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

Anonymous authors
Paper under double-blind review

Abstract

We study risk-sensitive reinforcement learning in finite discounted MDPs with recursive entropic risk measures (ERM), where the risk parameter $\beta \neq 0$ controls the agent's risk attitude: $\beta > 0$ for risk-averse and $\beta < 0$ for risk-seeking behavior. A generative model of the MDP is assumed to be available. Our focus is on the sample complexities of learning the optimal state–action value function (value learning) and an optimal policy (policy learning) under recursive ERM. We introduce a model-based algorithm, called Model-Based ERM Q-Value Iteration (MB-ERM-QVI), and derive PAC-type bounds on its sample complexity for both value and policy learning. Both PAC bounds scale exponentially with $|\beta|/(1-\gamma)$, where γ is the discount factor. We also establish corresponding lower bounds for both value and policy learning, showing that exponential dependence on $|\beta|/(1-\gamma)$ is unavoidable in the worst case. The bounds are tight in the number of states and actions (S and A), providing the first rigorous sample complexity guarantees for recursive ERM across both risk-averse and risk-seeking regimes.

1 Introduction

In reinforcement learning (RL), the aim of the agent is to conventionally maximize the expected return, which is defined in terms of a (possibly discounted) sum of rewards [71]. The environment is typically modeled via the Markov Decision Process (MDP) framework [60], wherein efficient computation of an optimal policy, thanks to optimal Bellman equations, renders possible. However, as a *risk-neutral* objective, the expected return is inadequate to capture the true needs of many high-stake applications arising in, e.g., medical treatment [24], finance [65, 10], operations research [20], and transportation [39]. In these domains, decision making must take into account the variability of returns, and risks thereof. To address this limitation, one may opt to maximize a risk measure of the return distribution. Alternatively, one may model the entire distribution of return, as in distributional RL [9], which has received significant attention over the last decade. In this paper, we focus on the former.

Within the first approach, the risk is elegantly quantified via concave risk measures, which lead to well-defined optimization problems. Notable risk measures include mean-variance [46], value-at-risk (VaR) [21], Conditional VaR (CVaR) [66], entropic risk [33], and entropic VaR (EVaR) [2], all of which have been applied to a wide-range of scenarios. Among these, CVaR has become particularly popular for modeling risk-sensitivity in MDPs [18, 11, 15, 7], mainly due to a delicate control it offers for the undesirable tail of return distribution. Despite its popularity and appealing interpretation, learning in MDPs with CVaR-based objectives may pose technical challenges [7]. ERM, as another popular notion, has long been considered for risk-sensitive control in MDPs and RL [33, 14, 31, 34, 27]. However, much of the existing literature focuses on undiscounted settings, despite the prevalence of discounted MDPs; see, e.g., [8, 31, 55] for notable exceptions.

In risk-sensitive RL with a specified risk measure, objectives can be formulated in two fundamentally different ways, depending on how the risk functional is applied to the reward sequence $(r_t)_{t\geq 0}$. The first approach, referred to as *non-recursive* (also called non-iterated or static), consists in directly applying the risk functional to the total return (e.g., $\sum_{t=0}^{\infty} \gamma^t r_t$ in the discounted case) [13, 14, 32]. The second, termed *recursive* (also called iterated, nested, or dynamic), applies the risk functional at every step t to the reward-to-go [4, 5, 22];

see Section 3 for details. In other words, the non-recursive formulation captures trajectory-level risk, whereas the recursive formulation deals with risk locally at each step. The two formulations are qualitatively different and should be viewed as orthogonal modeling choices. While trajectory-level risk is often easier to interpret, it may lead to policies that allow the agent to visit high-risk states, even though the risk of the entire trajectory is still controlled, which might be unacceptable in many safety-critical applications. In contrast, the recursive formulation may lead to more cautious behavior by discouraging entry into high-risk states at every step, which can be either desirable or overly conservative depending on the application [22]. From a theoretical perspective, another key distinction is that non-recursive formulations do not generally admit Bellman-type optimality equations and may lead to time-inconsistent optimal policies (see [36]), whereas recursive formulations preserve Bellman-type optimality structures. Motivated by these considerations, we study risk-sensitive discounted RL with objectives defined via the recursive ERM.

1.1 Main Contributions and Paper Organization

We consider risk-sensitive RL in tabular discounted MDPs under recursive ERM, encompassing both the risk-averse and risk-seeking regimes. We assume that the agent has access to a generative model of the MDP, namely, a simulator that generates samples from the true MDP for arbitrary state-action pairs. Learning performance is assessed in terms of sample complexity, defined as the total number T of samples required, for given (ε, δ), to obtain either an ε-optimal policy (*policy learning*) or an ε-close approximation of the optimal Q-value in the max-norm (*value learning*), with probability exceeding $1-\delta$.

We make the following contributions. We develop a model-based algorithm, called Model-Based ERM Q-Value Iteration (MB-ERM-QVI), and establish PAC-type bounds on its sample complexity for both value learning and policy learning. MB-ERM-QVI is based on a plug-in estimation of the transition kernel combined with a Q-value iteration scheme adapted to recursive ERM objectives. This QVI structure is inspired by the value iteration method of [5], which considers the risk-averse planning setting with known dynamics. We show that its sample complexity for value learning (Theorem 1) and policy learning (Theorem 2) scales as[1]

$$\widetilde{O}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2|\beta|^2}e^{2|\beta|/(1-\gamma)}\right) \qquad \text{and} \qquad \widetilde{O}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2|\beta|^2}\min\left\{S, \frac{1}{(1-\gamma)^2}\right\}e^{2|\beta|/(1-\gamma)}\right),$$

respectively. These bounds hold for any discounted MDP with S states, A actions, discount factor γ, and risk parameter β, with $\beta > 0$ (respectively, $\beta < 0$) corresponding to a risk-averse (respectively, risk-seeking) agent; see Section 3 for a precise definition. Moreover, the bounds are valid over the entire ε-range, namely $\varepsilon \in (0, \frac{1}{1-\gamma}]$. A notable feature of these results is the exponential dependence on the effective horizon $1/(1-\gamma)$, which is absent in the conventional risk-neutral setting, wherein $\beta = 0$.

