + 2\gamma I_1 I_2 \middle| x, a \right] \\ &= \mathsf{E}_k \left[I_1^2 + \gamma^2 I_2^2 + 2\gamma I_1 \mathsf{E}_k \ _1 [I_2 j X_1, A_1] \middle| x, a \right] \\ &= \mathsf{E}_k \left[I_1^2 \middle| x, a \right] + \gamma^2 \mathsf{E}_k \left[I_2^2 \middle| x, a \right] \\ &= \mathsf{E}_k \left[I_1^2 \middle| x, a \right] + \gamma^2 P^{\pi_k \ 1} \Sigma_k^2 \ _1(x,a) , \end{aligned}$$

where the second line follows from the law of total expectation, and the third line follows since $\mathsf{E}_k \ _1 [I_2 j X_1, A_1] = 0$ due to the Markov property. The first term in the last line is $\gamma^2 \sigma_k^2(x,a)$ because

$$\begin{aligned} \mathsf{E}_k \left[I_1^2 \middle| x, a \right] &\overset{(a)}{=} \gamma^2 \mathsf{E}_k \left[\Big(\underbrace{q^{\pi_k^0 \ 1}(X_1, A_1)}_{v^{\pi_k^0 \ 1}(X_1) \text{ from (b)}} \quad (Pv^{\pi_k^0 \ 1})(X_0, A_0) \Big)^2 \middle| x, a \right] \\ &= \gamma^2 \left(P \left(v^{\pi_k^0 \ 1} \right)^2 \right)(x,a) + \gamma^2 (Pv^{\pi_k^0 \ 1})^2(x,a) \quad 2(Pv^{\pi_k^0 \ 1})^2(x,a) \\ &= \gamma^2 \left(P \left(v^{\pi_k^0 \ 1} \right)^2 \right)(x,a) \quad \gamma^2 (Pv^{\pi_k^0 \ 1})^2(x,a) , \end{aligned}$$

where (a) follows from the definition that $q^{\pi_k^0} = r + \gamma P v^{\pi_k^0 \ 1}$, and (b) follows since the policies are deterministic. From this argument, it is clear that $\Sigma_k^2 = \gamma^2 \sigma_k^2 + \gamma^2 P^{\pi_k \ 1} \Sigma_k^2 \ _1$, which is the desired result. □

Now, we are ready to prove Lemma 22.

Proof of Lemma 22. Let $H_k \coloneqq \sum_{j=0}^{k-1} \gamma^j$. Using Jensen's inequality twice,

$$\sum_{j=0}^{k-1} \gamma^{j+1} P_k^{k-1-j} \sigma_{k-1-j} \quad \sum_{j=0}^{k-1} \gamma^{j+1} \sqrt{P_k^{k-1-j} \sigma_{k-1-j}^2}$$
$$\gamma H_k \sum_{j=0}^{k-1} \frac{\gamma^{j+1}}{H_k} \sqrt{P_k^{k-1-j} \sigma_{k-1-j}^2}$$
$$\sqrt{H_k \sum_{j=0}^{k-1} \gamma^{j+2} P_k^{k-1-j} \sigma_{k-1-j}^2} \quad \sqrt{H \sum_{j=0}^{k-1} \gamma^{j+2} P_k^{k-1-j} \sigma_{k-1-j}^2}.$$

From Lemma 23, we have that

$$\begin{aligned}
&\sum_{j=0}^{k-1} \gamma^{j+2} P_k^{k-1-j} \sigma_{k-1-j}^2 \\
&= \sum_{j=0}^{k-1} \gamma^j P_k^{k-1-j} \left(\Sigma_{k-j}^2 \quad \gamma^2 P^{\pi_{k-1} \cdots \pi_{k-j}} \Sigma_{k-1-j}^2 \right) \\
&= \sum_{j=0}^{k-1} \gamma^j P_k^{k-1-j} \left(\Sigma_{k-j}^2 \quad \gamma P^{\pi_{k-1} \cdots \pi_{k-j}} \Sigma_{k-1-j}^2 + \gamma(1 \quad \gamma) P^{\pi_{k-1} \cdots \pi_{k-j}} \Sigma_{k-1-j}^2 \right) \\
&= \sum_{j=0}^{k-1} \gamma^j P_k^{k-1-j} \Sigma_{k-j}^2 \quad \sum_{j=1}^{k} \gamma^j P_k^{k-j} \Sigma_{k-j}^2 + \gamma(1 \quad \gamma) \sum_{j=0}^{k-1} \gamma^j P_k^{k-1-j} \Sigma_{k-1-j}^2.
\end{aligned}$$

The final line is equal to $\Sigma_k^2 \quad \gamma^k P_0^k \Sigma_0^2 + \gamma(1 \quad \gamma) \sum_{j=0}^{k-1} \gamma^j P_k^{k-1-j} \Sigma_{k-1-j}^2$. Finally, from the monotonicity of stochastic matrices and that $\mathbf{0} \quad \Sigma_j^2 \quad H^2 \mathbf{1}$ for any j,

$$\sum_{j=0}^{k-1} \gamma^{j+1} P_k^{k-1-j} \sigma_{k-1-j} \quad p\overline{2H^3}.$$

This concludes the proof. □

F Proof of Lemmas for Theorem 1 (Bound for a Non-Stationary Policy)

Before starting the proof, we introduce some notations and facts frequently used in the proof.

Frequently Used Facts. We frequently use the following fact, which follows from definitions:

$$s_k = A_k r + \gamma P w_{k-1} + E_k \quad \text{for any } k \; 2 \; [K]. \tag{4}$$

Indeed, $s_k = \sum_{j=1}^{k} \alpha^{k-j} (r + \gamma P(w_{j-1} \quad \alpha w_{j-2}) + \varepsilon_j) = A_k r + \gamma P w_{k-1} + E_k$. In addition, we often mention the "monotonicity" of stochastic matrices: any stochastic matrix ρ satisfies that $\rho v \quad \rho u$ for any vectors v, u such that $v \quad u$. Examples of stochastic matrices in the proof are P, π, P^π, and πP. The monotonicity property is so frequently used that we do not always mention it.

