Entropic Risk Optimization in Discounted MDPs: Sample Complexity Bounds with a Generative Model

Oliver Mortensen[1] **Mohammad Sadegh Talebi**[1]
[1]University of Copenhagen

Abstract

In this paper we analyze the sample complexities of learning the optimal state-action value function Q^* and an optimal policy π^* in a discounted Markov decision process (MDP) where the agent has recursive entropic risk-preferences with risk-parameter $\beta \neq 0$ and where a generative model of the MDP is available. We provide and analyze a simple model based approach which we call model-based risk-sensitive Q-value-iteration (MB-RS-QVI) which leads to (ϵ, δ)-PAC-bounds on $\|Q^* - Q^k\|$, and $\|V^* - V^{\pi_k}\|$ where Q_k is the output of MB-RS-QVI after k iterations and π_k is the greedy policy with respect to Q_k. Both PAC-bounds have exponential dependence on the effective horizon $\frac{1}{1-\gamma}$ and the strength of this dependence grows with the learners risk-sensitivity $|\beta|$. We also provide two lower bounds which shows that exponential dependence on $|\beta|\frac{1}{1-\gamma}$ is unavoidable in both cases. The lower bounds reveal that the PAC-bounds are both tight in ε and δ and that the PAC-bound on Q-learning is tight in the number of actions A, and that the PAC-bound on policy-learning is nearly tight in A.

1 Introduction

In reinforcement learning (RL), the agent conventionally optimizes the expected return, which is defined in terms of a (discounted) sum of rewards [50]. In the majority of RL literature, the environment is modeled via the Markov Decision Process (MDP) framework [41], wherein efficient computation of an optimal policy, thanks to optimal Bellman equations, renders possible. However, as a *risk-neutral* objective, the expected return fails to capture the true needs of many high-stake applications arising in, e.g., medical treatment [19], finance [44, 9], and operations research [16]. Decision making in such applications must take into account the variability of returns, and risks thereof. To account for this, one may opt to to maximize a risk measure of the return distribution, while another approach could be to consider the entire distribution of return, as is done under the distributional RL framework [8] that has received much attention over the last decade.

Within the first approach, the risk is quantified via concave risk measures, which yield well-defined mathematical optimization frameworks. Notably, they include value-at-risk (VaR), Conditional VaR (CVaR) [45], entropic risk [23], and entropic VaR (EVaR) [2], all of which have been applied to a wide-range of scenarios. CVaR appears to be the most popular one used to model risk-sensitivity in MDPs [15, 10, 12, 7], mainly due to a delicate control it offers for the undesirable tail of return. Despite its popularity and rich interpretation, solving and learning MDPs with CVaR-defined objectives has rendered technically challenging [7]. This has been a key motivation of adopting weaker notions such as nested CVaR [6], at the expense of sacrificing the interpretability. Entropic risk, as another popular notion, has been considered for risk control in MDPs [11, 39, 22, 24, 20]. In the RL literature, it has been mainly considered for the undiscounted settings, despite the popularity of dis-

18th European Workshop on Reinforcement Learning (EWRL 2025).

counted MDPs. A notable exception is [22] that studies, among other things, planning in discounted MDPs under the entropic risk.

In this paper, we study risk-sensitive discounted RL where the agent's objective is formulated using the entropic risk measure. In discounted RL, where future rewards are discounted by a factor of γ, one may identify two main approaches to apply the entropic risk to sequence of rewards, $(r_t)_{t\geq 0}$, collected by an RL agent. The first and most intuitive one, which we may call the *non-recursive* approach, consists in directly applying the entropic risk functional to return $\sum_{t=0}^{\infty} \gamma^t r_t$. The other approach, which we may call the *recursive* approach, works by applying the risk functional at every step t (see Section 2 for details). The non-recursive approach (e.g., [22]), while being most intuitive, has several drawbacks; e.g., the optimal policy might not be time-consistent (see [25]). In contrast, the recursive approach yields a form of Bellman optimality equation, which is key to developing learning algorithms with provable sample complexity. Therefore, we restrict attention to the RL problem defined using recursive risk-preferences.

1.1 Main Contributions

We study risk-sensitive RL in finite discounted MDPs under the recursively applied entropic risk measure, assuming that the agent is given access to a generative model of the environment. The agent's learning performance is assessed via sample complexity defined as the total number T of samples needed to learn, for input (ε, δ), either an ε-optimal policy (which we call policy learning), or an ε-close approximation (in max-norm) to the optimal Q-value (which we call Q-value learning), with probability exceeding $1 - \delta$. Specifically, we make the following contributions: We present an algorithm called Model-Based Risk-Sensitive Q-Value Iteration (MB-RS-QVI), a model-based RL algorithms for the RL problem considered, which is derived using a simple plug-in estimator. It is provably sample efficient, despite its simple design. Notably, we report sample complexity bounds on the performance of (MB-RS-QVI) under Q-value learning (Theorem 1) and policy learning (Theorem 2), both scaling as $\widetilde{\mathcal{O}}\Big(\frac{S^2 A}{\varepsilon^2(1-\gamma)^2|\beta|^2} e^{2|\beta|\frac{1}{1-\gamma}}\Big)$, in any discounted MDP with S states, A action, and discount factor γ, with $\widetilde{\mathcal{O}}$ hiding logarithmic factors. Here, β denotes the risk parameter (see Section 2 for precise definition), where $\beta > 0$ (resp. $\beta < 0$) corresponds to a risk-averse (resp. risk-seeking) agent. An interesting property of these bounds is the exponential dependence on the effective horizon $\frac{1}{1-\gamma}$, which is absent in the conventional (risk-neutral) RL, wherein $\beta = 0$. Another key contribution of this paper is to derive the first, to our knowledge, sample complexity lower bounds for the discounted RL under entropic risk measure. Our lower bounds establish that for both Q-value learning (Theorem 3) and policy learning (Theorem 4), one needs at least $\widetilde{\Omega}\Big(\frac{SA}{\varepsilon^2|\beta|^2} e^{|\beta|\frac{1}{1-\gamma}}\Big)$ samples to come ε-close to optimality. Interestingly, the derived lower bounds assert that exponential dependence on $\frac{|\beta|}{1-\gamma}$ in both sample complexities are unavoidable.

1.2 Related Work

Finite-sample guarantees for risk-neutral RL. There is a large body of papers studying provably-sample efficient learning algorithms in discounted MDPs. These papers consider a variety of settings, including the generative setting [27, 21, 1], the offline (or batch) setting [42, 34], and the online setting [48, 32]. In particular, in the generative setting – which we consider in this paper – some notable developments include, but not limited to [28, 21, 1, 33, 47, 52, 14, 26]. Among these, [21, 1, 47, 33, 26] present algorithms, attaining minimax-optimal sample complexity bounds, although some of these result do not cover the full range of ε. We also mention a line of work, comprising e.g. [40, 46], that investigate discounted RL in the generative setting but under distributional robustness. Let us also remark that some recent works – notably [56, 57] – study sample complexity of average-reward MDPs in the generative setting.

Finally, we mention that some studies consider adaptive sampling in the generative setting to account for the heterogeneity across the various state-action pairs in the MDP; see, e.g., [3, 54]. This line of works stand in contrast to the papers cited above that strive for optimizing the performance, in the worst-case sense, via uniformly sampling various state-action pairs.

Risk-sensitive RL. There exists a rich literature on decision making under a risk measure. We refer to [43, 36, 30] for some developments in bandits, where the performance is assessed via regret.

Extensions to episodic MDPs were pursued in a recent line of work, including [20, 35, 24], which establish near-optimal guarantees on regret. The two lines of work (on bandits and episodic MDPs) constitute the majority of work on risk-sensitive RL, thoroughly studying a variety of risk measures including CVaR, entropic risk, and entropic value-at-risk [2, 49].

Among the various risk measures, CVaR has arguably received a great attention; see, for instance, [17, 18, 13]. For instance, [18, 13] study episodic RL in the regret setting and present algorithms with sub-linear regret in tabular MDPs ([18]) and under function approximation ([13]), while [17] investigates the sample complexity analysis in the generative setting (similar to ours), We also provide a pointer to some work which deal with a class of measures called coherent risk measures that include CVaR as a special case. In this category, we refer to, for instance, [51], which studied policy gradient algorithms, and to [31] that investigates regret minimization in the episodic setting and with function approximation. In the case of infinite-horizon setting, the entropic risk measure is mostly considered in the undiscounted (i.e., the average-reward) setting; examples include [11, 39, 37, 38]. In contrast, little attention is paid to the discounted setting, which is mainly due to technical difficulties caused by discounting. The work [22] is among the few papers investigating solving discounted MDPs under the entropic risk.

Notations. For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. 1_A denotes the indicator function of an event A. Given a set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the probability simplex over $\mathcal{X}$. We use the convention that $\|\cdot\| := \|\cdot\|_\infty$ and explicitly use the subscript $\|\cdot\|_p$ when using p-norms for which $1 \le p < \infty$. We use $Z = \mathcal{S} \times \mathcal{A}$ to denote the set of all state-action pairs.

2 Background

2.1 Markov Decision Processes

We write the 6-tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \beta)$ as a finite, discounted infinite-horizon Markov decision process (MDP), where $\mathcal{S} = \{1, 2, ..., S\}$ is the finite state space of size $S := |\mathcal{S}|$, $\mathcal{A} = \{1, 2, ..., A\}$ is the finite action space of size $A := |\mathcal{A}|$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function, $\gamma \in (0, 1)$ is the discount factor, and $\beta \neq 0$ is the risk-parameter. The agent interacts with the MDP M as follows. At the beginning of the process, M is in some initial state $s_0 \in \mathcal{S}$. At each time $t \ge 0$, the agent is in state $s_t \in \mathcal{S}$ and decides on an action $a_t \in \mathcal{A}$ according to some rule. The MDP generates a reward $r_t := R(s_t, a_t)$ and a next-state $s_{t+1} \sim P(\cdot|s_t, a_t)$. The MDP moves to s_{t+1} when the next time slot begins, and this process continues ad infinitum. This process yields a growing sequence $(s_t, a_t, r_t)_{t \ge 0}$. The agent's goal is to maximize an objective function, as function of reward values $(r_t)_{t \ge 0}$ in expectation, which involves the two parameters γ and β; in addition to the discount factor γ that makes future rewards less valuable than the present ones, the risk-parameter β quantifies to what degree the agent seeks or avoids strategies that have more variability in the rewards obtained over time. To concretely define the objective function of the agent, we shall introduce some necessary concepts.

2.2 Entropic Risk Preferences

The entropic risk preferences can be justified by expected utility theory. Consider for $\beta \neq 0$ the class of utility functions $u(t) = \frac{1}{\beta}(1 - e^{-\beta t})$ defined for $t \in \mathbb{R}$. The utility u is supposed to describe the preferences of some economic agent in the form of how much utility $u(t)$ she derives from some monetary quantity $t \in \mathbb{R}$. For any bounded random variable $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, *the certainty equivalent* $u^{-1}(\mathbb{E}[u(X)]) = \frac{-1}{\beta}\log(\mathbb{E}[e^{-\beta X}])$ expresses the amount of money that would give the same utility as that of entering in the bet given by the random variable X. We thus define the functional $\Phi_\beta : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ by

$$\Phi_\beta(X) = -\frac{1}{\beta}\log(\mathbb{E}[e^{-\beta X}])$$

We note that there does not seem to be consensus on wether the functional is parametrized with β of $-\beta$. We follow this convention considering its widespread use in the actuarial literature [4], but we remark that some other lines of work appear to prefer the other parametrization. For the case $\beta \to 0$, we recover the risk-neutral case that is simply the expectation: $\lim_{\beta \to 0} \Phi_\beta(X) = \mathbb{E}[X]$. It is straightforward to see that Φ_β admits the following properties:

$$\Phi_\beta(X) \le \Phi_\beta(Y), \quad \text{for any } X \le Y, \tag{1}$$

$$\Phi_\beta(c) = c, \quad \text{for any } c \in \mathbb{R}, \tag{2}$$

$$\Phi_\beta(X) \le \mathbb{E}[X], \quad \text{for } \beta > 0, \tag{3}$$

$$\Phi_\beta(X) \ge \mathbb{E}[X] \quad \text{for } \beta < 0, \tag{4}$$

where properties (3)-(4) follow from Jensen's inequality. Using Φ_β as a measure of the preference for different random variables, it follows directly from (2)-(4) that $\Phi_\beta(X) \le \Phi_\beta(\mathbb{E}[X])$ for $\beta > 0$ and $\Phi_\beta(X) \ge \Phi_\beta(\mathbb{E}[X])$ for $\beta < 0$, which shows that $\beta > 0$ is associated with risk-aversion, while $\beta < 0$ is associated with risk-seeking behavior.