We further establish worst-case lower bounds on the sample complexity of recursive ERM. Specifically, we show that for value learning (Theorem 3) and policy learning (Theorem 4), at least

$$\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2|\beta|^2}e^{|\beta|/(1-\gamma)}\right)$$

samples are required to achieve ε-optimality. These lower bounds demonstrate that the exponential dependence on $|\beta|/(1-\gamma)$ in sample complexity upper bounds is unavoidable, thereby establishing that learning under recursive ERM is fundamentally more challenging than in the risk-neutral case. To the best of our knowledge, these results constitute the first upper and lower bounds on the sample complexity of recursive ERM in discounted MDPs. A summary of our results is provided in Table 1.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the necessary background and formal problem setup. Section 4 presents the MB-ERM-QVI algorithm, while Section 5 reports its sample complexity guarantees, with proofs deferred to Section 6. Lower bounds are presented in Section 7. Section 8 presents numerical results to demonstrate the performance of MB-ERM-QVI. Finally, Section 9 concludes with a discussion and directions for future research. Additional background on risk measures, along with omitted proofs, is provided in the appendix.

[1]Throughout the paper, $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ suppress logarithmic factors in the relevant problem parameters.

Problem	Upper Bound	Lower Bound
ERM (value learning)	$\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2\lvert\beta\rvert^2}e^{2\lvert\beta\rvert/(1-\gamma)}\right)$ [Theorem 1]	$\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2\lvert\beta\rvert^2}e^{\lvert\beta\rvert/(1-\gamma)}\right)$ [Theorem 3]
ERM (policy learning)	$\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2\lvert\beta\rvert^2}\min\left\{S, \frac{1}{(1-\gamma)^2}\right\}e^{2\lvert\beta\rvert/(1-\gamma)}\right)$ [Theorem 2]	$\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2\lvert\beta\rvert^2}e^{\lvert\beta\rvert/(1-\gamma)}\right)$ [Theorem 4]
Risk-neutral (value learning)	$\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3}\right)$ [30]	$\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3}\right)$ [30]
Risk-neutral (policy learning)	$\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3}\right)$ [30, 1, 47, 37]	$\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3}\right)$ [30]

Table 1: Summary of upper and lower bounds presented in this paper. β denotes the risk parameter, where $\beta > 0$ (respectively, $\beta < 0$) corresponds to a risk-averse (respectively, risk-seeking) agent.

2 Related Work

Risk-neutral discounted RL. There is a large body of papers on provably-sample efficient learning algorithms in tabular discounted MDPs, encompassing a variety of settings such as the generative setting [38, 30, 1, 67, 47, 37], the offline (or batch) setting [61, 49], and the online setting [70, 45]. In our overview of risk-neutral work, we restrict attention to the generative setting —which is the setting we consider— with the aim of collecting most notable developments and key results. In this line, [40] reports the earliest known sample complexity bounds, which is achieved by a model-free method. Azar et al. [30] substantially improve upon this by showing that a simple model-based method attains optimal sample complexity bounds scaling as $\widetilde{\mathcal{O}}(\frac{SA}{\varepsilon^2(1-\gamma)^3})$ for both value learning and policy learning, albeit for substantially limited ε-ranges. It also establishes a lower bound of $\widetilde{\Omega}(\frac{SA}{\varepsilon^2(1-\gamma)^3})$ for value learning. Further algorithms and results are reported in more recent subsequent work, notably including [67, 74, 48, 37]. Among these, [67, 74, 37] present model-free methods, with [67, 37] presenting minimax-optimal bounds, although valid for restricted ε-ranges. Under model-based methods, minimax-optimal bounds, beyond [30], are reported in [1, 48]. In particular, [1] uses empirical MDP combined with a black-box planner, and reports a minimax-optimal bound for $\varepsilon \in \left(0, \frac{1}{\sqrt{1-\gamma}}\right)$, thus expanding that in [30]. More recently, [48] establishes minimax-optimal bounds for the entire ε-range, which are achieved by model-based methods built via the empirical MDP but with reward perturbations or conservative planning. It is worth emphasizing that existing optimal sample complexities rely on techniques that crucially exploit the additivity of the return with respect to rewards; this structural property generally fails for risk-sensitive objectives, and the corresponding techniques do not carry over.

We remark that the abovementioned works look at the learning performance in a worst-case scenario, which yield sample complexity bounds that hold for a model class. This is typically done via uniformly sampling various state-action pairs. In contrast, some studies (e.g., [3, 76]) consider adaptive sampling to account for the heterogeneity across state-action space of the MDP, typically resulting in instance-dependent bounds.

Risk-sensitive RL. There exists a substantial literature on decision making under a risk measure in bandit and RL settings. In bandits, risk-sensitive objectives are typically studied through regret minimization; see, e.g., [64, 51, 42]. Extensions to MDPs introduce substantially richer structural and algorithmic challenges, which are the focus of this work. The literature on RL under risk measures may be broadly categorized by the type of risk measure studied. Representative examples include CVaR [22, 23, 16, 44], ERM [14, 56, 52, 53, 31], mean-variance risk [69, 35, 43], and EVaR [57, 29]. Among these, CVaR has been the most extensively studied. Under non-recursive CVaR, [23] and [16] investigate online episodic RL in the regret setting for tabular MDPs and MDPs with function approximation, respectively, while reward-free RL is studied in [58]. Under recursive CVaR, [22] analyzes sample complexity in the generative setting; however, its analysis relies on structural properties specific to CVaR and does not extend to ERM.