F.1 Proof of Lemma 1 (Error Propagation Analysis)

Proof. Note that

$$\mathbf{0} \quad v \quad v^{\pi_k^0} = \frac{A_k}{A_1}\left(v \quad v^{\pi_k^0}\right) + \alpha^k\left(v \quad v^{\pi_k^0}\right) \quad \frac{A_k}{A_1}\left(v \quad v^{\pi_k^0}\right) + 2H\alpha^k\mathbf{1}$$

since $v \quad v^{\pi_k^0} \quad 2H\mathbf{1}$. Therefore, we need an upper bound for $A_k(v \quad v^{\pi_k^0})$. We decompose $A_k(v \quad v^{\pi_k^0})$ to $A_k v \quad w_k$ and $w_k \quad A_k v^{\pi_k^0}$. Then, we derive upper bounds for each of them (inequalities (5) and (6), respectively). The desired result is obtained by summing up those bounds.

Upper bound for $A_k v \quad w_k$. We prove by induction that for any $k \in [K]$,

$$A_k v \quad w_k \quad H A_{\gamma,k}\mathbf{1} \quad \sum_{j=0}^{k \quad 1} \gamma^j \pi \quad P^j E_{k \quad j}\,. \tag{5}$$

We have that

$$\begin{aligned} A_k v \quad w_k &\overset{(a)}{\quad} \pi \ (A_k q \quad s_k) \\ &\overset{(b)}{=} \pi \ (A_k q \quad A_k r \quad \gamma P w_{k \ 1} \quad E_k) \\ &\overset{(c)}{=} \pi \ (\gamma P(A_k v \quad w_{k \ 1}) \quad E_k) \\ &\overset{(d)}{\quad} \pi \ \left(\gamma P(A_{k \ 1} v \quad w_{k \ 1}) + \alpha^{k \ 1}\gamma H\mathbf{1} \quad E_k\right)\,, \end{aligned}$$

where (a) is due to the greediness of π_k, (b) is due to the equation (4), (c) is due to the Bellman equation for q , and (d) is due to the fact that $(A_k \quad A_{k \ 1})v \ = \alpha^{k \ 1} v \quad \alpha^{k \ 1} H\mathbf{1}$. From this result and the fact that $w_0 = \mathbf{0}$, $A_1 v \quad w_1 \quad \gamma H\mathbf{1} \quad \pi \ E_1$. Therefore, the inequality (5) holds for $k = 1$. From the step (d) above and induction, it is straightforward to verify that the inequality (5) holds for other k.

Upper bound for $w_k \quad A_k v^{\pi_k^0}$. We prove by induction that for any $k \in [K]$,

$$w_k \quad A_k v^{\pi_k^0} \quad H A_{\gamma,k}\mathbf{1} + \sum_{j=0}^{k \quad 1} \gamma^j \pi_k P_k^{k \ 1} {}_j^{1} E_{k \quad j}\,. \tag{6}$$

Recalling that $v^{\pi_k^0} = \pi_k T_0^{k \ 1} q^{\pi_0}$, we deduce that

$$\begin{aligned} w_k \quad A_k v^{\pi_k^0} &\overset{(a)}{=} \pi_k\left(s_k \quad A_k T_0^{k \ 1} q^{\pi_0}\right) \\ &\overset{(b)}{=} \pi_k\left(A_k r + \gamma P w_{k \ 1} \quad A_k T_1^{k \ 1} q^{\pi_0} + E_k\right) \\ &\overset{(c)}{=} \pi_k\left(\gamma P\left(w_{k \ 1} \quad A_k v^{\pi_k^0 \ 1}\right) + E_k\right) \\ &\overset{(d)}{\quad} \pi_k\left(\gamma P(w_{k \ 1} \quad A_{k \ 1} v^{\pi_k^0 \ 1}) + \alpha^{k \ 1}\gamma H\mathbf{1} + E_k\right)\,, \end{aligned}$$

where (a) follows from the definition of w_k, (b) is due to the equation (4), (c) follows from the definition of the Bellman operator, and (d) is due to the fact that $(A_k \quad A_{k \ 1})v^{\pi_k^0 \ 1} = \alpha^{k \ 1} v^{\pi_k^0 \ 1} \quad \alpha^{k \ 1} H\mathbf{1}$. From this result and the fact that $w_0 = \mathbf{0}$,

$$w_1 \quad A_1 v^{\pi_1^0} \quad \pi_1\left(\gamma P w_0 + \gamma H\mathbf{1} + E_1\right) \quad \gamma H\mathbf{1} + \pi_1 E_1\,.$$

Therefore, the inequality (6) holds for $k = 1$. From the step (d) above and induction, it is straightforward to verify that the inequality (6) holds for other k. □

F.2 Proof of Lemmas 2 and 3 (Coarse State-Value Bound)

The next lemma is necessary to bound E_k by using the Azuma-Hoeffding inequality (Lemma 17).

Lemma 24. *For any k 2 $[K]$, $v_{k\ 1}$ is bounded by H.*

Proof. We prove the claim by induction. The claim holds for $k = 1$ since $v_0 = \mathbf{0}$ by definition. Assume that $v_{k\ 1}$ is bounded by H for some k 1. Then, from the greediness of the policies π_k and $\pi_{k\ 1}$,

$$\pi_{k\ 1} q_k = \pi_{k\ 1}(s_k \quad \alpha s_{k\ 1}) \quad v_k \quad \pi_k(s_k \quad \alpha s_{k\ 1}) = \pi_k q_k$$

Since $q_k = r + \gamma \widehat{P}_{k\ 1} v_{k\ 1}$ is bounded by H due to the induction hypothesis, the claim holds. □

Proof of Lemma 2. Consider a fixed k 2 $[K]$ and (x, a) 2 $\mathbf{X}$ $\mathbf{A}$. Since

$$E_k(x,a) = \frac{\gamma}{M} \sum_{j=1}^{k} \alpha^{k\ j} \sum_{m=1}^{M} \underbrace{(v_{j\ 1}(y_{j\ 1,m,x,a}) \quad P v_{j\ 1}(x,a))}_{\text{bounded by } 2H \text{ from Lemma 24}},$$