Applying risk in a stochastic dynamic process can be done in several ways and is thus more complicated than for a single-period problem. Two approaches to this end exist in the literature. The first and most intuitive one, which we name the *non-recursive* approach, is to apply the functional to the total discounted sum of rewards $\Phi_\beta\big(\sum_{t=0}^{\infty} \gamma^t r_t\big)$, which is well-defined under the bounded rewards assumption, i.e., $r_t \in [0,1]$ for all t. The other approach, which we may call the *recursive* approach, works by applying the risk at every step t where we give the details below. The non-recursive approach is probably most intuitive but has several drawbacks. Even though there is no obvious optimality equation for the non-recursive case, a solution to the planning problem has been proposed in [22]. In comparison, the planning problem is more straightforward with recursive risk-preferences due to the availability of an optimality equation that allows for simple value-iteration type algorithms. The recursive approach also guarantees the existence of an optimal stationary deterministic policy whereas with non-recursive risk preferences the optimal policy might not be time-consistent (see [25]). In this paper, we study the problem with recursive risk-preferences.

2.3 Value Function and Q-function

A bit of notation is required in order to define the state value function V (henceforth V-function) and state-action value function Q (henceforth Q-function) of a policy.

We follow the approach of [4] although as they prove that among all history dependent policies there exists an optimal stationary policy we will reduce the notation by considering only stationary policies from the outset.

Let $v \in \mathbb{R}^S$ and $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary deterministic policy. We then define $\rho_s^a : \mathbb{R}^S \to \mathbb{R}$ by

$$\rho_s^a(v) = -\frac{1}{\beta} \log\left(\mathbb{E}_{s' \sim P(\cdot|s,a)}[e^{-\beta v(s')}]\right) \tag{5}$$

and where we write ρ_s^π when $a = \pi(s)$, that is $\rho_s^\pi := \rho_s^{\pi(s)}$. We then define the operator $J_\pi : \mathbb{R}^S \to \mathbb{R}^S$ given by $J_\pi(v)(s) = r(s, \pi(s)) + \gamma \rho_s^\pi(v)$. The N-step total discounted utility $J_N(s,\pi)$ is defined as applying J_π recursively N-times to the 0-map, that is $J_N(s,\pi) := J_\pi^N(\mathbf{0})(s)$. Note that the outer-most iteration corresponds to the immediate time-step. By properties $(1)-(2)$ of Φ_β it follows that $J_N(s,\pi)$ is increasing in N and that $J_N(s,\pi) \le \frac{1}{1-\gamma}$ so the $N \to \infty$ limit exists and this limit is considered the value of state s under the policy π of the agent: $V^\pi(s) = \lim_{N\to\infty} J_N(s,\pi)$.

The problem of the agent is then for all initial states $s \in \mathcal{S}$ to find a policy π^* that solves $J(s,\pi^*) = \sup_\pi J(s,\pi)$. In [4] the authors consider a more general framework than is not restricted to finite MDPs or stationary policies but they prove that under some conditions (that are trivially fulfilled in the case of finite MDPs) there exists a stationary policy π^* that maximizes the state-value function for all states $s \in \mathcal{S}$ simultaneously. The authors only prove this in the risk-averse case $\beta > 0$ but the proof carries over verbatim in the case of $\beta < 0$. They also show that any optimal policy π^* that solves the agent satisfies the optimality equation

$$V^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) - \frac{\gamma}{\beta} \log\left(\mathbb{E}_{s' \sim P(\cdot|s,a)}[e^{-\beta V^*(s')}]\right)\right) \tag{6}$$

where V^* is the optimal V-function.

Also for any stationary deterministic policy π we have the Bellman recursion:

$$V^\pi(s) = R(s,\pi(s)) - \frac{\gamma}{\beta} \log\left(\mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}[e^{-\beta V^\pi(s')}]\right) \tag{7}$$

Since we are not interested only in the planning problem but that of learning, we also introduce the state-action value function Q. The approach is very similar to that of the value-function. Given π, we define the operator $L_\pi : \mathbb{R}^S \to \mathbb{R}^{S\times A}$ as follows: for all $v : \mathcal{S} \to \mathbb{R}$, $L_\pi(v)(s,a) = R(s,\pi(s)) + \gamma\rho_s^\pi(v)$. Further, we define the operator $L : \mathbb{R}^S \to \mathbb{R}^{S\times A}$ for all $v : \mathcal{S} \to \mathbb{R}$ as: $L(v)(s,a) = R(s,a) + \gamma\rho_s^a(v)$ We define the N-step total discounted utility of the state-action pair (s,a) under π as $L_N(s,a,\pi) := (L \circ J_\pi^{N-1}(\mathbf{0}))(s,a)$ and the limit is denoted $Q^\pi(s,a)$: $Q^\pi(s,a) = \lim_{N\to\infty} L_N(s,a,\pi)$. Although the authors do not consider state-action value functions in their paper, repeating the arguments of [4] it suffices to consider stationary policies when wanting to solve $\max_\pi Q^\pi(s,a)$ for all (s,a) and that the solution Q^* solves the optimality equation

$$Q^*(s,a) = R(s,a) - \frac{\gamma}{\beta}\log\left(\mathbb{E}_{s'\sim P(\cdot|s,a)}[e^{-\beta\max_{a'} Q^*(s',a')}]\right) \tag{8}$$

Similarly it is clear that the Q-functions satisfy the Bellman recursion relations

$$Q^\pi(s,a) = R(s,a) - \frac{\gamma}{\beta}\log\left(\mathbb{E}_{s'\sim P(\cdot|s,a)}[e^{-\beta V^\pi(s')}]\right) \tag{9}$$

2.4 Learning Performance

We consider two types of RL algorithms $\mathcal{U}$, namely those that output a Q-function $Q_T^{\mathcal{U}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and those that output a policy $\pi_T^{\mathcal{U}} : \mathcal{S} \to \mathcal{A}$ using T transition samples. Note that any algorithm that outputs a Q-function also outputs a policy, namely the one obtained by acting greedy with respect to the Q-function. There are also ways to obtain a Q-function from a policy. There is however no canonical way to do this as the algorithm cannot simply output $Q^{\pi^{\mathcal{U}}}$ since the algorithm does not have access to the true MDP. The way we evaluate the quality of an algorithm that outputs a Q-function is by $\|Q^* - Q_T^{\mathcal{U}}\|$. For algorithms that instead outputs a policy we evaluate the policy in terms of $\|V^* - V^{\pi_T^{\mathcal{U}}}\|$. Often we will suppress T from the notation.

Definition 1 ((ε,δ)-correctness)**.** *We say that a algorithm $\mathcal{U}$ that outputs a Q-function $Q^{\mathcal{U}}$ is (ε,δ)-correct on a set of MDPs $\mathbb{M}$ if $\mathbb{P}(\|Q^* - Q^{\mathcal{U}}\| \le \epsilon) \ge 1-\delta$ for all $M \in \mathbb{M}$. Similarly we say that an algorithm $\mathcal{U}$ that outputs a policy $\pi^{\mathcal{U}}$ is is (ε,δ)-correct on a set of MDPs $\mathbb{M}$ if $\mathbb{P}(\|V^* - V^{\pi^{\mathcal{U}}}\| \le \epsilon) \ge 1-\delta$ for all $M \in \mathbb{M}$.*

3 Model-Based Risk-Sensitive Q-Value Iteration

In this section we describe the model-based value iteration algorithm which aims at finding the optimal Q-function Q^*. We then give an upper bound on the total number of calls to the generative model needed in order for this algorithm to be $(\varepsilon,\delta)-$correct. The model based approach is based on working on a model MDP which can disagree with the true MDP because it does not use the true transition probabilities but an estimate of the transition functions obtained from n calls to each of the state-action pairs in $\mathcal{S}\times\mathcal{A}$ as described in the algorithm below.

Algorithm 1: Model estimation

Input: Generative model P
Output: Model estimate $\widehat{P}$

```
Function EstimateModel(n):
    ∀ (s,z) ∈ S × Z : m(s,z) = 0
    for each z ∈ Z do
        for i = 1,2,...,n do
            s ~ P(·|z)
            m(s,z) := m(s,z) + 1
        end
        ∀s ∈ S : P̂(s,z) = m(s,z)/n
    end
    return P̂
```

The model-based approach we describe is general in the sense that any oracle that for any $\varepsilon > 0$ can find an ε-optimal policy can be used. We prove the existence of one such oracle in the form of a

Q-value iteration very like the one from the classical risk-neutral setting. The proof is very similar to Part (a) in Theorem 3.1 in [4] but is nevertheless provided in Appendix B.

Lemma 1 (Q-value iteration)**.** *Fix a map $\pi : \mathcal{A} \to \mathcal{S}$. We then define the operators $\mathcal{T}^\pi, \mathcal{T} : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ which for $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is given by*

$$(\mathcal{T}f)(s,a) = R(s,a) + \frac{-\gamma}{\beta}\log\left(\sum_{s'} P(s'|s,a)e^{-\beta \max_{a'} f(s',a')}\right)$$

$$(\mathcal{T}^\pi f)(s,a) = R(s,a) + \frac{-\gamma}{\beta}\log\left(\sum_{s'} P(s'|s,a)e^{-\beta f(s',\pi(s'))}\right)$$

The operators $\mathcal{T}, \mathcal{T}^\pi$ are γ-contractions with respect to the max-norm, i.e. for value-functions f_1 and f_2 it holds that $\|\mathcal{T}f_1 - \mathcal{T}f_2\| \leq \gamma\|f_1 - f_2\|$ and $\|\mathcal{T}^\pi f_1 - \mathcal{T}^\pi f_2\| \leq \gamma\|f_1 - f_2\|$.

The above lemma is the basis for the Q-value iteration algorithm:

Algorithm 2: RS-QVI(M, k)

Input: MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \beta)$ and number of iterations k.
Output: Estimate Q_k of optimal Q-function Q^*

Initialization: $\forall(s,a)$ set $Q(s,a) = 0$
for $j = 0, 1, ..., k-1$ **do**
  for *all* $s \in \mathcal{S}$ **do**
    $\pi_j(s) = \operatorname{argmax}_{a\in\mathcal{A}} Q_j(s,a)$
    for *all* $a \in \mathcal{A}$ **do**
      $\mathcal{T}Q_j(s,a) = R(s,a) - \frac{\gamma}{\beta}\log(\mathbb{E}_{s'\sim P(\cdot|s,a)}[e^{-\beta Q_j(s,\pi_j(s))}])$
      $Q_{j+1}(s,a) = \mathcal{T}Q_j(s,a)$
    end
  end
end
$\forall s \in \mathcal{S} : \pi_k(s) = \operatorname{argmax}_{a\in\mathcal{A}} Q_k(s,a)$
return Q_k and π_k

The next lemma shows that if we choose k large enough in the RS-QVI algorithm, we can obtain Q_k and V^{π_k} that are as close to Q^* and V^* as we desire. The proof is postponed to Appendix B.

Lemma 2. *Fix $\varepsilon > 0$. Then there exists some $k(\varepsilon)$ such that if the number of iterations in RS-QVI uses more than $k(\varepsilon)$ iterations then, the output of the Algorithm 2 (RS-QVI) satisfies $\|Q_k - Q^*\| < \varepsilon$ and $\|V^{\pi_k} - V^*\| < \varepsilon$*

Using Algorithm 2, we introduce the MB-RS-QVI algorithm, which consists in building an empirical model $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, \gamma, \beta)$ via calling the generative model n times – namely, $\widehat{P} =$ `EstimateModel`(n) – and then solving it via RS-QVI.