ERM has also been widely studied, beginning with early work such as [33] and followed by a rich literature across diverse settings [14, 12, 13, 63, 27, 25, 26, 52, 53, 56, 31]. Under non-recursive ERM, most work focuses on the average-reward or episodic settings, with planning studied in [33, 14, 53] and learning in [13, 56, 52]. The discounted setting has received comparatively little attention, largely due to technical challenges introduced by discounting; notable exceptions include planning results in [8, 31] and learning results in [55],

which modify the problem formulation to address time inconsistency. Under recursive ERM, recent works such as [27, 26, 25, 34, 50] study online episodic RL in the regret setting. To the best of our knowledge, existing work on discounted MDPs under recursive ERM is limited to planning; a notable example is [5], which provides a thorough theoretical treatment but does not propose learning algorithms.

Some work in risk-sensitive RL and control studies broader classes of risk measures. Two notable classes studied in this context are coherent risk measures and optimized certainty equivalent (OCE) measures, both of which include important special cases such as mean–variance and CVaR. While ERM is not coherent, it belongs to the OCE class for $\beta > 0$; a brief overview of risk measures is provided in Appendix A. Results for coherent risks [59, 72, 44, 77] do not apply to ERM, and existing results for OCE risks [73, 75, 62] do not address provably sample-efficient learning under recursive ERM in discounted MDPs. In particular, [62] considers offline RL in discounted MDPs under recursive OCE but does not provide sample-complexity guarantees. We also note that a connection between MDPs with recursive coherent risks and distributionally robust MDPs has been established in [5]. Finally, we note that approaches such as safe RL and constrained MDPs [19, 17] incorporate risk awareness into policies via constraints, but without altering the definition of return; they are therefore generally regarded as orthogonal to the present setting.

3 Background

Notations. For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. $\mathbb{1}_A$ 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 $1 \le p < \infty$. The notation $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ denotes the space of essentially bounded random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

3.1 Entropic Risk Preferences

The entropic risk measure (ERM) is rooted in expected utility theory [54]. 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 $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, it is easy to verify that the associated *certainty equivalent* to u is $u^{-1}(\mathbb{E}[u(X)]) = \frac{-1}{\beta}\log(\mathbb{E}[e^{-\beta X}])$, which 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 $\rho : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ by[2]

$$\rho(X) := \rho(X; \beta) := -\frac{1}{\beta}\log\left(\mathbb{E}[e^{-\beta X}]\right). \tag{1}$$

Evidently, when $\beta \to 0$ we recover the risk-neutral case, which simply coincides with the expectation: $\lim_{\beta \to 0} \rho(X) = \mathbb{E}[X]$. Further, it is straightforward to see that ρ admits the following:

$$\rho(X) \le \rho(Y), \quad \text{for any } X \le Y, \tag{2}$$
$$\rho(c) = c, \quad \text{for any } c \in \mathbb{R}, \tag{3}$$
$$\rho(X) \le \mathbb{E}[X], \quad \text{for } \beta > 0, \tag{4}$$
$$\rho(X) \ge \mathbb{E}[X], \quad \text{for } \beta < 0, \tag{5}$$

where properties (4)-(5) follow from Jensen's inequality. Using ρ as a measure of the preference for different random variables, it follows directly from (3)-(5) that $\rho(X) \le \rho(\mathbb{E}[X])$ for $\beta > 0$ and that $\rho(X) \ge \rho(\mathbb{E}[X])$ for $\beta < 0$. It further shows that $\beta > 0$ is associated with *risk-aversion*, while $\beta < 0$ is associated with a *risk-seeking* behavior. It is well-known that the ERM, unlike CVaR, does not belong to the nice class of coherent risk measures; we refer the reader to Appendix A for a primer on risk measures, where we collect some definitions and concrete examples.

[2]We note that there is a lack of consensus regarding the sign of β in the definition of ERM. We follow this convention considering its widespread use in the actuarial literature [4]. We refer to Appendix A for a related discussion.

3.2 Discounted Markov Decision Processes with Entropic Risk

We write the 6-tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \beta)$ to denote a finite, discounted infinite-horizon Markov decision process (MDP), where $\mathcal{S} = \{1, 2, \ldots, S\}$ is the finite state space of size $S := |\mathcal{S}|$, $\mathcal{A} = \{1, 2, \ldots, A\}$ 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 reward function, $\gamma \in (0, 1)$ is the discount factor, and $\beta \neq 0$ is the risk-parameter. We use $Z = \mathcal{S} \times \mathcal{A}$ to denote the set of all state-action pairs, and write $P_{s,a}(s')$ as short-hand notation for $P(s'|s, a)$ for any $(s, a) \in Z$. For simplicity of exposition, we consider a deterministic reward function, as is standard in the literature. 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 $t \geq 0$, 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 $(s_t, a_t, r_t)_{t\geq 0}$. The agent's goal is to maximize an objective function, as a function of the collected rewards $(r_t)_{t\geq 0}$, which depends on both γ and β.