$E_k(x, a)$ is a sum of bounded martingale differences with respect to the filtraion $(\mathbf{F}_{j,m})_{j=1,m=1}^{k,M}$. Therefore, using the Azuma-Hoeffding inequality (Lemma 17),

$$\mathsf{P}\left(jE_k j(x,a) \quad 3H\sqrt{\frac{A_1 \iota_1}{M}} \right) \quad \frac{\delta}{4KXA},$$

where the bound in $\mathsf{P}(\ )$ is simplified by $2^{p}\overline{2}\gamma$ 3 and $\sum_{j=1}^{k} \alpha^{2(k\ j)}$ $\sum_{j=1}^{k} \alpha^{k\ j} = A_1$. Taking the union bound over (x, a, k) 2 $\mathbf{X}$ $\mathbf{A}$ $[K]$,

$$\mathsf{P}(E_1) \quad 1 \quad \sum_{(x,a)2\mathbf{X}} \sum_{\mathbf{A}\,k=1}^{K} \mathsf{P}\left(jE_k j(x,a) \quad 3H\sqrt{\frac{A_1 \iota_1}{M}} \right) \quad 1 \quad \frac{\delta}{4},$$

and thus $\mathsf{P}(E_1^c)$ $\delta/4$, which is the desired result. □

Proof of Lemma 3. We condition the proof by the event E_1. This event occurs with probability at least 1 $\delta/4$. Note that under the current setting of α, $A_1 = H$. From Lemma 2 and the settings of α and M,

$$\sum_{j=0}^{k\ 1} \gamma^j \left(\pi_k P_k^{k\ 1\ j} \quad \pi \ P^j \right) E_{k\ j} \quad 2\sum_{j=0}^{k\ 1} \gamma^j k E_{k\ j} k_1 \quad \frac{H^{p}\overline{H}\varepsilon}{p\overline{c_2}}.$$

Thus, from Lemma 1, v $v^{\pi_k^0}$ $\sqrt{H/c_2}\varepsilon + 2(H+k)\gamma^k \mathbf{1}$. Finally, using Lemma 15,

$$2(H+K)\gamma^K \quad \frac{\varepsilon}{c_1},$$

and thus,

$$kv \quad v^{\pi_K^0} k_1 \quad \varepsilon\sqrt{\frac{H}{c_2}} + \frac{\varepsilon}{c_1} \quad \left(\frac{1}{c_1} + p\frac{1}{\overline{c_2}} \right)^{p} \overline{H}\varepsilon.$$

Therefore, for some c_1 and c_2, the claim holds. □

F.3 Proof of Lemma 4 (Value Estimation Error Bound)

We first prove an intermediate result.

Lemma 25. *For any $k \in [K]$,*

$$v^{\pi_k^0} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} \le \gamma^k H \mathbf{1} \le v_k - v^{\pi_k^0} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} + \gamma^k H \mathbf{1}.$$

Proof. From the greediness of π_{k-1}, $v_k = w_k - \alpha w_{k-1} \le \pi_k(s_k - \alpha s_{k-1}) = \pi_k(r + \gamma P v_{k-1} + \varepsilon_k)$. By induction on k, therefore,

$$v_k - \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} (r + \varepsilon_{k-j}) + \underbrace{\gamma^k \pi_k P_0^{k-1} \cdots P_1^0 v_0}_{=\mathbf{0}} = \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} (r + \varepsilon_{k-j}),$$

Note that

$$T_0^{k-1} \cdots T_0^1 q^{\pi_0} = \sum_{j=0}^{k-1} \gamma^j P_k^{k-1} \cdots P_{k-j+1}^{k-j} r + \gamma^k \underbrace{P_0^{k-1} \cdots P_1^0 q^{\pi_0}}_{H\mathbf{1}} \le \sum_{j=0}^{k-1} \gamma^j P_k^{k-1} \cdots P_{k-j+1}^{k-j} r \le T_0^{k-1} \cdots T_0^1 q^{\pi_0} + \gamma^k H.$$

Accordingly, $v_k - \pi_k T_0^{k-1} \cdots T_0^1 q^{\pi_0} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} \le \gamma^k H \mathbf{1}$.

Similarly, from the greediness of π_k, $v_k = w_k - \alpha w_{k-1} \ge \pi_{k-1}(s_k - \alpha s_{k-1}) \ge \pi_{k-1}(r + \gamma P v_{k-1} + \varepsilon_k)$. By induction on k, therefore,

$$v_k - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_k^{k-1} \cdots P_{k-j+1}^{k-j} (r + \varepsilon_{k-j}) + \underbrace{\gamma^{k-1} \pi_{k-1} P_0^{k-1} \cdots P_1^0 P v_0}_{=\mathbf{0}} \ge 0.$$

Note that $T_0^{k-2} \cdots T_0^1 q^{\pi_0} = T_0^{k-2} \cdots T_0^1 (r + \gamma P v^{\pi_0})$, and

$$T_0^{k-1} \cdots T_0^1 q^{\pi_0} = \sum_{j=0}^{k-1} \gamma^j P_k^{k-1} \cdots P_{k-j+1}^{k-j} r + \gamma^k \underbrace{P_0^{k-1} \cdots P_1^0 P v^{\pi_0}}_{H\mathbf{1}} \ge \sum_{j=0}^{k-1} \gamma^j P_k^{k-1} \cdots P_{k-j+1}^{k-j} r \ge T_0^{k-1} \cdots T_0^1 q^{\pi_0} - \gamma^k H.$$

Accordingly, $v_k - \pi_{k-1} T_0^{k-1} \cdots T_0^1 q^{\pi_0} + \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} \ge -\gamma^k H \mathbf{1}$. □