3.1 Analysis of MB-RS-QVI

With RS-QVI in place we need a set of lemmas for the analysis of the sample-complexity of the model based RS-QVI algorithm.

An important result for the analysis is a risk-sensitive version of the simulation lemma [29, 48], which describes how different two Q-functions for the same policy are in two different MDPs that differ only slightly in their rewards and transition functions. The proof is postponed to Appendix B.

Lemma 3 (Simulation Lemma with Entropic Risk)**.** *Consider two MDPs $M_1 = (S, A, P_1, R, \gamma, \beta)$ and $M_2 = (S, A, P_2, R, \gamma, \beta)$ differing only in their transition functions. Fix a stationary policy π, and let Q_1^π and Q_2^π be respective the corresponding Q-functions of π in M_1 and M_2. If $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_1(\cdot|s,a) - P_2(\cdot|s,a)\|_1 \leq \tau$. Then, it holds that*

$$\|V_1^\pi - V_2^\pi\| \leq \|Q_1^\pi - Q_2^\pi\| \leq \frac{\gamma}{1-\gamma}\frac{e^{|\beta|\frac{1}{1-\gamma}}}{|\beta|}\tau$$

It then follows that if for some $\varepsilon > 0$ it holds that $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_1(\cdot|s,a) - P_2(\cdot|s,a)\|_1 \leq \varepsilon\frac{1-\gamma}{\gamma}|\beta|e^{-|\beta|\frac{1}{1-\gamma}}$, then $\|Q_1^\pi - Q_2^\pi\| \leq \varepsilon$. Using the decompositions from Lemma 5 we get for any $s \in \mathcal{S}$ that

$$V^{\pi_k}(s) \geq V^*(s) - \|V^{\pi_k} - \widehat{V}^{\pi_k}\| - \|\widehat{V}^{\pi^*} - V^*\| - \|\widehat{V}^{\pi_k} - \widehat{V}^*\| \tag{10}$$

and similarly we get for any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ that

$$Q_k(s,a) \geq Q^*(s,a) - \|\widehat{Q}^{\pi^*} - Q^*\| - \|Q_k - \widehat{Q}^*\| \tag{11}$$

The last of the distances on the right-hand side of (10) and (11) can be made arbitrarily small by any optimization oracle for the problem. One such is value-iteration using the model MDP as demonstrated by Lemma 2. Making these terms small enough is thus purely a computational matter and not a statistical one. We thus focus on bounding the remaining distances on the right-hand sides.

Lemma 4. *Fix $\tau > 0$. If every state-action pair has been tried at least $m = 8\frac{S+\log(\frac{SA}{\delta})}{\tau^2}$ times, then it holds that $\max_{(s,a)} \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \leq \tau$ with probability at least $1-\delta$.*

Proof. Using the Weissman inequality [53], any confidence interval for a state-action pair that have been tried m times have size $2\sqrt{2[\log(2^S-2) - \log(\delta_P)]/m}$. Setting this size to be smaller than τ and solving for m, we find that

$$m \geq \frac{8}{\tau^2}\Big(\log(2^S-2) - \log(\delta_P)\Big). \tag{12}$$

Noting that $8\big(S + \log(1/\delta_P)\big)/\tau^2 \geq 8[\log(2^S-2) - \log(\delta_P)]/\tau^2$ and substituting $\delta_P = \frac{\delta}{SA}$, the result follows by the union bound since each of the SA confidence balls contain $P(\cdot|s,a)$ with probability at least $1-\delta$. □

Theorem 1. *There exist a universal constant c such that for any $\varepsilon > 0, \delta \in (0,1)$ and any MDP M with S states and A actions if the learner is allowed to make*

$$T = c\frac{SA(S+\log(SA/\delta))}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2} \tag{13}$$

calls to the generative model then $\mathbb{P}(\|Q^* - Q_k\| \leq \varepsilon) \geq 1-\delta$

Proof. For any $\varepsilon > 0$, we can get $\|Q_k - \widehat{Q}^*\| < \varepsilon/2$ using enough iterations of the optimization oracle by Lemma 2. The term $\|\widehat{Q}^{\pi^*} - Q^*\|$ can also be made smaller than $\varepsilon/2$ by the simulation lemma if $\max_{(s,a)} \|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 < \tau$ where $\tau = \frac{\varepsilon}{2}\left[\frac{\gamma}{1-\gamma}\frac{e^{|\beta|\frac{1}{1-\gamma}}}{|\beta|}\right]^{-1}$ which can be ensured with probability larger than $1-\delta$ by picking $m = 8\frac{S+\log(\frac{SA}{\delta})}{\tau^2}$. Using that the total calls to the generative model is $T = SAm$ and substituting in the value for τ we can ensure for all (s,a) that $Q^\pi(s,a) > Q^*(s,a) - \varepsilon$ with probability larger than $1-\delta$ by using a total number of samples

$$T = 32\left[S + \log\left(\frac{SA}{\delta}\right)\right]\frac{SA}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2} \tag{14}$$

□

Theorem 2. *There exist a universal constant c such that for any $\varepsilon > 0, \delta \in (0,1)$ and any MDP M with S states and A actions if the learner is allowed to make*

$$T = c\frac{SA(S+\log(SA/\delta))}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2} \tag{15}$$

calls to the generative model then $\mathbb{P}(\|V^* - V^{\pi_k}\| \leq \varepsilon) \geq 1-\delta$.

Proof. By Lemma 2 we can pick k large enough so that $\|\hat{V}^{\pi_k} - \hat{V}^*\| \leq \varepsilon/3$. The two terms $\|V^{\pi_k} - \widehat{V}^{\pi_k}\|$ and $\|\hat{V}^{\pi^*} - V^*\|$ can simultaneously be made smaller than $\varepsilon/3$ by sampling each state-action pair $m = 8\frac{S+\log(\frac{SA}{\delta})}{\tau^2}$ times where $\tau = \frac{\varepsilon}{3}\left[\frac{\gamma}{1-\gamma}\frac{e^{|\beta|\frac{1}{1-\gamma}}}{|\beta|}\right]^{-1}$ such that if we sample

$$T = 72\left[S + \log\left(\frac{SA}{\delta}\right)\right]\frac{SA}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2} \tag{16}$$

times in total, then we can ensure that for all s simultaneously it holds that $V^{\pi_k}(s) > V^* - \varepsilon$. $\square$

Remark 1. *The overall proof technique for the upper bounds can also be used to derive upper bounds in the risk-neutral case by using the classical simulation lemma [29]. Doing this will however yield a suboptimal upper bound. State-of-the-art techniques for proving optimal upper bounds in the risk-neutral case are however critically exploiting linearity of the expectation operator and are thus not readily available in the risk-sensitive case due to the non-linearity of the entropic risk measure.*

4 Sample Complexity Lower Bounds

In this section, we provide two sample complexity lower bounds. The first one, presented in Theorem 3, concerns the sample complexity of learning the optimal Q-value function Q^*, whereas the second one, in Theorem 4, is on learning an optimal policy π^*. The proofs are both postponed to Appendix D.

Theorem 3 (Lower bound for learning Q^*)**.** *There exist constants $c_1, c_2 > 0$ such that for any RL algorithm $\mathcal{U}$ that outputs a Q-function $Q^{\mathcal{U}}$ and any $\delta \in (0, \frac{1}{4})$ and $\varepsilon \in \left(0, \frac{1}{40}\frac{\gamma}{|\beta|}(1 - e^{-|\beta|\frac{1}{1-\gamma}})\right)$, the following holds: if the total number T of transitions satisfies*

$$T \leq \frac{SA\gamma^2}{c_1\varepsilon^2}\frac{(e^{|\beta|\frac{1}{1-\gamma}} - 3)}{|\beta|^2}\log\left(\frac{SA}{c_2\delta}\right),$$

then there is some MDP M with S states and A actions for which $\mathbb{P}(\|Q^*_M - Q^{\mathcal{U}}_T\| > \varepsilon) \geq \delta$.

Theorem 4 (Lower bound for learning π^*)**.** *There exist constants $c_1, c_2 > 0$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $Q^{\mathcal{U}}$ and any $\delta \in (0, \frac{1}{4})$ and $\varepsilon \in \left(0, \frac{1}{50}\frac{\gamma}{|\beta|}(1 - e^{-|\beta|\frac{1}{1-\gamma}})\right)$, it holds that if the total number T of transitions satisfies*

$$T \leq \frac{SA\gamma^2}{c_1\varepsilon^2}\frac{(e^{|\beta|\frac{1}{1-\gamma}} - 3)}{|\beta|^2}\log\left(\frac{S}{c_2\delta}\right),$$

then there is some MDP M with S states and A actions for which $\mathbb{P}(\|V^*_M - V^{\pi^{\mathcal{U}}_T}\| > \varepsilon) \geq \delta$.

The lower bounds in Theorems 3-4 establish that an exponential dependence of the sample complexity on the effective horizon $\frac{1}{1-\gamma}$; more precisely, they assert that a scaling of $e^{|\beta|\frac{1}{1-\gamma}}$ is unavoidable in both Q-learning and policy-learning. Recalling the sample complexity bounds of MB-RS-QVI (Theorems 1-2), we observe a similar exponential dependence. However, there remains a gap of order $\frac{1}{(1-\gamma)^2}e^{|\beta|\frac{1}{1-\gamma}}$. This remains as an interesting open question as to whether closing this gap can be done via a more elegant analysis of MB-RS-QVI or it calls for more novel algorithmic ideas.

The proofs of lower bounds are given in Section D in the appendix, but we briefly sketch below the ideas here for the case where the algorithm outputs a Q-function; the other case is proven using quite similar ideas.

We consider a class of hard-to-learn MDPs, admitting a structure similar to that used in the lower bound derivation of [5]. In such MDPs, the state-action pairs are organized into triplets $\{z^0, z^G, z^B\}$, where z^G yields a reward of 1, z^B yields a reward of 0, and both z^G and z^B are absorbing. In z_0, the reward is also zero and the transition function satisfies $P(z^G|z^0) = q$ and $P(z^B|z^0) = 1-q$, for some $q > 0$. This construction critically allows us to calculate explicitly $Q^*(z)$ for a given parameter

q and for two different MDPs M_0, M_1 in the class where $q_0 = p$ and $q_1 = p + \alpha$ for appropriately chosen values of p and α we are able to ensure that $Q^*_{M_1}(z) - Q^*_{M_0}(z) > 2\varepsilon$ which means that any specific algorithmic output $Q^{\mathcal{U}}(z)$ cannot be ε-close to both $Q^*_{M_1}(z)$ and $Q^*_{M_0}(z)$. We then show by a likelihood ratio argument that any algorithm $\mathcal{U}$ that is (ε, δ)-correct on M_0, i.e. that $\mathbb{P}_0(|Q^*_{M_0}(z) - Q^{\mathcal{U}}(z)| \leq \varepsilon) > \delta$, will also satisfy that $\mathbb{P}_1(|Q^*_{M_0}(z) - Q^{\mathcal{U}}(z)| \leq \varepsilon) > \delta$ provided that the algorithm does not try out z enough times on M_0 and exactly because $Q^*_{M_1}(z) - Q^*_{M_0}(z) > 2\varepsilon$, the event $\{|Q^*_{M_0}(z) - Q^{\mathcal{U}}(z)| \leq \varepsilon\}$ is disjoint from the event on being ε-close to $Q^*_{M_1}$. The final part of the proof is to exploit that the different triplets of state-action pairs contain no information about each other which allows for an independence argument for the estimation of $Q^{\mathcal{U}}(z)$ and $Q^{\mathcal{U}}(z')$ for $z \neq z'$. While doing this analysis, we fix an inaccuracy in the proof of Lemma 17 in [21] that arises where they lower-bound the likelihood ratio of two Bernoulli random variables with biases $p \geq \frac{1}{2}$ and $p + \alpha$ on a high probability event. We also mention that we extend the result to hold for $p < \frac{1}{2}$.