To concretely define the agent's objective using ERM, we discuss two approaches of applying the functional ρ in (1) in the context of MDPs. The first approach, called *non-recursive* (or static or non-iterated) [31, 53], consists in applying ρ to the total discounted sum of rewards $\rho\big(\sum_{t=0}^{\infty} \gamma^t r_t\big)$, which is well-defined under the bounded rewards assumption, i.e., $r_t \in [0, 1]$ for all t. This problem admits no obvious optimality equation, although solution and approximation schemes for planning have been proposed in the literature [31, 53]. The other approach where ρ is applied at every step is called *recursive* (also called dynamic or iterated) [4]. The planning problem in this case is tractable thanks to existence of Bellman-type optimality equations. The recursive approach also guarantees the existence of an optimal stationary deterministic policy, whereas the non-recursive approach may lead to optimal policies that are not time-consistent (see [36]). In this paper, we study the case where the RL objective is defined via the recursive ERM.

3.3 Value Function and Q-function

We shall introduce some notation and definitions to formally define the value function V and state-action value function Q (henceforth, Q-value) of a policy. We follow the approach of [4] and [6], but since none of their cases include our $\beta < 0$ case and also only cover value function, we present in Appendix B the full setup with history-dependent policies as well as a thorough definition of the value and Q-values. There, we prove existence of a stationary optimal policy, and show that the value functions of this policy satisfy a Bellman optimality equation and that the value of any policy satisfies a Bellman recursion. We give an outline here that only deals with stationary policies, which is justified by the results of Appendix B.

Let $v \in \mathbb{R}^S$ and $\pi : \mathcal{S} \to \mathcal{A}$ be a stationary deterministic policy. We define $\rho_{s,a} : \mathbb{R}^S \to \mathbb{R}$ as

$$\rho_{s,a}(v) = -\frac{1}{\beta} \log \Big(\mathbb{E}_{s' \sim P_{s,a}} [e^{-\beta v(s')}] \Big) \tag{6}$$

and slightly abusing the notation, we write $\rho_{s,\pi}$ when $a = \pi(s)$, i.e., $\rho_{s,\pi} := \rho_{s,\pi(s)}$. We then introduce the operator $J_\pi : \mathbb{R}^S \to \mathbb{R}^S$ with $J_\pi(v)(s) = R(s, \pi(s)) + \gamma \rho_{s,\pi}(v)$. The N-step total discounted utility $J_N(s, \pi)$ is defined as applying J_π recursively N times to the 0-map: $J_N(s, \pi) := J_\pi^N(\mathbf{0})(s)$. Note that the outer-most iteration corresponds to the immediate time-step. Then, the value of policy π is defined as: $V^\pi(s) = \lim_{N\to\infty} J_N(s, \pi)$ for all $s \in \mathcal{S}$. By properties (2)-(3), it follows that $J_N(s, \pi)$ is increasing in N and that $J_N(s, \pi) \leq \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, so that the limit above exists and the value function is thus well-defined. The optimal state-values are defined as $V^*(s) = \sup_\pi V^\pi(s)$ for all $s \in \mathcal{S}$, where the sup is taken over all possible policies. Any policy achieving $V^*(s)$ at all states is called optimal and V^* is called the optimal value function. Further, given $\varepsilon > 0$, a policy achieving $V^\pi(s) \geq V^*(s) - \varepsilon$ for all $s \in \mathcal{S}$ is called ε-optimal.

In [4], the authors consider a more general MDP framework that is not restricted to finite MDPs or stationary policies; they prove that under some conditions —that are trivially fulfilled in the case of finite MDPs— there exists a stationary deterministic optimal policy in the $\beta > 0$ case. This result is easily extended to the $\beta < 0$ case and we give a unified proof for completeness in Appendix B. The optimal value function V^*

satisfies the Bellman optimality equation:

$$V^*(s) = \max_{a\in\mathcal{A}} \left(R(s,a) - \frac{\gamma}{\beta} \log \left(\mathbb{E}_{s'\sim P_{s,a}}[e^{-\beta V^*(s')}] \right) \right), \quad \forall s \in \mathcal{S}.$$

Further, for any stationary deterministic policy π, the value function satisfies the Bellman recursion:

$$V^\pi(s) = R(s,\pi(s)) - \frac{\gamma}{\beta} \log \left(\mathbb{E}_{s'\sim P_{s,\pi(s)}}[e^{-\beta V^\pi(s')}] \right), \quad \forall s \in \mathcal{S}. \tag{7}$$

We introduce the Q-value functions using a similar approach. Given $\pi : \mathcal{S} \to \mathcal{A}$, we define the operator $L_\pi : \mathbb{R}^S \to \mathbb{R}^{S\times A}$ as follows: for all $v : \mathcal{S} \to \mathbb{R}$, for all (s,a), $L_\pi(v)(s,a) = R(s,\pi(s)) + \gamma\rho_{s,\pi}(v)$. Also, we define the operator $L : \mathbb{R}^S \to \mathbb{R}^{S\times A}$ as follows: for all $v : \mathcal{S} \to \mathbb{R}$, for all (s,a), $L(v)(s,a) = R(s,a) + \gamma\rho_{s,a}(v)$. 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 [4] does not consider Q-value functions, building on their arguments we show in Appendix B that it suffices to consider stationary policies when solving $\max_\pi Q^\pi(s,a) =: Q^*(s,a)$ for all (s,a) and that Q^* satisfies the optimality equation:

$$Q^*(s,a) = R(s,a) - \frac{\gamma}{\beta} \log \left(\mathbb{E}_{s'\sim P_{s,a}} \left[e^{-\beta \max_{a'} Q^*(s',a')} \right] \right), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Similarly, the Q-value of a policy π satisfies the Bellman recursion:

$$Q^\pi(s,a) = R(s,a) - \frac{\gamma}{\beta} \log \left(\mathbb{E}_{s'\sim P_{s,a}} \left[e^{-\beta Q^\pi(s',\pi(s'))} \right] \right), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Let us introduce the operators $\mathcal{T}^\pi, \mathcal{T} : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, which for $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ are defined as

$$(\mathcal{T}f)(s,a) = R(s,a) - \frac{\gamma}{\beta} \log \sum_{s'} P_{s,a}(s') e^{-\beta \max_{a'} f(s',a')},$$
$$(\mathcal{T}^\pi f)(s,a) = R(s,a) - \frac{\gamma}{\beta} \log \sum_{s'} P_{s,a}(s') e^{-\beta f(s',\pi(s'))}.$$

Then, the Bellman equations above can be written as $Q^* = \mathcal{T}Q^*$ and $Q^\pi = \mathcal{T}^\pi Q^\pi$.[3]

We conclude this section by a remark about the case of rewards bounded in $[R_{\min}, R_{\max}]$.