Proof of Lemma 4. From Lemma 25 and $\pi_k T^{\pi_k - 1:1} q^{\pi_0} = v^{\pi_k^0}$ v, we have that

$$v^{\pi_k^0} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} - 2\gamma^k H \mathbf{1} \le v_k - v + \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} + 2\gamma^k H \mathbf{1},$$

where we loosened the bound by multiplying $\gamma^k H$ by 2. By simple algebra, the lower bound for $v - v_k$ is obtained. On the other hand, from Lemma 1,

$$v^{\pi_k^0} - v \le \frac{1}{A_{1-1}} \sum_{j=0}^{k-2} \gamma^j \left(\pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} - \pi P^j \right) E_{k-1-j} \le 2H \left(\alpha^{k-1} + \frac{A_{\gamma,k-1}}{A_{1-1}} \right) \mathbf{1}$$

for any $k \in \{2, \ldots, K\}$. Therefore, we have that

$$\begin{aligned} v - v_k &\le 2H \left(\alpha^{k-1} + \gamma^k + \frac{A_{\gamma,k-1}}{A_{1-1}} \right) \mathbf{1} \\ &\quad + \frac{1}{A_{1-1}} \sum_{j=0}^{k-2} \gamma^j \left(\pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} - \pi P^j \right) E_{k-1-j} - \sum_{j=0}^{k-1} \gamma^j \pi_k P_k^{k-1} \cdots P_{k-j+1}^{k-j} \varepsilon_{k-j} \end{aligned}$$

for any $k \in \{2,\dots,K\}$.

Finally, for $k=1$, since $v_1 = \pi_1 q_1 = \pi_1 r$,

$$\gamma H \mathbf{1} \quad \pi \ (q \quad r) \quad v \quad v_1 \quad \gamma\pi \ Pv \quad \gamma H \mathbf{1}.$$

As $\Gamma_1 \quad \mathbf{0}$, the claim holds for $k=1$ too. □

F.4 Proof of Lemmas 5 and 6 (Value Estimation Variance Bound)

Proof of Lemma 5. Consider a fixed $k \in [K]$ and $(x,a) \in \mathbf{X} \quad \mathbf{A}$. Since

$$\varepsilon_k(x,a) = \frac{\gamma}{M} \sum_{m=1}^{M} \underbrace{(v_k \ _1(y_k \ _{1,m,x,a}) \quad Pv_k \ _1(x,a))}_{\text{bounded by } 2H \text{ from Lemma 24}},$$

$\varepsilon_k(x,a)$ is a sum of martingale differences with respect to the filtraion $(\mathbf{F}_{k,m})_{m=1}^{M}$ and bounded by $2\gamma H/M$. Therefore, using the Azuma-Hoeffding inequality (Lemma 17),

$$\mathsf{P}\left(|\varepsilon_k|(x,a) \quad 3H\sqrt{\frac{\iota_1}{M}}\right) \quad \frac{\delta}{4KXA},$$

where the bound in $\mathsf{P}(\ )$ is simplified by $2^{p}\overline{2} \quad 3$. Taking the union bound over $(x,a,k) \in \mathbf{X} \quad \mathbf{A} \quad [K]$,

$$\mathsf{P}(E_2) \quad 1 \quad \sum_{(x,a)\in\mathbf{X}} \ \sum_{\mathbf{A}}^{K} \ \sum_{k=1} \mathsf{P}\left(|\varepsilon_k|(x,a) \quad 3H\sqrt{\frac{\iota_1}{M}}\right) \quad 1 \quad \frac{\delta}{4},$$

and thus $\mathsf{P}(E_2^c) \quad \delta/4$, which is the desired result. □

Next, we prove a uniform bound on $v \quad v_k$.

Lemma 26. *Conditioned on $E_1 \setminus E_2$,*

$$\|v \quad v_k\|_1 < 2H \min\left\{1, \gamma^k + \alpha^k \ ^1 + \frac{A_{\gamma,k} \ _1}{A_1} + 6H\sqrt{\frac{\iota_1}{M}}\right\}$$

for all $k \in [K]$, where $1/0 \coloneqq 1$.

Proof. Let $e_k \coloneqq \gamma^k H + H \max_{j\in[k]} \|\varepsilon_j\|_1$. From Lemma 4, $v \quad v_k \quad 2e_k\mathbf{1}$ for any $k \in [K]$, and

$$v \quad v_k \quad 2H\left(\alpha^k \ ^1 + \frac{A_{\gamma,k} \ _1}{A_1} + \frac{1}{A_1} \max_{j\in[k \ 1]} \|E_j\|_1\right)\mathbf{1} + 2e_k\mathbf{1}$$

for any $k \in \{2,\dots,K\}$. Note that $\|v \quad v_k\|_1 \quad 2H$ from Lemma 24 for any k. Combining these results with Lemmas 2 and 5,

$$\begin{aligned}\|v \quad v_k\|_1 &< 2H \min\left\{1, \gamma^k + \alpha^k \ ^1 + \frac{A_{\gamma,k} \ _1}{A_1} + 3H\sqrt{\frac{\iota_1}{M}}\left(1 + \sqrt{\frac{1}{A_1}}\right)\right\} \\ &\quad 2H \min\left\{1, \gamma^k + \alpha^k \ ^1 + \frac{A_{\gamma,k} \ _1}{A_1} + 6H\sqrt{\frac{\iota_1}{M}}\right\}\end{aligned}$$

for all $k \in [K]$, where we used the fact that $1 \quad A_1$. This concludes the proof. □

Now, we are ready to prove Lemma 6.