Remark 2. *It is worth remarking that the best lower bound in the risk-neutral setting is derived in [21] using a richer construction than above. However, with a risk-sensitive learning objective, the optimal state-action value function in the construction of [21] does not admit an analytical solution, which is needed for the delicate tuning of the transition probabilities.*

5 Conclusion and Future Works

We have studied the sample-complexity of learning the optimal Q-function and that of learning an optimal policy in finite discounted MDPs, where the agent has recursive risk-preferences given by the entropic risk measure and has access to a generative model. We introduced an algorithm, called MB-RS-QVI, and derived PAC-type bounds on its sample complexity for both learning which have derived bounds from analyzing the MB-RS-QVI algorithm that uses the model given by the plug-in estimator from samples generated by a simulator. We also derive lower bounds that show that dependence on $e^{|\beta|\frac{1}{1-\gamma}}$ is unavoidable in both cases. The bounds that we derive on the sample complexity of learning the optimal Q-value are of order

$$\mathcal{O}\left(\left(S + \log(SA/\delta)\right)\frac{SA}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2}\right), \qquad \Omega\left(\frac{SA}{\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}}{|\beta|^2}\log(SA/\delta)\right) \tag{17}$$

while the bounds we derive on the sample complexity of learning an optimal policy are of order

$$\mathcal{O}\left(\left(S + \log(SA/\delta)\right)\frac{SA}{\varepsilon^2(1-\gamma)^2}\frac{e^{2|\beta|\frac{1}{1-\gamma}}}{|\beta|^2}\right), \qquad \Omega\left(\frac{SA}{\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}}{|\beta|^2}\log(S/\delta)\right) \tag{18}$$

These constitute the first bounds, to our knowledge, on the sample complexities of entropic risk-sensitive agents in the discounted MDP setting. The upper and lower bounds derived in this paper leave open gaps of sizes, at most, $\frac{S}{(1-\gamma)^2}e^{|\beta|\frac{1}{1-\gamma}}$. Since the constructions in the lower bounds are not the ones used in the tightest lower bounds derived in the risk-neutral setting, one possibility is that the lower bounds can be improved by considering a more carefully chosen set of hard-to-learn MDPs with the challenge being to control the gap in V-values or Q-values under different parameters. Also since the plugin-estimator model-based QVI algorithm is provably optimal in the risk-neutral setting, we believe that this might also be the case for risk-sensitive agents but that more tools are needed to develop a more careful analysis. Another future direction is that of developing model-free algorithms for this setting and analyzing their statistical efficiency. Another interesting research direction is to consider function approximation, as in [55]. As other future directions, one may consider more complicated RL settings such as offline RL [42], where data is collected under a fixed (but unknown) behavior policy, and online RL [48, 32], where the agent's learned policy impacts the data collection process. And finally one may also consider the problem where the learner have non-recursive risk-preferences instead.

6 Conflicts of interests

The authors claim no conflict of interests.

References

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

[2] Amir Ahmadi-Javid. "Entropic value-at-risk: A new coherent risk measure". In: *Journal of Optimization Theory and Applications* 155 (2012), pp. 1105–1123.

[3] Aymen Al Marjani and Alexandre Proutiere. "Adaptive sampling for best policy identification in Markov decision processes". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7459–7468.

[4] Hubert Asienkiewicz and Anna Jaśkiewicz. "A note on a new class of recursive utilities in Markov decision processes". In: *Applicationes Mathematicae* 44 (2017), pp. 149–161.

[5] Mohammad Gheshlaghi Azar et al. "Reinforcement learning with a near optimal rate of convergence". In: (2011).

[6] Nicole Bäuerle and Alexander Glauner. "Markov decision processes with recursive risk measures". In: *European Journal of Operational Research* 296.3 (2022), pp. 953–966.

[7] Nicole Bäuerle and Jonathan Ott. "Markov decision processes with average-value-at-risk criteria". In: *Mathematical Methods of Operations Research* 74 (2011), pp. 361–379.

[8] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[9] Tomasz R Bielecki and Stanley R Pliska. "Risk-sensitive dynamic asset management". In: *Applied Mathematics and Optimization* 39 (1999), pp. 337–360.

[10] Lorenzo Bisi et al. "Risk-averse policy optimization via risk-neutral policy optimization". In: *Artificial Intelligence* 311 (2022), p. 103765.

[11] Vivek S Borkar and Sean P Meyn. "Risk-sensitive optimal control for Markov decision processes with monotone cost". In: *Mathematics of Operations Research* 27.1 (2002), pp. 192–209.

[12] Daniel Brown, Scott Niekum, and Marek Petrik. "Bayesian robust optimization for imitation learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2479–2491.

[13] Yu Chen et al. "Provably Efficient Iterated CVaR Reinforcement Learning with Function Approximation and Human Feedback". In: *The Twelfth International Conference on Learning Representations*. 2024.

[14] Zaiwei Chen et al. "Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8223–8234.

[15] Yinlam Chow and Mohammad Ghavamzadeh. "Algorithms for CVaR optimization in MDPs". In: *Advances in neural information processing systems* 27 (2014).

[16] Erick Delage and Shie Mannor. "Percentile optimization for Markov decision processes with parameter uncertainty". In: *Operations research* 58.1 (2010), pp. 203–213.

[17] Zilong Deng, Simon Khan, and Shaofeng Zou. "Near-Optimal Sample Complexity for Iterated CVaR Reinforcement Learning with a Generative Model". In: *AISTATS*. 2025.

[18] Yihan Du, Siwei Wang, and Longbo Huang. "Provably Efficient Risk-Sensitive Reinforcement Learning: Iterated CVaR and Worst Path". In: *The Eleventh International Conference on Learning Representations*. 2023.

[19] Damien Ernst et al. "Clinical data based optimal STI strategies for HIV: A reinforcement learning approach". In: *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE. 2006, pp. 667–672.

[20] Yingjie Fei et al. "Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22384–22395.

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

[22] Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. "Entropic risk optimization in discounted MDPs". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 47–76.

[23] Ronald A Howard and James E Matheson. "Risk-sensitive Markov decision processes". In: *Management science* 18.7 (1972), pp. 356–369.

[24] Xiaoyan Hu and Ho-fung Leung. "A tighter problem-dependent regret bound for risk-sensitive reinforcement learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 5411–5437.

[25] Stratton C Jaquette. "A utility criterion for Markov decision processes". In: *Management Science* 23.1 (1976), pp. 43–49.

[26] Yujia Jin et al. "Truncated variance reduced value iteration". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 117481–117508.

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

[28] Michael Kearns and Satinder Singh. "Finite-sample convergence rates for Q-learning and indirect algorithms". In: *Advances in neural information processing systems* 11 (1998).

[29] Michael Kearns and Satinder Singh. "Near-optimal reinforcement learning in polynomial time". In: *Machine learning* 49 (2002), pp. 209–232.

[30] Najakorn Khajonchotpanya, Yilin Xue, and Napat Rujeerapaiboon. "A revised approach for risk-averse multi-armed bandits under CVaR criterion". In: *Operations Research Letters* 49.4 (2021), pp. 465–472.

[31] Thanh Lam et al. "Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation". In: *The Eleventh International Conference on Learning Representations*. 2022.

[32] Tor Lattimore and Marcus Hutter. "Near-optimal PAC bounds for discounted MDPs". In: *Theoretical Computer Science* 558 (2014), pp. 125–143.

[33] Gen Li et al. "Breaking the sample size barrier in model-based reinforcement learning with a generative model". In: *Advances in neural information processing systems* 33 (2020), pp. 12861–12872.

[34] Gen Li et al. "Settling the sample complexity of model-based offline reinforcement learning". In: *The Annals of Statistics* 52.1 (2024), pp. 233–260.

[35] Hao Liang and Zhiquan Luo. "Regret bounds for risk-sensitive reinforcement learning with lipschitz dynamic risk measures". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 1774–1782.

[36] Odalric-Ambrym Maillard. "Robust risk-averse stochastic multi-armed bandits". In: *Algorithmic Learning Theory: 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings 24*. Springer. 2013, pp. 218–233.

[37] Alexandre Marthe, Aurélien Garivier, and Claire Vernade. "Beyond average return in markov decision processes". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 56488–56507.

[38] Alexandre Marthe et al. "Efficient Risk-sensitive Planning via Entropic Risk Measures". In: *arXiv preprint arXiv:2502.20423* (2025).

[39] Gergely Neu, Anders Jonsson, and Vicenç Gómez. "A unified view of entropy-regularized markov decision processes". In: *arXiv preprint arXiv:1705.07798* (2017).

[40] Kishan Panaganti and Dileep Kalathil. "Sample complexity of robust reinforcement learning with a generative model". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 9582–9602.

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

[42] Paria Rashidinejad et al. "Bridging offline reinforcement learning and imitation learning: A tale of pessimism". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11702–11716.

[43] Amir Sani, Alessandro Lazaric, and Rémi Munos. "Risk-aversion in multi-armed bandits". In: *Advances in neural information processing systems* 25 (2012).

[44] Maria Grazia Scutella and Raffaella Recchia. "Robust portfolio asset allocation and risk measures". In: *Annals of Operations Research* 204.1 (2013), pp. 145–169.

[45] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. SIAM, 2021.

[46] Laixi Shi et al. “The curious price of distributional robustness in reinforcement learning with a generative model”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 79903–79917.

[47] Aaron Sidford et al. “Variance reduced value iteration and faster algorithms for solving Markov decision processes”. In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. 2018, pp. 770–787.

[48] Alexander L Strehl and Michael L Littman. “An analysis of model-based interval estimation for Markov decision processes”. In: *Journal of Computer and System Sciences* 74.8 (2008), pp. 1309–1331.

[49] Xihong Su, Marek Petrik, and Julien Grand-Clément. “Risk-averse Total-reward MDPs with ERM and EVaR”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 19. 2025, pp. 20646–20654.

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

[51] Aviv Tamar et al. “Policy gradient for coherent risk measures”. In: *Advances in neural information processing systems* 28 (2015).

[52] Mengdi Wang. “Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time”. In: *Mathematics of Operations Research* 45.2 (2020), pp. 517–546.

[53] Tsachy Weissman et al. “Inequalities for the L1 deviation of the empirical distribution”. In: *Hewlett-Packard Labs, Tech. Rep* (2003), p. 125.

[54] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. “Almost horizon-free structure-aware best policy identification with a generative model”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[55] Dongruo Zhou, Jiafan He, and Quanquan Gu. “Provably efficient reinforcement learning for discounted mdps with feature mapping”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12793–12802.

[56] Matthew Zurek and Yudong Chen. “Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[57] Matthew Zurek and Yudong Chen. “The Plugin Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis”. In: *36th International Conference on Algorithmic Learning Theory*. 2025.

Appendix

Table of Contents

A Technical Lemmas

Recall that Q_k is the Q-function output by the algorithm after k iterations, π_k is the greedy policy with respect to Q_k, and that π^* is an optimal policy of the true MDP M.

The first lemma establishes a decomposition result for MB-RS-QVI, whose proof follows very similar lines to the proof of Lemma 3 in [1].