Remark 1. *For rewards bounded in $[R_{\min}, R_{\max}]$, one can equivalently consider rewards in $[0,1]$ but with a risk parameter $\frac{\beta}{w}$, with $w := R_{\max} - R_{\min}$. This is verified by observing that (i) ρ is* translation invariant*, implying that the range w is important –not the absolute values– so that one must model rewards as belonging to $[0,w]$; and (ii) one has $\rho(wX,\beta) = -\frac{w}{w\beta}\log(\mathbb{E}[e^{-\beta wX}] = w\rho(X,w\beta)$ for any $w > 0$, so that scaling the range by w amounts to working with ρ with a risk parameter $w\beta$. This observation justifies our focus on rewards in $[0,1]$, but more importantly signals that the risk parameter is calibrated to a specific reward range.*

3.4 Learning Performance

We consider RL algorithms that aim to find an ε-optimal policy or an ε-optimal value function for input $\varepsilon > 0$ under ERM, while having access to a generative model (or simulator) of the MDP. Precisely speaking, the generative model can produce a sample $s' \sim P_{s,a}$ for any queried state-action (s,a). We consider two types of such algorithms, which we generically denote by $\mathcal{U}$: The first type outputs a Q-value $Q_T^{\mathcal{U}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, whereas the second outputs a policy $\pi_T^{\mathcal{U}} : \mathcal{S} \to \mathcal{A}$ using T samples. We evaluate the quality of an algorithm that outputs a Q-value by $\|Q^* - Q_T^{\mathcal{U}}\|$. For an algorithm that instead outputs a policy, we evaluate it in terms of $\|V^* - V^{\pi_T^{\mathcal{U}}}\|$. Often, we will suppress T from the notation. This leads to the notion of (ε,δ)-correct value and policy for input parameters (ε,δ):

[3] We note that our analysis only rests on the Bellman optimality equation; the Bellman equations for V^π and Q^π are included for completeness.

Definition 1 ((ε, δ)-correct value and policy)**.** *An algorithm $\mathcal{U}$ that outputs a Q-value $Q^{\mathcal{U}}$ is called* (ε, δ)-value-correct *on a set of MDPs $\mathbb{M}$ if $\mathbb{P}(\|Q^* - Q^{\mathcal{U}}\| \leq \varepsilon) \geq 1 - \delta$ for all $M \in \mathbb{M}$. Similarly, an algorithm $\mathcal{U}$ that outputs a policy $\pi^{\mathcal{U}}$ is called* (ε, δ)-policy-correct *on a set of MDPs $\mathbb{M}$ if $\mathbb{P}(\|V^* - V^{\pi^{\mathcal{U}}}\| \leq \varepsilon) \geq 1 - \delta$ for all $M \in \mathbb{M}$.*

The notion of (ε, δ)-value-correctness yields a sample complexity notion in the case of *value learning*, while (ε, δ)-policy-correctness serves a similar role for *policy learning*.

We remark that any algorithm that outputs a Q-value also outputs a policy, e.g., the one obtained by acting greedily with respect to the Q-value. However, it is well-known that such a greedy policy can be off by a factor of $1/(1-\gamma)$, which impacts the corresponding sample complexity of policy learning; see discussions in [68, 67, 1]. To remedy this, the literature resort to proof arguments, which may however come at a cost of limiting the ε-range, as briefly discussed in Section 2.

4 Model-Based ERM Q-Value Iteration

In this section, we present a model-based algorithm, called MB-ERM-QVI, for value and policy learning settings under ERM, assuming access to a generative model of the MDP. Then, we derive PAC-type bounds on its sample complexity.

We begin with introducing the protocol for obtaining N samples from each state-action pair in $Z = \mathcal{S} \times \mathcal{A}$; this is done by making a total of $T = NSA$ calls to the generative model (see Algorithm 1).

Algorithm 1: Model estimation

Input: Generative model P
Output: Model estimate $\widehat{P}$
Function `EstimateModel`(N):
  $\forall\ (s,z) \in \mathcal{S} \times Z : m(s,z) = 0$
  for *each* $z \in Z$ **do**
    for $i = 1, 2, \ldots, N$ **do**
      $s \sim P(\cdot|z)$
      $m(s,z) := m(s,z) + 1$
    end
    $\forall s \in \mathcal{S} : \widehat{P}(s,z) = \frac{m(s,z)}{N}$
  end
  return $\widehat{P}$

Algorithm 2: ERM-QVI

Input: Empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, \gamma, \beta)$, number of iterations k
Output: Estimate Q_k of optimal Q-function Q^*
Initialization: $\forall (s,a)$ set $Q(s,a) = 0$
for $j = 0, 1, \ldots, k-1$ **do**
  for *all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $Q_{j+1}(s,a) = R(s,a) - \frac{\gamma}{\beta} \log\left(\mathbb{E}_{s' \sim \widehat{P}_{s,a}}\left[e^{-\beta \max_{a'} Q_j(s',a')}\right]\right)$
  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