Proof of Lemma 6. Clearly $\sigma(v_0) = \mathbf{0}$ since $v_0 = \mathbf{0}$. From Lemma 21, $\sigma(v_k) \quad \sigma(v_k \quad v\ ) + \sigma(v\ )$. Using Popoviciu's inequality on variances (Lemma 20) together with Lemma 26,

$$\sigma(v_k \quad v\ ) \quad 2H \min\left\{1, \gamma^k + \alpha^{k\ \ 1} + \frac{A_{\gamma,k\ \ 1}}{A_1} + 6H\sqrt{\frac{\iota_1}{M}}\right\},$$

where we used a simple formula, $\min\{a, bg^2 = \min\{a^2, b^2\}g$ for any scalars $a, b \quad 0$. Finally, loosening the bound by replacing $\gamma^k + \alpha^{k\ \ 1}$ by $2\max\{\alpha, \gamma\}g^{k\ \ 1}$, the claim holds. □

F.5 Proof of Lemmas 7 and 8 (Error Bounds with Bernstein's Inequality)

Proof of Lemma 7. Consider a fixed k 2 $[K]$ and (x, a) 2 $\mathbf{X} \quad \mathbf{A}$. Since

$$E_k(x,a) = \frac{\gamma}{M}\sum_{j=1}^{k}\alpha^{k\ \ j}\sum_{m=1}^{M}\underbrace{(v_{j\ \ 1}(y_{j\ \ 1,m,x,a}) \quad Pv_{j\ \ 1}(x,a))}_{\text{bounded by } 2H \text{ from Lemma 24}},$$

$E_k(x, a)$ is a sum of bounded martingale differences with respect to the filtraion $(\mathbf{F}_{j,m})_{j=1,m=1}^{k,M}$. From the facts that $v_0 = \mathbf{0}$, and $\gamma \quad 1$,

$$V' = \frac{\gamma^2}{M}\sum_{j=1}^{k}\alpha^{2(k\ \ j)}\mathrm{Var}(v_{j\ \ 1})(x,a) \quad \underbrace{\frac{1}{M}\sum_{j=2}^{k}\alpha^{2(k\ \ j)}\mathrm{Var}(v_{j\ \ 1})(x,a)}_{:=\sim},$$

Since we are conditioned with the event $E_1 \setminus E_2$, the inequality (3) in Lemma 6 holds and implies that the predictable quadratic variation V' satisfies the following inequality:

$$\begin{aligned}
\sim \quad & \sum_{j=2}^{k}\alpha^{2(k\ \ j)}\left(\sigma(v\ )(x,a) + 2H\min\left\{1, 2\max\{\alpha,\gamma\}g^{j\ \ 2} + \frac{A_{\gamma,j\ \ 2}}{A_1} + 6H\sqrt{\frac{\iota_1}{M}}\right\}\right)^2 \\
& \sum_{j=2}^{k}\alpha^{2(k\ \ j)}\left(\sigma(v\ )(x,a) + 2H\left(2\max\{\alpha,\gamma\}g^{j\ \ 2} + \frac{A_{\gamma,j\ \ 2}}{A_1} + 6H\sqrt{\frac{\iota_1}{M}}\right)\right)^2 \\
& 4\sum_{j=2}^{k}\alpha^{2(k\ \ j)}\left(\mathrm{Var}(v\ )(x,a) + 4H^2\left(4\max\{\alpha,\gamma\}g^{2(j\ \ 2)} + \frac{A^2_{\gamma,j\ \ 2}}{A^2_1} + \frac{36H^2\iota_1}{M}\right)\right),
\end{aligned}$$

where the last line follows from Lemma 12. Consequently, V' is bounded by

$$V' \quad \frac{4}{M}\sum_{j=2}^{k}\alpha^{2(k\ \ j)}\left(\mathrm{Var}(v\ )(x,a) + 4H^2\left(4\max\{\alpha,\gamma\}g^{2(j\ \ 2)} + \frac{A^2_{\gamma,j\ \ 2}}{A^2_1} + \frac{36H^2\iota_1}{M}\right)\right),$$

which is equal to $V_k(x, a)$. Using Lemma 19 and taking the union bound over (x, a, k) 2 $\mathbf{X} \quad \mathbf{A} \quad [K]$,

$$\mathsf{P}\left(9(x,a,k)\ 2\ \mathbf{X} \quad \mathbf{A} \quad [K] \text{ s.t. } jE_Kj(x,a) \quad \frac{4H\iota_2}{3M} + \sqrt{2V_k(x,a)\iota_2}\,\middle|\, E_1 \setminus E_2\right) \quad \frac{\delta}{4}.$$

(Recall that $\mathsf{P}(E_1 \setminus E_2) \quad 1 \quad \frac{\delta}{2} \quad \frac{1}{2}$, and hence, we need to use ι_2.) Thus, $\mathsf{P}(E_3^c jE_1 \setminus E_2) \quad \frac{\delta}{4}$. □

Proof of Lemma 8. Consider a fixed k 2 $[K]$ and (x, a) 2 $\mathbf{X} \quad \mathbf{A}$. Since

$$\varepsilon_k(x,a) = \frac{\gamma}{M}\sum_{m=1}^{M}\underbrace{(v_{k\ \ 1}(y_{k\ \ 1,m,x,a}) \quad Pv_{k\ \ 1}(x,a))}_{\text{bounded by } 2H \text{ from Lemma 24}},$$

$\varepsilon_k(x,a)$ is a sum of bounded martingale differences with respect to $\mathbf{F}_{k,m}$. Since we are conditioned with the event $E_1 \setminus E_2$, the inequality (3) in Lemma 6 holds and implies that the predictable quadratic variation V^0 can be shown to satisfy the following inequality as in the proof of Lemma 7:

$$V^0 = \frac{\gamma^2}{M}\mathrm{Var}\left(v_{k\ 1}\right)(x,a) \quad \frac{4}{M}\overline{\mathrm{Var}}_k\,,$$

where the last line is equal to $W_k(x,a)$. (Note that $v_0 = \mathbf{0}$.)

Using Lemma 19 and taking the union bound over (x,a,k) 2 $\mathbf{X}$ $\mathbf{A}$ $[K]$,

$$\mathsf{P}\left(9(x,a,k)\ 2\ \mathbf{X} \quad \mathbf{A} \quad [K]\ \text{s.t.}\ j\varepsilon_k j(x,a) \quad \frac{4H\iota_2}{3M} + \sqrt{2W_k(x,a)\iota_2}\middle| E_1 \setminus E_2\right) \quad \frac{\delta}{4}\,.$$

(Recall that $\mathsf{P}(E_1 \setminus E_2)$ 1 $\frac{\delta}{2}$ $\frac{1}{2}$, and hence, we need to use ι_2.) Thus, $\mathsf{P}\left(E_4^c j E_1 \setminus E_2\right)$ $\frac{\delta}{4}$. □

G Proof of Lemmas for Theorem 2 (Bound for a Stationary Policy)

We use the same notations as those used in Appendix F.