Lemma 5. *For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$Q_k(s,a) \geq Q^*(s,a) - \|Q_k - \widehat{Q}^*\| - \|\widehat{Q}^{\pi^*} - Q^*\| .$$

Further, for any state $s \in \mathcal{S}$,

$$V^{\pi_k}(s) \geq Q^*(s) - \|V^{\pi_k} - \hat{V}^{\pi_k}\| - \|\hat{V}^{\pi_k} - \hat{V}^*\| - \|\hat{V}^{\pi^*} - V^*\| .$$

Proof. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned} Q_k(s,a) - Q^*(s,a) &= Q_k(s,a) - \widehat{Q}^*(s,a) + \widehat{Q}^*(s,a) - Q^*(s,a) \\ &\geq Q_k(s,a) - \widehat{Q}^*(s,a) + \widehat{Q}^{\pi^*}(s,a) - Q^*(s,a) \\ &\geq -\|Q_k - \widehat{Q}^*\| - \|\widehat{Q}^{\pi^*} - Q^*\| . \end{aligned}$$

Similarly, for any $s \in \mathcal{S}$, we have

$$\begin{aligned} V^{\pi_k}(s) - V^*(s) &= V^{\pi_k}(s) - \hat{V}^{\pi_k}(s) + \hat{V}^{\pi_k}(s) - \hat{V}^*(s) + \hat{V}^*(s) - V^*(s) \\ &\geq V^{\pi_k}(s) - \hat{V}^{\pi_k}(s) + \hat{V}^{\pi_k}(s) - \hat{V}^*(s) + \hat{V}^{\pi *}(s) - V^*(s) \\ &\geq -\|V^{\pi_k} - \hat{V}^{\pi_k}\| - \|\hat{V}^{\pi_k} - \hat{V}^*\| - \|\hat{V}^{\pi^*} - V^*\| , \end{aligned}$$

and the lemma follows. ☐

Next, we present two lemmas that collect a few useful inequalities. Some of these may be standard results, but for concreteness, we collect them here.

Lemma 6. *It holds that*

$$\begin{aligned} &\log(1-x) \geq -x - x^2 + x^3 && \forall x \in [0, \frac{1}{5}] \\ &\log(1-x) \geq -x - 2x && \forall x \in [0, \frac{1}{2}] \\ &\log(1+x) \geq x - x^2 && \forall x \in [0, \infty) \\ &\log(1+x) \geq \frac{x}{2} && \forall x \in [0, 1] . \end{aligned}$$

Proof. We only prove the first claim, as the rest could be proven using the technique after some elementary calculations.

Let $f(x) = (1-x)$ and $g(x) = -x - x^2 + x^3$. It holds that $f(0) = g(0)$, and since we have $f'(x) = \frac{1}{1-x}$ and $g'(x) = -1 - 2x + 3x^2$, it follows easily that

$$f'(x) \geq g'(x) \Leftrightarrow 0 \leq x(1 - 5x + 3x^2)$$

where the inequality is satisfied for all $x \in [0, \frac{5-\sqrt{13}}{6}] \subseteq [0, \frac{1}{5}]$. The result then follows from the fundamental theorem of calculus. ☐

Lemma 7. *Let $\alpha > 1$. For any $x \in [0, \frac{1}{\alpha}]$, it holds that*

$$1 - (1-x)^\alpha \geq \frac{x\alpha}{2} .$$

Proof. Define $f(x) = 1 - (1-x)^\alpha - \frac{x\alpha}{2}$. Since $f''(x) = -\alpha(\alpha-1)(1-x)^{\alpha-2} < 0$, f is strictly concave. Further, since $f(0) = 0$ and $f(\frac{1}{\alpha}) = \frac{1}{2}(1 - \frac{1}{\alpha})^\alpha > \frac{1}{2} - \frac{1}{e} > 0$, f is positive on the interval $[0, \frac{1}{\alpha}]$ and the result follows. ☐

B Analysis of MB-RS-QVI: Missing Proofs

B.1 Proof of Lemma 1

Proof. We only give the proof for $\mathcal{T}$ as the claim for $\mathcal{T}^\pi$ could be proven using extremely similar lines.

Consider two maps $Q : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ and $W : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, and let $Q' = \mathcal{T}Q$ and $W' = \mathcal{T}W$ be their respective $\mathcal{T}$-transforms. Let (s,a) be any pair such that $|Q'(s,a) - W'(s,a)| = \|Q' - W'\|_\infty$, and assume without loss of generality that $Q'(s,a) \geq W'(s,a)$. Further, define

$$V(s) := \max_a Q(s,a)\,, \qquad X(s) := \max_a W(s,a)\,.$$

Assuming that $\beta > 0$ (the case $\beta < 0$ is completely similar), we then have

$$\begin{aligned}
\|Q' - W'\| &= Q'(s,a) - W'(s,a) \\
&= -\frac{\gamma}{\beta}\log\left(\sum_{s'} P(s'|s,a)e^{-\beta V(s')}\right) + \frac{\gamma}{\beta}\log\left(\sum_{s'} P(s'|s,a)e^{-\beta X(s')}\right) \\
&= \frac{\gamma}{\beta}\log\left(\frac{\sum_{s'} P(s'|s,a)e^{-\beta X(s')}}{\sum_{s'} P(s'|s,a)e^{-\beta V(s')}}\right) \\
&= \frac{\gamma}{\beta}\log\left(\frac{\sum_{s'} P(s'|s,a)e^{-\beta V(s') + \beta(V(s') - X(s'))}}{\sum_{s'} P(s'|s,a)e^{-\beta V(s')}}\right) \\
&\leq \frac{\gamma}{\beta}\log\left(\frac{\sum_{s'} P(s'|s,a)e^{-\beta V(s') + \beta\|V - X\|_\infty}}{\sum_{s'} P(s'|s,a)e^{-\beta V(s')}}\right) \\
&= \frac{\gamma}{\beta}\log\left(e^{\beta\|V-X\|}\right) \\
&= \gamma\|V - X\| \\
&\leq \gamma\|Q - W\|\,,
\end{aligned}$$

and the lemma follows. □

B.2 Proof of Lemma 2

Proof. By Lemma 1, we have that $\mathcal{T}$ is a γ-contraction and that Q^* is its unique fixed point. We thus have $\|Q_k - Q^*\| = \|\mathcal{T}Q_{k-1} - \mathcal{T}Q^*\| \leq \gamma\|Q_{k-1} - Q^*\|$. Applying this inequality k times yields

$$\|Q_k - Q^*\| \leq \gamma^k\|Q_0 - Q^*\| \leq \frac{\gamma^k}{1-\gamma}\,.$$

Solving $\frac{\gamma^k}{1-\gamma}$ for k, we get that if $k > \frac{\log(\frac{1}{(1-\gamma)\varepsilon})}{\log(1/\gamma)}$, then $\|Q_k - Q^*\| < \varepsilon$, thus proving the first claim.

To show the other claim, we start by noting that $\|V^{\pi_k} - V^*\| \leq \|Q^{\pi_k} - Q^*\|$. Note also that by design we have that $\mathcal{T}^{\pi_k}Q^{\pi_k} = Q^{\pi_k}$ and that $\mathcal{T}Q_k = \mathcal{T}^{\pi_k}Q_k$. Thus,

$$\|Q^{\pi_k} - Q^*\| \leq \|Q^{\pi_k} - Q_k\| + \|Q_k - Q^*\|\,. \tag{19}$$

The first term in the right-hand side is bounded as follows:

$$\begin{aligned}
\|Q^{\pi_k} - Q_k\| &= \|\mathcal{T}^{\pi_k}Q^{\pi_k} - Q_k\| \\
&\leq \|\mathcal{T}^{\pi_k}Q^{\pi_k} - \mathcal{T}Q_k\| + \|\mathcal{T}Q_k - Q_k\| \\
&= \|\mathcal{T}^{\pi_k}Q^{\pi_k} - \mathcal{T}^{\pi_k}Q_k\| + \|\mathcal{T}Q_k - \mathcal{T}Q_{k-1}\| \\
&\leq \gamma\|Q^{\pi_k} - Q_k\| + \gamma\|Q_k - Q_{k-1}\|\,,
\end{aligned}$$

which means that

$$\|Q^{\pi_k} - Q_k\| \leq \frac{\gamma}{1-\gamma}\|Q_k - Q_{k-1}\| \leq \frac{\gamma^k}{1-\gamma}\|Q_1 - Q_0\| \leq \frac{\gamma^k}{(1-\gamma)^2}\,. \tag{20}$$

The proof is completed by observing that picking $k > \frac{\log(\frac{2}{(1-\gamma)^2\varepsilon})}{\log(1/\gamma)}$ implies $\|V^{\pi_k} - V^*\| < \varepsilon$. □

B.3 Proof of Lemma 3

Proof. We report the proof in the case $\beta > 0$, as the other case is proven using completely similar lines.

Since the policy π is fixed, we omit it from the notation. Let (s,a) be a state-action pair that satisfies $\|Q_1 - Q_2\| = |Q_1(s,a) - Q_2(s,a)|$ and without loss of generality that $Q_1(s,a) \geq Q_2(s,a)$. Then,

$$\begin{aligned}
\|Q_1 - Q_2\| &= \frac{\gamma}{\beta} \ln\left(\frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_2(s')}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \right) \\
&= \frac{\gamma}{\beta} \ln\left(\frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_1(s') + \beta (V_2(s') - V_1(s'))}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \right) \\
&\leq \frac{\gamma}{\beta} \ln\left(e^{\beta \|V_1 - V_2\|} \frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_1(s')}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \right) \\
&= \gamma \|V_1 - V_2\| + \frac{\gamma}{\beta} \ln\left(\frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_1(s')}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \right) \\
&\leq \gamma \|Q_1 - Q_2\| + \frac{\gamma}{\beta} \ln\left(1 + \frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_1(s')} - \sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \right) \\
&\leq \gamma \|Q_1 - Q_2\| + \frac{\gamma}{\beta} \frac{\sum_{s'} P_2(s'|s,a) e^{-\beta V_1(s')} - \sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}}{\sum_{s'} P_1(s'|s,a) e^{-\beta V_1(s')}} \\
&\leq \gamma \|Q_1 - Q_2\| + \frac{\gamma}{\beta} \frac{\sum_{s'} [P_2(s'|s,a) - P_1(s'|s,a)] e^{-\beta V_1(s')}}{e^{-\beta \frac{1}{1-\gamma}}} \\
&\leq \gamma \|Q_1 - Q_2\| + \frac{\gamma}{\beta} e^{\beta \frac{1}{1-\gamma}} \sum_{s'} |P_2(s'|s,a) - P_1(s'|s,a)| \\
&\leq \gamma \|Q_1 - Q_2\| + \frac{\gamma}{\beta} e^{\beta \frac{1}{1-\gamma}} \tau \,.
\end{aligned}$$

Rearranging the terms yields the asserted result:

$$\|Q_1 - Q_2\| \leq \frac{\gamma}{1-\gamma} \frac{e^{\beta \frac{1}{1-\gamma}}}{\beta} \tau \,.$$

□

C Lower Bound on Bernoulli Likelihood Ratio

In this section, we revisit and develop a technical result that bounds the likelihood ratio of two samples under different hypotheses on a high probability event. Parts of the proof closely resembles parts of Lemma 17 in [21]; however, we stress that our treatment fixes an error in the proof, which however requires slightly stronger assumptions than those imposed in [21]. In addition, while the result in [21] only considers $p \geq \frac{1}{2}$, ours deal with both cases of $p \geq \frac{1}{2}$ and $p < \frac{1}{2}$.

Let $p \in (0,1)$ and $\tilde{p} = \max\{p, 1-p\}$. Let $\alpha \in (0, \frac{1-\tilde{p}}{5}]$. Consider two coins (Bernoulli random variables), one with bias $q = p$ and one with bias $q = p + \alpha$. We name the two statistical hypotheses $H_0 : q = p$ and $H_1 : q = p + \alpha$.

Let W be the outcome of flipping one of the coins t times and the associated likelihood function under hypothesis m as

$$L_m(w) := \mathbb{P}_m(W = w) \tag{21}$$

for hypothesis H_m with $m \in \{0,1\}$ and for every possible history of outcomes w, and where $\mathbb{P}_m(W = w)$ denotes the probability of observing the history w under the hypothesis H_m. The likelihood function defines a random variable $L_m(W)$, where W is the stochastic process of realized coin tosses.

Let $t \in \mathbb{N}$ and $\theta = \exp\big(-\frac{c_1\alpha^2 t}{p(1-p)}\big)$. Let k be the number of successes in the t trials and

$$\tilde{k} = \begin{cases} k & \text{if } p \geq \frac{1}{2} \\ t-k & \text{if } p < \frac{1}{2}\,. \end{cases}$$

Finally, we define the event $\mathcal{E}$ as

$$\mathcal{E} = \left\{ \tilde{p}t - \tilde{k} \leq \sqrt{2p(1-p)\log(\frac{c_2}{2\theta})} \right\},$$

where $c_2 \geq 2$ is any constant.