Let $\widehat{P}$ denote the plug-in estimator built using the $T = NSA$ independent samples obtained from the generative model; that is, for $(s,a,s') \in Z \times \mathcal{S}$, $\widehat{P}(s'|s,a) = \frac{n(s,a,s')}{N}$, where $n(s,a,s')$ denotes the number of times s' was observed under the queried pair $(s,a) \in Z$. The model-based approach we describe relies on the empirical MDP formulated using $\widehat{P}$, $\widehat{M} = (\mathcal{S}, \mathcal{A}, R, \widehat{P}, \gamma, \beta)$, but it is otherwise general in the sense that it can use any oracle that outputs an ε-optimal policy for any $\varepsilon > 0$. We prove the existence of one such oracle in the analysis (cf. Lemma 2) in the form of a Q-value iteration akin to that of the classical risk-neutral setting. It is the basis for ERM-QVI (Algorithm 2), which is a Q-value iteration for discounted MDPs with ERM with correctness guarantees.

Equipped with these, we introduce MB-ERM-QVI. For an input $\varepsilon > 0$, the algorithm consists in:

(i) building an empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, R, \gamma, \beta)$ via calling the generative model N times (namely, $\widehat{P} = \texttt{EstimateModel}(N)$);

(ii) returning ERM-QVI($\widehat{M}, k$).

We show in Lemma 2 how to pick $k := k(\varepsilon)$ to ensure ε-value-correctness (i.e., $\|Q^* - Q_k\| \leq \varepsilon$) and ε-policy-correctness (i.e., $\|V^* - V^{\pi_k}\| \leq \varepsilon$).

5 Sample Complexity Analysis of MB-ERM-QVI

In this section, we present a sample complexity analysis of MB-ERM-QVI under both policy learning and value learning.

5.1 Properties of ERM-QVI

We first state two results for ERM-QVI, establishing its convergence properties. Their proofs are presented in Appendix D.

Lemma 1 (Contraction properties)**.** *The operators $\mathcal{T}$ and $\mathcal{T}^\pi$ are γ-contractions with respect to the max-norm, i.e., $\|\mathcal{T}f_1 - \mathcal{T}f_2\| \leq \gamma\|f_1 - f_2\|$ and $\|\mathcal{T}^\pi f_1 - \mathcal{T}^\pi f_2\| \leq \gamma\|f_1 - f_2\|$ for value functions f_1 and f_2.*

The proof of this result is very similar to that of Part (a) in Theorem 3.1 in [4]; nevertheless, we include it for completeness. The next lemma shows that for large enough k in ERM-QVI, we can obtain Q_k and V^{π_k} that are as close to, respectively, Q^* and V^* as desired:

Lemma 2. *Fix $\varepsilon > 0$. Under ERM-QVI (Algorithm 2), one has: (i) $\|Q_k - Q^*\| < \varepsilon$ if $k \geq \frac{-\log((1-\gamma)\varepsilon)}{\log(1/\gamma)}$; (ii) $\|V^{\pi_k} - V^*\| < \varepsilon$ if $k \geq \frac{\log 2 - \log((1-\gamma)^2\varepsilon)}{\log(1/\gamma)}$.*

5.2 Sample Complexity Upper Bounds

We are ready to present sample complexity bounds for MB-ERM-QVI: Theorem 1 states such a result for the case of value learning, while Theorem 2 offers a bound for policy learning.

Theorem 1 (Sample complexity, value learning)**.** *For any $\varepsilon > 0$, $\delta \in (0,1)$, and any MDP M with S states and A actions, if the learner makes*

$$T \geq \frac{2SA\gamma^2}{\varepsilon^2(1-\gamma)^2}\left(\frac{e^{|\beta|/(1-\gamma)} - 1}{|\beta|}\right)^2 \log\left(\frac{SA}{\delta}\right)$$

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

Theorem 2 (Sample complexity, policy learning)**.** *For any $\varepsilon > 0, \delta \in (0,1)$, and any MDP M with S states and A actions, if the learner makes*

$$T \geq \frac{9SA\gamma^2}{\varepsilon^2(1-\gamma)^2}\left(\frac{e^{|\beta|/(1-\gamma)} - 1}{|\beta|}\right)^2 \min\left\{\frac{\gamma^2}{(1-\gamma)^2}\log\left(\frac{4SA}{\delta}\right), \log\left(\frac{4SA|\Pi|}{\delta}\right)\right\}$$

model calls, then $\mathbb{P}(\|V^* - V^{\pi_k}\| \leq \varepsilon) \geq 1 - \delta$. Here, Π denotes the set of stationary deterministic policies.

The sample complexity bound offered by Theorem 2 can be further simplified using the worst-case bound $|\Pi| \leq A^S$ to $\widetilde{\mathcal{O}}\big(\frac{SA}{\varepsilon^2(1-\gamma)^2}\min\big\{S, \frac{1}{(1-\gamma)^2}\big\}\big)L^2\big)$ with $L = \frac{1}{|\beta|}\big(e^{|\beta|/(1-\gamma)} - 1\big)$. To be more precise, by including log-terms, depending on whether $S \ll 1/(1-\gamma)^2$ in the problem at hand, one may obtain a bound of

$$\mathcal{O}\left(\frac{SA}{\varepsilon^2(1-\gamma)^4}\log\left(\frac{4SA}{\delta}\right)L^2\right) \qquad \text{or} \qquad \mathcal{O}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2}\left(S + \log\left(\frac{4SA}{\delta}\right)\right)L^2\right).$$

Let us remark however that in problems where $|\Pi|$ grows polynomially with S, one will get a substantially better bound.