G.1 Proof of Lemma 9 (Error Propagation Analysis)

To prove Lemma 9, we need the following lemma.

Lemma 27. *For any k 2 $[K]$, let $\Delta_k \coloneqq w_k$ $w_{k\ 1}$. Then, for any k 2 $[K]$,*

$$\pi_{k\ 1}\sum_{j=0}^{k\ 1}\gamma^j P_{k\ 1\ j}^{k\ 2} E_{k\ j}^0 \quad A_{\gamma,k}\mathbf{1} \quad \Delta_k \quad \pi_k\sum_{j=0}^{k\ 1}\gamma^j P_{k\ j}^{k\ 1} E_{k\ j}^0 + A_{\gamma,k}\mathbf{1}\,.$$

Proof. We prove only the upper bound by induction as the proof for a lower bound is similar. We have that $\Delta_k = \pi_k s_k$ $\pi_{k\ 1} s_{k\ 1}$ $\pi_k(s_k$ $s_{k\ 1})$, where the inequality follows from the greediness of $\pi_{k\ 1}$. Let $\sim_k \coloneqq s_k$ $s_{k\ 1}$. Since $s_0 = \mathbf{0}$, $\sim_1 = r + E_1^0$ $\mathbf{1} + E_1^0$. From the monotonicity of π_1, the claim holds for $k = 1$. Assume that for some k 1 1, the claim holds. Then, from the equation (4), the induction hypothesis, and the monotonicity of P,

$$\begin{aligned}\sim_k &= (A_k \quad A_{k\ 1})\,r + \gamma P\Delta_{k\ 1} + E_k^0 \\ &\quad \sum_{j=0}^{k\ 1}\gamma^j P_{k\ j}^{k\ 1} E_{k\ j}^0 + (\alpha^{k\ 1} + \gamma A_{\gamma,k\ 1})\mathbf{1} = \sum_{j=0}^{k\ 1}\gamma^j P_{k\ j}^{k\ 1} E_{k\ j}^0 + A_{\gamma,k}\mathbf{1}\,.\end{aligned}$$

The claimed upper bound follows from the monotonicity of π_k. □

Now, we are ready to prove Lemma 9.

Proof of Lemma 9. Note that

$$\mathbf{0} \quad v \quad v^{\pi_k} = \frac{A_k}{A_1}(v \quad v^{\pi_k}) + \alpha^k(v \quad v^{\pi_k}) \quad \frac{A_k}{A_1}(v \quad v^{\pi_k}) + 2H\alpha^k\mathbf{1}$$

since v v^{π_k} $2H\mathbf{1}$. Therefore, we need an upper bound for $A_k(v$ $v^{\pi_k})$. We decompose $A_k(v$ $v^{\pi_k})$ to $A_k v$ w_k and w_k $A_k v^{\pi_k}$. Then, we derive upper bounds for each of them. The desired result is obtained by summing up those bounds.

Upper bound for $A_k v \quad w_k$. Note that

$$\begin{aligned} A_k v \quad w_k &\overset{(a)}{=} N^{\pi} \quad (\pi \quad (A_k r + \gamma P w_k) \quad w_k) \\ &\overset{(b)}{=} N^{\pi} \quad \pi \quad (A_k r + \gamma P w_k \quad s_k) \\ &\overset{(c)}{=} N^{\pi} \quad \pi \quad (\gamma P (w_k \quad w_{k \quad 1}) \quad E_k) \\ &\overset{(d)}{=} N^{\pi} \quad \pi \quad \left(\sum_{j=1}^{k} \gamma^j P^k_{k+1 \quad j} E^0_{k+1 \quad j} \quad E_k \right) + H A_{\gamma,k} \mathbf{1}, \end{aligned}$$

where (a) is due to the fact that $I = N^{\pi}(I \quad \gamma\pi P)$ and $v^{\pi} = N^{\pi}\pi r$ for any policy π, (b) is due to the greediness of π_k, (c) follows from the equation (4), and (d) follows from Lemma 27.

Upper bound for $w_k \quad A_k v^{\pi_k}$. We have that

$$\begin{aligned} w_k \quad A_k v^{\pi_k} &\overset{(a)}{=} N^{\pi_k} (w_k \quad \pi_k (A_k r + \gamma P w_k)) \\ &\overset{(b)}{=} N^{\pi_k} \pi_k (w_k \quad A_k r \quad \gamma P w_k) \\ &\overset{(c)}{=} N^{\pi_k} \pi_k (\quad \gamma P (w_k \quad w_{k \quad 1}) + E_k) \\ &\overset{(d)}{=} N^{\pi_k} \pi_k \left(E_k \quad \sum_{j=1}^{k} \gamma^j P^k_{k \quad j} {}^{1} E^0_{k+1 \quad j} \right) + H A_{\gamma,k} \mathbf{1}, \end{aligned}$$

where (a) is due to the fact that $I = N^{\pi}(I \quad \gamma\pi P)$ and $v^{\pi} = N^{\pi}\pi r$ for any policy π, (b) is due to the definition of w_k, (c) follows from the equation (4), and (d) follows from Lemma 27. □

G.2 Proof of Lemma 10 (Coarse State-Value Bounds)

Before starting the proof, we note that $A_{\,1} = H^2$ under the current setting.