Theorem 5. *For $c_1 = 32$, it holds that $\frac{L_1(W)}{L_0(W)} 1_{\mathcal{E}} \geq \frac{2\theta}{c_2} 1_{\mathcal{E}}$.*

Proof. We distinguish two cases depending on the value of p.

Case 1: $p \geq \frac{1}{2}$. The likelihood ratio can be written as

$$\frac{L_1(W)}{L_2(W)} = \frac{(p+\alpha)^k(1-p-\alpha)^{t-k}}{p^k(1-p)^{t-k}} = \left(1+\frac{\alpha}{p}\right)^k\left(1-\frac{\alpha}{1-p}\right)^{t-k}$$
$$= \left(1+\frac{\alpha}{p}\right)^k\left(1-\frac{\alpha}{1-p}\right)^{k\frac{1-p}{p}}\left(1-\frac{\alpha}{1-p}\right)^{t-\frac{k}{p}}.$$

We start by bounding the second factor using that $\log(1-x) \geq -x-x^2+x^3$ for $x \in [0, \frac{1}{5}]$ (Lemma 6) and that $\exp(x) \geq 1+x$ for all x along with our assumption that $\alpha \leq \frac{1-p}{5}$:

$$\begin{aligned}
\left(1-\frac{\alpha}{1-p}\right)^{\frac{1-p}{p}} &\geq \exp\left(\frac{1-p}{p}\left[-\frac{\alpha}{1-p}-\frac{\alpha^2}{(1-p)^2}+\frac{\alpha^3}{(1-p)^3}\right]\right) \\
&\geq 1-\frac{1-p}{p}\left[\frac{\alpha}{1-p}+\frac{\alpha^2}{(1-p)^2}-\frac{\alpha^3}{(1-p)^3}\right] \\
&= 1-\frac{\alpha}{p}-\frac{\alpha^2}{p(1-p)}+\frac{\alpha^3}{p(1-p)^2} \\
&\geq 1-\frac{\alpha}{p}-\frac{\alpha^2}{p(1-p)}+\frac{\alpha^3}{p^2(1-p)} \\
&= \left(1-\frac{\alpha}{p}\right)\left(1-\frac{\alpha^2}{p(1-p)}\right),
\end{aligned}$$

where we have used that $p \geq 1-p$.

Using this along with the fact that $k \leq t$ and $p \geq 1-p$, it follows that

$$\begin{aligned}
\frac{L_1(W)}{L_0(W)} &\geq \left(1-\frac{\alpha^2}{p^2}\right)^k\left(1-\frac{\alpha^2}{p(1-p)}\right)^k\left(1-\frac{\alpha}{1-p}\right)^{t-\frac{k}{p}} \\
&\geq \left(1-\frac{\alpha^2}{p(1-p)}\right)^{2k}\left(1-\frac{\alpha}{1-p}\right)^{t-\frac{k}{p}} \\
&\geq \left(1-\frac{\alpha^2}{p(1-p)}\right)^{2t}\left(1-\frac{\alpha}{1-p}\right)^{t-\frac{k}{p}}.
\end{aligned}$$

Note that we have $\alpha^2 \leq \frac{(1-p)^2}{25} \leq \frac{p(1-p)}{25} \leq \frac{p(1-p)}{2}$. Using this and the fact that $\log(1-x) \geq -2x$ for $x \in [0, \frac{1}{2}]$, we obtain

$$\begin{aligned}
\left(1-\frac{\alpha^2}{p(1-p)}\right)^{2t} &\geq \exp\left(-4t\frac{\alpha^2}{p(1-p)}\right) \\
&= \theta^{\frac{4}{c_1}} \\
&\geq \left(\frac{2\theta}{c_2}\right)^{\frac{4}{c_1}},
\end{aligned}$$

where we have used that $\frac{2}{c_2} \geq 1$.

Now on the event $\mathcal{E}$, we have that $t - \frac{k}{p} \leq \sqrt{2\frac{1-p}{p} t \log(\frac{c_2}{2\theta})}$. Using this along with the fact that $\frac{1}{c_1}\log(\frac{c_2}{2\theta}) \leq \frac{\alpha^2 t}{p(1-p)}$, which follows since

$$\log(\frac{c_2}{2\theta}) = \log\left(\frac{c_2}{2}\exp\left[\frac{c_1\alpha^2 t}{p(1-p)}\right]\right) \leq \log\left(\exp\left[\frac{c_1\alpha^2 t}{p(1-p)}\right]\right) = \frac{c_1\alpha^2 t}{p(1-p)},$$

we obtain that

$$\begin{aligned}
\left(1 - \frac{\alpha}{1-p}\right)^{t-\frac{k}{p}} &\geq \left(1 - \frac{\alpha}{1-p}\right)^{\sqrt{2\frac{1-p}{p} t \log(c_2/(2\theta))}} \\
&\geq \exp\left(- 2\frac{\alpha}{1-p}\sqrt{2\frac{1-p}{p} t \log(c_2/(2\theta))}\right) \\
&= \exp\left(- 2\sqrt{2}\sqrt{\frac{\alpha^2 t}{p(1-p} \log(c_2/(2\theta))}\right) \\
&\geq \exp\left(- \frac{2\sqrt{2}}{\sqrt{c_1}} \log(c_2/(2\theta))\right) \\
&= \left(\frac{2\theta}{c_2}\right)^{\frac{2\sqrt{2}}{\sqrt{c_1}}}.
\end{aligned}$$

Putting these together, we see that

$$\frac{L_1(W)}{L_2(W)} 1_{\mathcal{E}} \geq \left(\frac{2\theta}{c_2}\right)^{\frac{2\sqrt{2}}{\sqrt{c_1}} + \frac{2(1-p)}{p\cdot c_1} + \frac{2}{c_1}} 1_{\mathcal{E}},$$

so that choosing $c_1 = 32$ yields the claimed result:

$$\frac{L_1(W)}{L_2(W)} 1_{\mathcal{E}} \geq \frac{2\theta}{c_2} 1_{\mathcal{E}}.$$

Case 2: $p < \frac{1}{2}$**.** Define $m = t - k$, which is now the number of failed coin flips. Hence,

$$\begin{aligned}
\frac{L_1(W)}{L_0(W)} &= \frac{(1-p-\alpha)^m (p+\alpha)^{t-m}}{(1-p)^m p^{t-m}} = \left(1 - \frac{\alpha}{1-p}\right)^m \left(1 + \frac{\alpha}{p}\right)^{t-m} \\
&= \left(1 - \frac{\alpha}{1-p}\right)^m \left(1 + \frac{\alpha}{p}\right)^{m\frac{p}{1-p}} \left(1 + \frac{\alpha}{p}\right)^{t - \frac{m}{1-p}}.
\end{aligned}$$

Again, using $\exp(1+x) \geq x$ for all $x \in \mathbb{R}$ and using that $\log(1+x) \geq x - x^2$ for all $x \geq 0$, we get that

$$\begin{aligned}
\left(1 + \frac{\alpha}{p}\right)^{\frac{p}{1-p}} &\geq \exp\left(\frac{p}{1-p}\left[\frac{\alpha}{p} - \frac{\alpha^2}{p^2}\right]\right) \\
&\geq 1 + \frac{\alpha}{1-p} - \frac{\alpha^2}{p(1-p)} \\
&\geq 1 + \frac{\alpha}{1-p} - \frac{\alpha^2}{p(1-p)} - \frac{\alpha^3}{p(1-p)^2} \\
&= \left(1 + \frac{\alpha}{1-p}\right)\left(1 - \frac{\alpha^2}{p(1-p)}\right).
\end{aligned}$$

Using this along with the fact that $(1-p) > p$ and $m \leq t$, we have

$$\begin{aligned}
\frac{L_1(W)}{L_2(W)} &\geq \left(1 - \frac{\alpha^2}{(1-p)^2}\right)^m \left(1 - \frac{\alpha^2}{p(1-p)}\right)^m \left(1 + \frac{\alpha}{p}\right)^{t-\frac{m}{1-p}} \\
&\geq \left(1 - \frac{\alpha^2}{p(1-p)}\right)^{2t} \left(1 - \frac{\alpha}{p}\right)^{t - \frac{m}{1-p}}.
\end{aligned}$$

Again, using $\log(1-x) \geq -2x$ for $x \in [0, \frac{1}{2}]$, we get that

$$\begin{aligned}\left(1 - \frac{\alpha^2}{p(1-p)}\right)^{2t} &\geq \exp\left(-4t\frac{\alpha^2}{p(1-p)}\right) \\ &\geq \theta^{\frac{4}{c_1}} \\ &\geq \left(\frac{2\theta}{c_2}\right)^{\frac{4}{c_1}}.\end{aligned}$$

On the event $\mathcal{E}$, we have that $t - \frac{m}{1-p} \leq \sqrt{\frac{2pt\alpha^2}{1-p}\log(\frac{c_2}{2\theta})}$. Using this along with the fact that $\frac{1}{c_1}\log(\frac{c_2}{2\theta}) \leq \frac{\alpha^2 t}{p(1-p)}$, we get on the event $\mathcal{E}$ that

$$\begin{aligned}\left(1 - \frac{\alpha}{p}\right)^{t - \frac{m}{1-p}} &\geq \left(1 - \frac{\alpha}{p}\right)^{\sqrt{\frac{2p}{1-p}t\log(\frac{c_2}{2\theta})}} \\ &\geq \exp\left(-2\sqrt{\frac{2t}{p(1-p)}\log(\frac{c_2}{2\theta})}\right) \\ &\geq \exp\left(-\frac{2\sqrt{2}}{\sqrt{c_1}}\log(\frac{c_2}{2\theta})\right) \\ &= \left(\frac{2\theta}{c_2}\right)^{\frac{2\sqrt{2}}{\sqrt{c_1}}}.\end{aligned}$$

We thus get the desired result for $c_1 = 32$:

$$\frac{L_1(W)}{L_0(W)}1_{\mathcal{E}} \geq 1_{\mathcal{E}}\left(\frac{2\theta}{c_2}\right)^{\frac{4}{c_1}+\frac{2\sqrt{2}}{\sqrt{c_1}}} \geq 1_{\mathcal{E}}\left(\frac{2\theta}{c_2}\right).$$

□

D Proofs of Lower Bounds

D.1 Lower Bound for Q-value Learning

For a lower bound we let us inspire by the class of MDPs used in [5] where each MDP is of size $SA = 3n$ for some $n \in \mathbb{N}$ and where each MDP consists of n triplets of state-action pairs $\{z_i^0, z_i^G, z_i^B\}$ where the transition from z_i^0 to z_i^G is given by $P(z_i^G|z_i^0) = q_i$ and the transition from z_i^0 to z_i^B is given by $P(z_i^B|z_i^0) = 1-q_i$. Both z_i^G and z_i^B are absorbing and the reward is 1 in z_i^G and 0 in both z_i^0 and z_i^B. For all three types of state-actions we can explicitly calculate the state-action value-functions

$$\begin{aligned}Q(z_i^0) &= \frac{-\gamma}{\beta}\log(q_i e^{-\beta\frac{1}{1-\gamma}} + 1 - q_i) \\ Q(z_i^G) &= \frac{1}{1-\gamma} \\ Q(z_i^B) &= 0\end{aligned}$$

for all $i = 1, \ldots, n$. Denote the collection of all such MDPs by $\mathbb{M}$.