Remark 2. *Taking the limit in the PAC bound of Theorems 1–2, as $\beta \to 0$, yields corresponding sample complexity bounds for the risk-neutral case. The resulting bound for value learning is off the optimal bound by a factor of $(1-\gamma)^{-1}$, and for policy learning by a factor of $(1-\gamma)^{-1}\min\{S, (1-\gamma)^{-2}\}$; we again refer to Table 1 for a related comparison. It is worth stressing, however, that these implied bounds are valid for the entire ε-range, unlike the results in, e.g., [30, 1, 67].*

Remark 3. *Existing derivations of minimax sample complexity bounds in the risk-neutral setting (e.g., [30, 1, 67, 47]) rely on techniques that crucially exploit the additive structure of the return with respect to rewards, such as variance lemmas establishing Bellman consistency of the variance of cumulative discounted rewards. These tools do not extend to ERM due to its intrinsic non-linearity, and are therefore not applicable in our setting.*

6 Proofs: Sample Complexity Upper Bounds

In this section, we prove Theorems 1 and 2. As preliminarily, we present some results that will be used in the proofs. The first one concerns basic decompositions of the error terms associated to V^{π_k} and Q_k. Let $\widehat{V}^*$ and $\widehat{Q}^*$ denote the optimal value and Q-value in $\widehat{M}$, respectively, and π^* denote an optimal policy in M. Then, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$Q_k(s,a) \geq Q^*(s,a) - \|\widehat{Q}^{\pi^*} - Q^*\| - \|Q_k - \widehat{Q}^*\|\,, \tag{8}$$

$$V^{\pi_k}(s) \geq V^*(s) - \|V^{\pi_k} - \widehat{V}^{\pi_k}\| - \|\widehat{V}^{\pi^*} - V^*\| - \|\widehat{V}^{\pi_k} - \widehat{V}^*\|\,. \tag{9}$$

These follow from standard techniques, but for completeness, we derive them in Lemma 6 in Appendix C.

The second result concerns smoothness of Q-values under ERM when the transition function is perturbed. More specifically, it asserts how different Q-values of a fixed policy are in two different MDPs that differ only slightly in their transition functions. This parallels the result in [41, 70] to ERMs.

Lemma 3 (Q-value smoothness under ERM)**.** *Consider two MDPs $M = (S, A, P, R, \gamma, \beta)$ and $\widetilde{M} = (S, A, \widetilde{P}, R, \gamma, \beta)$ differing only in their transition functions. Fix a stationary policy π, and let Q^π and $\widetilde{Q}^\pi$ be respective Q-values of π in M and $\widetilde{M}$. It holds that $\|Q^\pi - \widetilde{Q}^\pi\| \leq \xi W_1$ for $\beta < 0$, and $\|Q^\pi - \widetilde{Q}^\pi\| \leq \xi W_2$ for $\beta > 0$, where $\xi := \frac{\gamma}{(1-\gamma)|\beta|} e^{|\beta|/(1-\gamma)}$ and*

$$W_1 := \max_{s,a} \Big| \sum_{s' \in \mathcal{S}} [P_{s,a}(s') - \widetilde{P}_{s,a}(s')] e^{-|\beta|(\frac{1}{1-\gamma} - V^\pi(s'))} \Big|,$$

$$W_2 := \max_{s,a} \Big| \sum_{s' \in \mathcal{S}} [P_{s,a}(s') - \widetilde{P}_{s,a}(s')] e^{-|\beta| V^\pi(s')} \Big|, \qquad \textit{with } V^\pi(s) = \max_a Q^\pi(s,a).$$

6.1 Proof of Theorem 1

Proof. Let $\beta > 0$ and $\varepsilon > 0$. In view of the error decomposition in (8), to establish ε-value-correctness it suffices to ensure $\|\widehat{Q}^{\pi^*} - Q^*\| \leq \varepsilon/2$ and $\|Q_k - \widehat{Q}^*\| \leq \varepsilon/2$. By Lemma 2, we can have $\|Q_k - \widehat{Q}^*\| < \varepsilon/2$ using enough iterations of the optimization oracle. Further, by Lemma 3, if

$$\max_{s,a} \Big| \sum_{s' \in \mathcal{S}} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] e^{-|\beta| V^*(s')} \Big| < \frac{\varepsilon(1-\gamma)|\beta|}{2\gamma} e^{-|\beta|/(1-\gamma)} =: \tau, \tag{10}$$

then $\|\widehat{Q}^{\pi^*} - Q^*\| \leq \varepsilon/2$. To ensure this, we use the following lemma (proven in Appendix D):

Lemma 4. *Let π be any fixed policy and $\tau > 0$. If $N > \frac{1}{2\tau^2}\big(1 - e^{-|\beta|/(1-\gamma)}\big)^2 \log(2SA/\delta)$, then it holds that*

$$(i) \qquad \max_{s,a} \Big| \sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] e^{-\beta V^\pi(s')} \Big| < \tau\,, \quad \textit{with probability} \geq 1-\delta, \quad \beta > 0;$$

$$(ii) \qquad \max_{s,a} \Big| \sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] e^{-|\beta|(V^\pi(s') - \frac{1}{1-\gamma})} \Big| < \tau \quad \textit{with probability} \geq 1-\delta, \quad \beta < 0.$$

Applying Lemma 4, inequality (i), with $\pi = \pi^\star$, we observe that (10) holds with probability at least $1-\delta$ by picking $N \geq \frac{1}{2\tau^2}\big(1 - e^{-|\beta|/(1-\gamma)}\big)^2 \log(2SA/\delta)$. Noting that the total calls to the generative model is $T = SAN$ and substituting in the value for τ, we can ensure for all (s,a) that $Q_k(s,a) > Q^*(s,a) - \varepsilon$ with probability larger than $1-\delta$ by using a total number of samples

$$T \geq \frac{2SA\gamma^2}{\varepsilon^2(1-\gamma)^2} \left(\frac{e^{|\beta|/(1-\gamma)} - 1}{|\beta|} \right)^2 \log\left(\frac{2SA}{\delta}\right).$$