Proof of Lemma 10. From Lemma 2, $\|E_k\|_{\,1} \quad 3H\sqrt{A_{\,1}\,\iota_1/M} \quad 3\varepsilon\sqrt{H^3/c_4}$ for any $k \in [K]$. On the other hand, from Lemma 5, $\|\varepsilon_k\|_{\,1} \quad 3H\sqrt{\iota_1/M} \quad 3\varepsilon\sqrt{H/c_4}$ for any $k \in [K]$. Combining these bounds with Lemma 1,

$$v \quad v^{\pi^0_k} \quad \overline{H} \max_{j \in [k]} \|E_j\|_{\,1} \mathbf{1} + \quad H\alpha^k \mathbf{1} \quad \left(\varepsilon\sqrt{\frac{H}{c_4}} + H\alpha^k \right) \mathbf{1}$$

for any $k \in [K]$, where we used the fact that $A_{\gamma,k}/A_{\,1} \quad \alpha^k/H \quad \alpha^k$, which follows from Lemma 13, is used. Furthermore, combining previous upper bounds for errors with Lemma 9,

$$\begin{aligned} v \quad v^{\pi_k} \quad & 2H \underbrace{\left(\alpha^k + \frac{A_{\gamma,k}}{A_{\,1}} \right)}_{2\alpha^k \text{ from (a)}} \mathbf{1} + \frac{1}{A_{\,1}} \underbrace{(N^{\pi_k}\pi_k \quad N^{\pi} \quad \pi \quad) E_k}_{2HkE_k\|_{\,1}\,\mathbf{1} \text{ from (b)}} \\ & + \frac{1}{A_{\,1}} \sum_{j=1}^{k} \gamma^j \underbrace{\left(N^{\pi} \quad \pi \quad P^k_{k+1 \quad j} \quad N^{\pi_k}\pi_k P^k_{k \quad j} {}^{1} \right) E^0_{k+1 \quad j}}_{2H(\|\varepsilon_{k+1 \quad j}\|_{\,1} + (1 \quad \alpha)\|E_{k \quad j}\|_{\,1})\mathbf{1} \text{ from (c)}} \\ \overset{(d)}{\quad} & 4H\alpha^k \mathbf{1} + \frac{2}{H}\|E_k\|_{\,1} + 2 \max_{j \in [k]} \left(\|\varepsilon_j\|_{\,1} + \frac{1}{H^2}\|E_j\|_{\,1} \right) \\ & 4H\alpha^k \mathbf{1} + 6\varepsilon\sqrt{\frac{H}{c_4}} + \frac{6\varepsilon}{\cancel{p}\overline{c_4}} \left(\cancel{p}\overline{H} + \cancel{p}\frac{1}{\overline{H}} \right) \mathbf{1} = \quad \left(\varepsilon\sqrt{\frac{H}{c_4}} + H\alpha^k \right) \mathbf{1} \end{aligned}$$

for any $k \in [K]$, where (a) follows as $A_{\gamma,k}/A_1 \quad \alpha^k/H \quad \alpha^k$ from Lemma 13, (b) is due to the monotonicity of stochastic matrices, and $\|E_k\|_1 \mathbf{1} \quad E_k \quad \|E_k\|_1 \mathbf{1}$ for any $k \in [K]$, (c) is due to the monotonicity of stochastic matrices, and $(\|\varepsilon_k\|_1 + (1 \quad \alpha)\|E_k \quad {}_1\|_1)\mathbf{1} \quad E_k^0 \quad (\|\varepsilon_k\|_1 + (1 \quad \alpha)\|E_k \quad {}_1\|_1)\mathbf{1}$ for any $k \in [K]$, and (d) follows by taking the maximum over j. □

H Details on empirical illustrations

This appendix details the settings used for the illustrations of Section 6. It provides

- a precise definition of the Garnet setting and pseudo-code for Q-LEARNING in Appendix H.1;
- additional numerical experiments illustrating the effects of α and M on the algorithm in Appendix H.2.

H.1 Detailed setting

Garnets. We use the Garnets (Archibald et al., 1995) class of random MDPs. A Garnet is characterized by three integer parameters, X, A, and B, that are respectively the number of states, the number of actions, and the branching factor – the maximum number of accessible new states in each state. For each $(x,a) \in \mathbf{X} \quad \mathbf{A}$, we draw B states $(y_1,\dots,y_B)$ from $\mathbf{X}$ uniformly without replacement. Then, we draw $B \quad 1$ numbers uniformly in $(0,1)$, denoting them sorted as $(p_1,\dots,p_{B \quad 1})$. We set the transition probability $P_{x,a}^{y_k} = p_k \quad p_{k \quad 1}$ for each $1 \quad k \quad B$, with $p_0 = 0$ and $p_B = 1$. Finally, the reward function, depending only on the states, is drawn uniformly in $(\quad 1,1)$ for each state. In our examples, we used $X = 8$, $A = 2$, and $B = 2$. We compute our experiments with $\gamma = 0.9$.

Q-learning. For illustrative purposes, we compare the performance of MDVI to the one of a sampled version of Q-LEARNING, that we know is not minimax-optimal. For completeness, the pseudo-code for this method is given in Algorithm 2. It shares the time complexity of MDVI, but has a lower memory complexity, since it does not need to store an additional XA table.

Algorithm 2: Q-LEARNING(K, M, w)

Input: number of iterations K, number of samples per iteration M, $w \in [0.5,1]$ a learning rate parameter.

Let $q_0 = \mathbf{0} \in \mathbf{R}^{XA}$;
for k from 0 to $K \quad 1$ do
  for each state-action pair $(x,a) \in \mathbf{X} \quad \mathbf{A}$ do
    Sample $(y_{k,m,x,a})_{m=1}^M$ from the generative model $P(\cdot|x,a)$;
    Let $m_{k+1}(x,a) = r(x,a) + \gamma M \quad {}^1 \sum_{m=1}^M \max_{a'} q_k(y_{k,m,x,a}, a')$;
  end
  Let $\eta_k = (k+1) \quad {}^w$;
  Let $q_{k+1} = (1 \quad \eta_k) q_k + \eta_k m_{k+1}$;
end
return π_K , *a greedy policy with respect to* q_K;

H.2 Additional numerical illustrations

Additional experiment for sample complexity. In Figure 1, we plot the sample complexity of a standard version of Q-LEARNING using $w = 1$ (*i.e* performing an exact average of q-values). However, we know (Even-Dar et al., 2003) that we can reach a better sample complexity by choosing a more appropriate w in $(0.5,1)$. In Figure 2, we provide the sample complexity for MDVI, and Q-LEARNING with $w = 1$ and $w = 0.7$.