Fix a triplet with index i and consider the two hypotheses $H_0^i : q_i = p$ and $H_1^i : q_i = p + \alpha$ where p and α are given by

$$p = \begin{cases} 1 - e^{-\beta\frac{1}{1-\gamma}} & \text{for } \beta > 0 \\ e^{-|\beta|\frac{1}{1-\gamma}} & \text{for } \beta < 0 \end{cases}$$

and

$$\alpha = 8\varepsilon\frac{|\beta|}{\gamma}\frac{1}{e^{|\beta|\frac{1}{1-\gamma}} - 1}$$

for any ε in the range

$$\varepsilon < \frac{1}{40}\frac{\gamma}{|\beta|}(1 - e^{-|\beta|})\,.$$

We use M_0 to denote an MDP where H_0^i holds for our fixed triplet, and M_1 to denote an MDP where instead H_0^i holds and $\mathbb{E}_0$ and $\mathbb{P}_0$ as the expectations operator and probability operator under H_1^i and similarly $\mathbb{E}_1$ and $\mathbb{P}_1$ under H_0^i. Fix any (ε, δ)-correct Q-algorithm $\mathcal{U}$. We start by showing that with these parameter we have that $Q^*_{M_1}(z_i^0) - Q^*_{M_0}(z_i^0) > 2\varepsilon$, which we do by casing on the sign of β :

Case 1: $\beta < 0$**.** In this case $p = e^{-|\beta|\frac{1}{1-\gamma}}$. We then have

$$\begin{aligned}
Q^*_{M_1}(z_i^0) - Q^*_{M_0}(z_i^0) &= \frac{\gamma}{|\beta|}\log\left(\frac{(p+\alpha)e^{|\beta|\frac{1}{1-\gamma}} + 1 - p - \alpha}{pe^{|\beta|\frac{1}{1-\gamma}} + 1 - p}\right) \\
&= \frac{\gamma}{|\beta|}\log\left(1 + \frac{\alpha(e^{|\beta|\frac{1}{1-\gamma}} - 1)}{pe^{|\beta|\frac{1}{1-\gamma}} + 1 - p}\right) \\
&\geq \frac{\gamma}{|\beta|}\frac{\alpha}{2}\frac{e^{|\beta|\frac{1}{1-\gamma}} - 1)}{pe^{|\beta|\frac{1}{1-\gamma}} + 1 - p} \\
&> \frac{\gamma}{|\beta|}\frac{\alpha}{4}(e^{|\beta|\frac{1}{1-\gamma}} - 1) \\
&= 2\varepsilon\,,
\end{aligned}$$

where we have used that $p = e^{-|\beta|\frac{1}{1-\gamma}}$ and the fact that $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$.

Case 2: $\beta > 0$**.** The case for $\beta > 0$ is similar, although in this case we have $p = 1 - e^{-\beta\frac{1}{1-\gamma}}$ and use the inequality $\log(1+x) \leq x$ for all $x > -1$ to get that

$$\begin{aligned}
Q^*_{M_1}(z_i^0) - Q^*_{M_0}(z_i^0) &= -\frac{\gamma}{\beta}\log\left(\frac{(p+\alpha)e^{-\beta\frac{1}{1-\gamma}} + 1 - p - \alpha}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p}\right) \\
&= -\frac{\gamma}{\beta}\log\left(1 - \frac{\alpha(1 - e^{-\beta\frac{1}{1-\gamma}})}{1 - p + pe^{-\beta\frac{1}{1-\gamma}}}\right) \\
&= -\frac{\gamma}{\beta}\log\left(1 - \frac{\alpha(1 - e^{-\beta\frac{1}{1-\gamma}})}{(1-p)e^{-\beta\frac{1}{1-\gamma}}}\right) \\
&\geq \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{(1+p)e^{-\beta\frac{1}{1-\gamma}}} \\
&\geq \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{2e^{-\beta\frac{1}{1-\gamma}}} \\
&\geq \frac{\gamma}{\beta}\alpha\frac{e^{\beta\frac{1}{1-\gamma}} - 1}{2} \\
&= 4\varepsilon\,.
\end{aligned}$$

In particular this means that the events $B_0 := \{|Q^*_{M_0}(z_i^0) - Q_t^{\mathcal{U}}(z_i^0)| \leq \varepsilon\}$ and $B_1 := \{|Q^*_{M_1}(z_i^0) - Q_t^{\mathcal{U}}(z_i^0)| \leq \varepsilon\}$ are disjoint events Let t be the number of times the algorithm tries z_i^0. Since $\mathcal{U}$ is (ε, δ)-correct it holds that $\mathbb{P}_0(B_0) \geq 1 - \delta \geq \frac{3}{4}$.

Let k be the number of transitions from z_i^0 to z_i^G in the t trials. We then define $\tilde{k}, \tilde{p}$ and θ by

$$\tilde{k} := \begin{cases} k & \text{if } p \geq \frac{1}{2} \\ t - k & \text{if } p < \frac{1}{2} \end{cases}$$

$$\theta := \exp\big(-\frac{32\alpha^2 t}{p(1-p)}\big)$$

$$\tilde{p} = \max\{p, 1-p\}$$

and the event

$$\mathcal{E} = \left\{ \tilde{p}t - \tilde{k} \leq \sqrt{2p(1-p)t\log(\frac{8}{2\theta})} \right\}$$

for which we have $\mathbb{P}_0(\mathcal{E}) > \frac{3}{4}$ by Lemma 16 in [21] and thus $\mathbb{P}_0(B_0 \cap \mathcal{E}) > \frac{1}{2}$. Now by theorem 5 we get that

$$\mathbb{P}_1(B_0) \geq \mathbb{P}_1(B_0 \cap \mathcal{E}) = \mathbb{E}_1[1_{\mathcal{E}} 1_{B_0}] = \mathbb{E}_0\left[\frac{L_1}{L_0} 1_{\mathcal{E}} 1_{B_0}\right] \geq \frac{\theta}{4}\mathbb{E}_0[1_{\mathcal{E}} 1_{B_0}] = \frac{\theta}{4}\mathbb{P}_0(\mathcal{E} \cap B_0) \geq \frac{\theta}{8}$$

Solving for t in $\frac{\theta}{8} > \delta$ we find

$$t < \frac{p(1-p)}{32\alpha^2}\log(\frac{1}{8\delta})$$

and since

$$\begin{aligned}\frac{p(1-p)}{\alpha^2} &= \frac{\gamma^2}{|\beta|^2}\frac{e^{-|\beta|\frac{1}{1-\gamma}}(1-e^{-|\beta|\frac{1}{1-\gamma}})}{64\varepsilon^2}(e^{|\beta|\frac{1}{1-\gamma}}-1)^2 \\ &\geq \frac{\gamma^2}{64\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}-3}{|\beta|^2}\end{aligned}$$

we conclude that if the algorithm $\mathcal{U}$ tries the state-action pair z_i^0 less than

$$\tilde{T}(\varepsilon, \delta) := \frac{\gamma^2}{2048\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}-3}{|\beta|^2}\log(\frac{1}{8\delta})$$

times under the hypothesis H_0^i then $\mathbb{P}_1(B_0) > \delta$ and $B_0 \subset B_1^c$.

Next we use the fact that the structure of the MDPs is such that information on the Q-value of any state-action pair carries no information on the Q-values of any other state-action pair.

If the number of total transition samples is less than $\frac{n}{2}\tilde{T}(\varepsilon, \delta) = \frac{S}{6}\tilde{T}(\varepsilon, \delta)$ there must be at least one $n/2$ state-action pairs in the set $\{z_i^0\}_{i=1}^n$ that has been tried no more than $\tilde{T}(\varepsilon, \delta)$ times which without loss of generality we might assume are the state-action pairs $\{z_i^0\}_{i=1}^{n/2}$.

Let T_i be the number of times the algorithm has tried z_i^0 for $i \leq n/2$ Due to the structure of the MDPs in $\mathbb{M}$ it is sufficient to consider only the algorithms that yields an estimate of $Q_{T_i}^{\mathcal{U}}$ based on samples from z_i^0 since any other samples can yield no information on $Q^*(z_i^0)$

Thus by defining the events $\Lambda_i := \{|Q_{M_1}^*(z_i^0) - Q_{T_i}^{\mathcal{U}}(z_i^0)| > \varepsilon\}$ we have that Λ_i and Λ_j are conditionally independent given T_i and T_j. We then have

$$\begin{aligned}&\mathbb{P}_1(\{\Lambda_i^c\}_{1\leq i\leq n/2} \cap \{T_i \leq \tilde{T}(\varepsilon,\delta)\}_{1\leq i\leq n/2}) \\ &= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{n}/2=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}_1(\{T_i = t_i\}_{1\leq i\leq n/2})\mathbb{P}_1(\{\Lambda_i^c\}_{1\leq i\leq n/2} \cap \{T_i = t_i\}_{1\leq i\leq n/2}) \\ &= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{n}/2=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}_1(\{T_i = t_i\}_{1\leq i\leq n/2}) \prod_{1\leq i\leq n/2} \mathbb{P}_1(\Lambda_i^c \cap \{T_i = t_i\}) \\ &= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{n}/2=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}_1(\{T_i = t_i\}_{1\leq i\leq n/2})(1-\delta)^{n/2}\end{aligned}$$

where we have used the law of total probability from line one to two and from two to three follows from independence. We now have directly that

$$\mathbb{P}_1(\{\Lambda_i^c\}_{1\leq i\leq n/2} | \{T_i \leq \tilde{T}(\varepsilon,\delta)\}_{1\leq i\leq n/2}) \leq (1-\delta)^{\frac{n}{2}}$$

Thus, if the total number of transitions T is less than $\frac{n}{2}\tilde{T}(\varepsilon, \delta)$, then

$$\begin{aligned}\mathbb{P}_1(\|Q^* - Q_T^{\mathcal{U}}\| > \varepsilon) &\geq \mathbb{P}_1\bigg(\bigcup_{z \in S \times A} \Lambda(z)\bigg) \\ &= 1 - \mathbb{P}_1\bigg(\bigcap_{1 \leq i \leq n/2} \Lambda_i^c\bigg) \\ &\geq 1 - \mathbb{P}_1(\{\Lambda_i^c\}_{1 \leq i \leq n/2} | \{T_{z_i} \leq \tilde{T}(\varepsilon, \delta)\}_{1 \leq i \leq n/2}) \\ &\geq 1 - (1 - \delta)^{n/2} \\ &\geq \frac{\delta n}{4}\,,\end{aligned}$$

when $\delta \frac{n}{2} \leq 1$ by Lemma 7. By setting $\delta' = \delta \frac{n}{4}$ we obtain the result. This shows that if the number of samples is smaller than

$$T = \frac{SA}{12288} \frac{\gamma^2}{\varepsilon^2} \frac{e^{|\beta| \frac{1}{1-\gamma}} - 3}{|\beta|^2} \log(\frac{SA}{32\delta}) \tag{22}$$

on the MDP corresponding to the hypothesis $H_0 : \{H_0^i | 1 \leq i \leq n\}$ it holds that $\mathbb{P}_1(\|Q_{M_1}^* - Q_T^{\mathcal{U}}\| > \varepsilon) > \delta'$

D.2 Lower Bound for Policy Learning

We consider a class of MDPs $\mathbb{M}$ of the following form: The state space $\mathcal{S}$ is of size $3n$ for some $n \in \mathbb{N}$ labelled $\mathcal{S} = \{s_i^0, s_i^G, s_i^B\}_{i=1}^n$ and $A + 1$ actions labelled $\mathcal{A} = \{a_0, a_1, ..., a_A\}$. For any i the states s_i^G and s_i^B are absorbing. Otherwise we have $P(s_i^G | s_i^0, a_j) = q_j$ and $P(s_i^B | s_i^0, a_j) = 1 - q_j$. Also $r(s_i^G, a_j) = 1$ for all i and j, and rewards are zero for all other state-action pairs. From the state s_i^0 with probabilities that depend on the action taken the agent will then end up in either a good state s_i^G which is absorbing and yields the maximal unit reward or in the bad state s_i^B which is also absorbing but which yields no reward. The different MDPs thus differ only in their transition probabilities in the choice states s_i^0.

Fix an index $1 \leq i \leq n$. We then consider the following set of possible parameters called hypotheses $H_l^i, l \in \{0, 1, 2, ..., A\}$ given by

$$\begin{aligned}&H_0^i : q(s_i^0, a_0) = p + \alpha \qquad q(s_i^0, a) = p \text{ for } a \neq a_0 \\ &H_l : q(s_i^0, a_0) = p + \alpha \qquad q(s_i^0, a) = p \text{ for } a \notin \{a_0, l\} \qquad q(s_i^0, a_l) = p + 2\alpha\,,\end{aligned}$$

where p and α are given by

$$p = \begin{cases} 1 - e^{-\beta \frac{1}{1-\gamma}} & \beta > 0 \\ e^{-|\beta| \frac{1}{1-\gamma}} & \beta < 0 \end{cases}$$

$$\alpha = \frac{5|\beta|}{\gamma} \frac{\varepsilon}{e^{|\beta| \frac{1}{1-\gamma}} - 1}\,,$$

where we allow for

$$0 < \varepsilon < \frac{\gamma}{50|\beta|}\big(1 - e^{-|\beta| \frac{1}{1-\gamma}}\big)\,,$$

which ensures that $\alpha \leq \frac{e^{-|\beta| \frac{1}{1-\gamma}}}{10}$.

Consider a fixed hypothesis H_l^i for some $l \neq 0$ and the sub-MDP that only consists of the states s_i^0, s_i^G, s_i^B. Here the optimal action is $a^* = a_l$, the second best action is a_0 and all other actions are even worse so the value-error over all states in the triplet for any suboptimal choice of actions will be at least as large as $V^*(s_i^C) - V^0(s_i^C)$ where V^0 is the value by choosing $a = 0$. We now show that any non-optimal action is ε-bad on s_i^C.

Case 1: $\beta > 0$**.**

$$
\begin{aligned}
V^*(s_i^0) - V^0(s_i^0) &= -\frac{\gamma}{\beta}\log\left(\frac{(p+2\alpha)e^{-\beta\frac{1}{1-\gamma}} + 1 - p - 2\alpha}{(p+\alpha)e^{-\beta\frac{1}{1-\gamma}} + 1 - p - \alpha}\right) \\
&= \frac{-\gamma}{\beta}\log\left(1 - \alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p - \alpha(1 - e^{-\beta\frac{1}{1-\gamma}})}\right) \\
&> \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p - \alpha(1 - e^{-\beta\frac{1}{1-\gamma}})} \\
&\geq \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p} \\
&= \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{(1+p)e^{-\beta\frac{1}{1-\gamma}}} \\
&\geq \frac{\gamma}{\beta}\alpha\frac{1 - e^{-\beta\frac{1}{1-\gamma}}}{2e^{-\beta\frac{1}{1-\gamma}}} \\
&= \frac{\gamma}{2\beta}\alpha(1 - e^{-\beta\frac{1}{1-\gamma}}) \\
&\geq \varepsilon\,,
\end{aligned}
$$

where we have used $\log(1 + x) > x$ for $x \in (-1, \infty) \setminus \{0\}$.

Case 2: $\beta < 0$**.**

$$
\begin{aligned}
V^*(s_i^0) - V^0(S_i^0) &= \frac{\gamma}{|\beta|}\log\left(\frac{(p+2\alpha)e^{|\beta|\frac{1}{1-\gamma}} + 1 - p - 2\alpha}{(p+\alpha)e^{|\beta|\frac{1}{1-\gamma}} + 1 - p - \alpha}\right) \\
&= \frac{\gamma}{|\beta|}\log\left(1 + \alpha\frac{e^{|\beta|\frac{1}{1-\gamma}} - 1}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p + \alpha(e^{|\beta|\frac{1}{1-\gamma}} - 1)}\right) \\
&> \frac{\gamma}{2|\beta|}\alpha\frac{e^{|\beta|\frac{1}{1-\gamma}} - 1}{pe^{-\beta\frac{1}{1-\gamma}} + 1 - p + \alpha(e^{|\beta|\frac{1}{1-\gamma}} - 1)} \\
&\geq \frac{\gamma}{2|\beta|}\alpha\frac{e^{|\beta|\frac{1}{1-\gamma}} - 1}{2 + \alpha(e^{|\beta|\frac{1}{1-\gamma}} - 1)} \\
&\geq \frac{\gamma}{2|\beta|}\alpha\frac{e^{|\beta|\frac{1}{1-\gamma}} - 1}{2 + \frac{1}{10}} \\
&\frac{5}{21}\frac{\gamma}{|\beta|}\alpha(e^{|\beta|\frac{1}{1-\gamma}} - 1) \\
&\geq \varepsilon\,,
\end{aligned}
$$

where we have used $\log(1 + x) > \frac{x}{2}$ for $x \in (0, 1)$.

Now haven shown that all non-optimal actions are ε-bad, we wish to show that any algorithm that is (ε, δ)-correct on H_0^i, i.e. choosing the action a_0 with probability at least $1 - \delta$, will also have a probability of choosing a_0 on H_l^i that is larger than δ provided that a_l is not tried sufficiently many times under H_0^i.

Let $\mathbb{P}_l$ and $\mathbb{E}_l$ denote the probability operator and expectation operator under the hypothesis H_l^i. Let $t := t_l^i$ be the number of times the algorithm tries action l in s_i^C under H_0. Assuming that $\delta \in (0, \frac{1}{4})$ and using that the algorithm is (ε, δ)-correct we have that $\mathbb{P}_0(B) \geq 1 - \delta \geq \frac{3}{4}$ where $B = \{\pi^{\mathcal{U}}(s_i^0) = a_0\}$ is the event that the algorithm outputs the action a_0.

Let $\theta = \exp\big(-\frac{32\alpha^2 t}{p(1-p)}\big)$. Fix some $t \in \mathbb{N}$ and let k be the number of transitions to s_i^G in the t trials and

$$\tilde{k} = \begin{cases} k & \text{if } p \geq \frac{1}{2} \\ t-k & \text{if } p < \frac{1}{2} \end{cases}$$

Finally we define the event $\mathcal{E}$ as

$$\mathcal{E} = \left\{ \tilde{p}t - \tilde{k} \leq \sqrt{2p(1-p)\log(\frac{8}{2\theta})} \right\} \tag{23}$$

Form the Chernoff-Hoeffding bound and as shown in [21] we have that $\mathbb{P}_0(\mathcal{E}) > \frac{3}{4}$ and so $\mathbb{P}_0(B \cap \mathcal{E}) > \frac{1}{2}$. From Theorem 5 we get that

$$\mathbb{P}_1(B) \geq \mathbb{P}_1(B \cap \mathcal{E}) = \mathbb{E}_1[1_B 1_{\mathcal{E}}] \geq \mathbb{E}_0\left[\frac{L_1(W)}{L_0(W)} 1_{\mathcal{E}} 1_B\right] \geq \mathbb{E}_0\left[\frac{\theta}{4} 1_{\mathcal{E}} 1_B\right] = \frac{\theta}{4}\mathbb{P}_0(\mathcal{E} \cap B) \geq \frac{\theta}{8} \tag{24}$$

Now solving for $\frac{\theta}{8} > \delta$ we see that if

$$t < \tilde{t}(\varepsilon, \delta) := \frac{1}{800}\log(\frac{1}{8\delta})\frac{\gamma^2}{\varepsilon^2} \cdot \frac{e^{|\beta|\frac{1}{1-\gamma}} - 3}{|\beta|^2} \tag{25}$$

then $\mathbb{P}_1(B) > \delta$ and the event B is containing the event that the algorithm does not choose the optimal action a_l.

Since this holds for all the A hypotheses $H_l^i, l = 1, 2, ..., A$ it follows that the algorithm needs at least $\tilde{T}(\epsilon, \delta) := A\tilde{t}(\varepsilon, \delta)$ samples to be (ε, δ)-correct on the sub-MDP that consists only of the states $\mathcal{S}_i = \{s_i^0, s_i^G, s_i^B\}$.

Next we use the fact that the structure of the MDPs is such that information used to determine $\pi^*(s_i^0)$ carries no information to determine $\pi^*(s_j^0)$ for $i \neq j$.

If the number of total transition samples is less than $\frac{S}{6}\tilde{T}(\epsilon, \delta)$ then there must be at least $\frac{S}{6}$ states in the set $\{s_i^0\}_{i=1}^S$ for which some action (apart from a_0) has been tried no more than $\tilde{T}(\varepsilon, \delta)$ times which without loss of generality we might assume are the states$\{s_i^0\}_{i=1}^{S/6}$. and that it is action a_1 that has been tried out at most $\tilde{T}(\varepsilon, \delta)$ times in each of these states.

Let T_i be the number of times the algorithm has tried sampled any action on s_i^0 for $i \leq S/6$ Due to the structure of the MDPs in $\mathbb{M}$ it is sufficient to consider only the algorithms that yields an estimate of $\pi_{T_i}^{\mathcal{U}}$ based on samples from s_i^0 since any other samples can yield no information on $\pi^*(s_i^0)$.

Thus by defining the events $\Lambda_i := \{|V_{M_1}^*(s_i^0) - V^{\pi_{T_i}^{\mathcal{U}}}(s_i^0)| > \varepsilon\}$ we have that Λ_i and Λ_j are conditionally independent given T_i and T_j. We then have that for the MDP $M_1 \in \mathbb{M}$ (The one corresponding to the hypothesis $H_1 := \{H_1^i | 1 \leq i \leq n\}$) it holds that

$$\begin{aligned}
&\mathbb{P}(\{\Lambda_i^c\}_{1\leq i\leq S/6} \cap \{T_i \leq \tilde{T}(\varepsilon,\delta)\}_{1\leq i\leq S/6}) \\
&= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{S}/6=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}(\{T_i = t_i\}_{1\leq i\leq S/6})\mathbb{P}(\{\Lambda_i^c\}_{1\leq i\leq S/6} \cap \{T_i = t_i\}_{1\leq i\leq S/6}) \\
&= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{S}/6=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}(\{T_i = t_i\}_{1\leq i\leq S/6}) \prod_{1\leq i\leq S/6} \mathbb{P}(\Lambda_i^c \cap \{T_i = t_i\}) \\
&= \sum_{t_1=0}^{\tilde{T}(\varepsilon,\delta)} \cdots \sum_{t_{S}/6=0}^{\tilde{T}(\varepsilon,\delta)} \mathbb{P}(\{T_i = t_i\}_{1\leq i\leq S/6})(1-\delta)^{S/6},
\end{aligned}$$

where we have used the law of total probability from line one to two and from two to three follows from independence. We now have directly that

$$\mathbb{P}\big(\{\Lambda_i^c\}_{1\leq i\leq S/6} | \{T_i \leq \tilde{T}(\varepsilon,\delta)\}_{1\leq i\leq S/6}\big) \leq (1-\delta)^{\frac{n}{2}}.$$

Thus, if the total number of transitions T is less than $\frac{S}{6}\tilde{T}(\varepsilon,\delta)$ on the MDP M_0 corresponding to the hypothesis $H_0 : \{H_0^i | 1 \leq i \leq n\}$, then on M_1 it holds that

$$\begin{aligned}
\mathbb{P}(\|V^* - V^{\pi_T^{\mathcal{U}}}\| > \varepsilon) &\geq \mathbb{P}\bigg(\bigcup_{1\leq i\leq S/6} \Lambda(z)\bigg) \\
&= 1 - \mathbb{P}\bigg(\bigcap_{1\leq i\leq S/6} \Lambda_i^c\bigg) \\
&\geq 1 - \mathbb{P}(\{\Lambda_i^c\}_{1\leq i\leq S/6} | \{T_{z_i} \leq \tilde{T}(\varepsilon,\delta)\}_{1\leq i\leq S/6}) \\
&\geq 1 - (1-\delta)^{S/6} \\
&\geq \frac{\delta S}{12},
\end{aligned}$$

when $\frac{\delta S}{6} \leq 1$ by Lemma 7. By setting $\delta' = \delta\frac{S}{12}$ we obtain the result. This shows that if the number of samples is smaller than

$$T = \frac{SA}{4800}\log(\frac{S}{96\delta})\frac{\gamma^2}{\varepsilon^2}\cdot\frac{e^{|\beta|\frac{1}{1-\gamma}} - 3}{|\beta|^2}$$

on M_0 then on M_1 it holds that $\mathbb{P}(\|V^* - V^{\pi_T^{\mathcal{U}}}\| > \varepsilon) > \delta$.