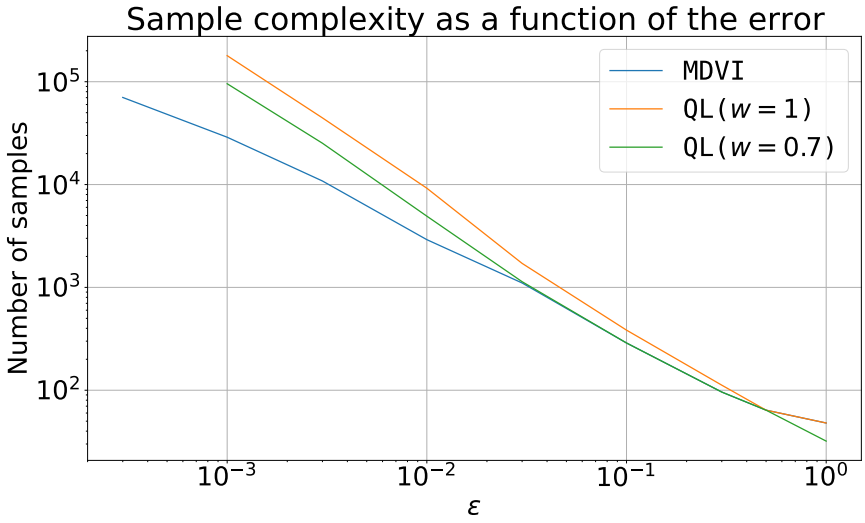

Figure 2: Number of samples needed to reach a certain error.

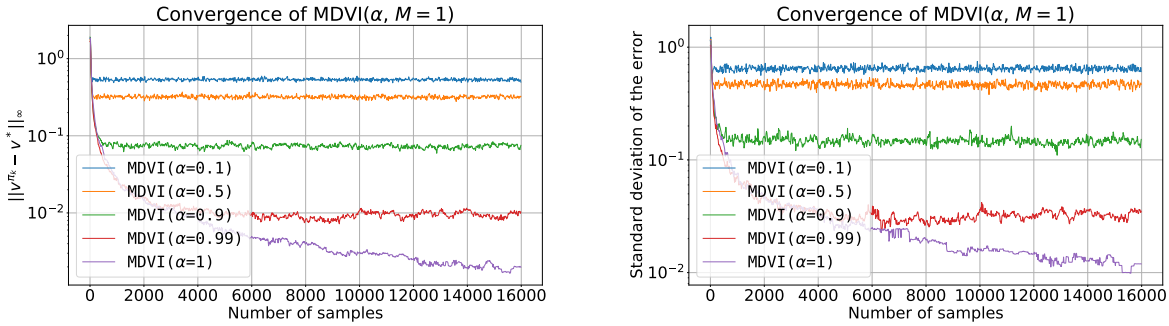

Figure 3: Error of the policy computed by MDVI in function of the number of samples used. Left: mean, Right: standard deviation; estimated over 1000 MDPs.

The version with $w = 0.7$ catches up with MDVI at high errors, but the difference is still quite large at higher precision. Note that we add additional data points for $\varepsilon < 10^{-3}$. Both versions of Q-LEARNING do not have sample complexity plotted for these errors, because they did not reach these ε in the number of iterations we ran them (up to 10^5 iterations).

Influence of α. We showcase the impact of α when $M = 1$ in Figure 3. With $\alpha = 1$, MDVI will asymptotically converge to π . With a $\alpha < 1$, MDVI will reach an ε-optimal policy, but will not actually converge to the optimal policy of the MDP (although this ε can be controlled by choosing a large enough value for α, or a larger value of M). Indeed, in the latter case, the distance to the optimal policy depends on a moving average of the errors (by a factor α). The moving average reduces the variance, but does not bring it zero, contrarily to the exact average implicitly performed when $\alpha = 1$. This behaviour is illustrated in Figure 3. We observe there that, with $M = 1$, one has to choose a large enough value of α to reach a policy close enough to the optimal one.

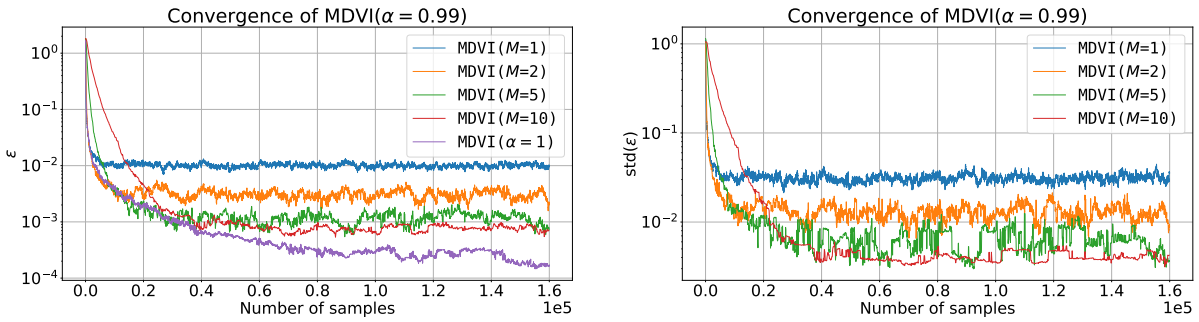

Figure 4: Error of the policy computed by MDVI in function of the number of samples used, for different values of M. **Left**: mean, **Right**: standard deviation; estimated over 1000 MDPs. For this value of $\gamma = 0.9$, choosing $\alpha = 0.99$ matches the condition $\alpha = 1 \quad (1 \quad \gamma)^2$.

Influence of M. Choosing the right M is not that obvious from the theory (it notably depends on an unknown constant c_2). We illustrate in Figure 4 the influence it has on the speed of convergence of MDVI. We run MDVI with $\alpha = 0.99$ (for a setting where $\gamma = 0.9$), and for different values of M. With a fixed α, a larger M allows MDVI to reach a lower asymptotic error, but slows down the learning in early iterations. M cannot however be chosen as large as possible: at one point it start to be useless to increase its value. For instance, moving from $M = 5$ to $M = 10$ does not allow for a noticeable lower error, but slows the learning. We compare this to the setting where $\alpha = 1$ for completeness.