The case of $\beta < 0$ is proven using very similar lines, but will use inequality (ii) in Lemma 4. □

6.1.1 Proof of Theorem 2

Proof. Let $\varepsilon > 0$ and $\beta \neq 0$. In view of the error decomposition in (9), to establish ε-policy-correctness it suffices to require: (i) $\|\widehat{V}^{\pi_k} - \widehat{V}^*\| \leq \varepsilon/3$, (ii) $\|\widehat{V}^{\pi^*} - V^*\| \leq \varepsilon/3$, and (iii) $\|V^{\pi_k} - \widehat{V}^{\pi_k}\| \leq \varepsilon/3$. For (i), we can have $\|\widehat{V}^{\pi_k} - \widehat{V}^*\| \leq \varepsilon/3$ using enough iterations of the optimization oracle as a consequence of Lemma 2. To control (ii) and (iii), let us define the event

$$E := \cap_{\pi \in \Pi} \big\{ \|V^\pi - \widehat{V}^\pi\| \leq \varepsilon/3 \big\}.$$

We can show that $\mathbb{P}(E) \geq 1 - \delta$ for sufficiently large N. Indeed, for a given $\pi \in \Pi$, an application of Lemma 3 and Lemma 4, identical to the treatment in the proof of Theorem 2, shows that $\|V^\pi - \widehat{V}^\pi\| \leq \varepsilon/3$ with probability at least $1 - \delta/|\Pi|$ if $N \geq \frac{1}{2\tau^2}\big(1 - e^{-\beta/(1-\gamma)}\big)^2 \log(2SA|\Pi|/\delta)$ with $\tau = \frac{\varepsilon(1-\gamma)|\beta|}{3\gamma} e^{-|\beta|/(1-\gamma)}$. Hence,

$$\mathbb{P}(E^c) = \mathbb{P}\big(\exists \pi \in \Pi : \|V^\pi - \widehat{V}^\pi\| > \varepsilon/3\big) \leq \sum_{\pi \in \Pi} \mathbb{P}\big(\|V^\pi - \widehat{V}^\pi\| > \varepsilon/3\big) \leq \delta,$$

so that $\mathbb{P}(E) \geq 1 - \delta$. It is evident that conditioned on E, (ii) and (iii) holds. Hence, with probability greater than $1 - \delta$, ε-policy-correctness is maintained using a total number of model calls of

$$T_{1,\delta} = \frac{9SA\gamma^2}{\varepsilon^2(1-\gamma)^2} \left(\frac{e^{|\beta|/(1-\gamma)} - 1}{|\beta|} \right)^2 \log\left(\frac{2SA|\Pi|}{\delta} \right). \tag{11}$$

To establish the second bound, first observe that Theorem 1 implies that $\|\widehat{V}^{\pi_k} - V^*\| \leq \varepsilon$ with probability exceeding $1 - \delta$ if $T \geq \frac{2SA\gamma^2}{\varepsilon^2(1-\gamma)^2|\beta|^2}\big(e^{|\beta|/(1-\gamma)} - 1\big)^2 \log\big(\frac{2SA}{\delta}\big)$; this is verified by noting that

$$\|\widehat{V}^{\pi_k} - V^*\| = \max_s \big| \max_a \widehat{Q}^{\pi_k}(s,a) - \max_a Q^*(s,a) \big| \leq \max_{s,a} \big| \widehat{Q}^{\pi_k}(s,a) - Q^*(s,a) \big| = \|Q_k - Q^*\|.$$

To proceed, we make use of the following lemma, which is proven in Appendix D:

Lemma 5. *Let $\alpha > 0$. Let $\overline{V} \in \mathbb{R}^S$ be a value function obeying $\|V^* - \overline{V}\| < \alpha$, and $\pi^G := \mathrm{argmax}_a [R(s,a) + \gamma \rho_{s,a}(\overline{V}(s'))]$ be a greedy policy with respect to $\overline{V}$. Then, $\|V^* - V^{\pi_G}\| \leq \frac{2\gamma}{1-\gamma}\alpha$.*

Choosing $\overline{V} = \widehat{V}^{\pi_k}$, we have shown that $\|\widehat{V}^{\pi_k} - V^*\| \leq \varepsilon$ with probability $1 - \delta$. Further, note that π_k by construction is the greedy policy with respect to $\widehat{V}^{\pi_k}$. Therefore, by Lemma 5, the true value of π_k satisfies, with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_k}\| \leq \frac{2\gamma}{1-\gamma}\varepsilon,$$

which yields the following bound:

$$T_{2,\delta} = \frac{8SA\gamma^2}{\varepsilon^2(1-\gamma)^4} \left(\frac{e^{|\beta|/(1-\gamma)} - 1}{|\beta|} \right)^2 \log\left(\frac{2SA}{\delta} \right). \tag{12}$$

Final Bound. To derive the final bound, we put together (11) and (12), while suitably adjusting the error probabilities. Therefore, ε-policy-correctness is guaranteed with probability exceeding $1 - \delta$ if $T \geq \min\{T_{1,\delta/2}, T_{2,\delta/2}\}$. □

7 Sample Complexity Lower Bounds

In this section, we provide two sample complexity lower bounds. The first (Theorem 3) concerns value learning, whereas the second (Theorem 4) addresses policy learning.

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

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

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

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

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

While analogous policy learning lower bounds are often stated in the risk-neutral literature, explicit proofs are typically omitted, to the best of our knowledge. For completeness, we provide a detailed, step-by-step derivation, emphasizing its close connection to the corresponding value-learning lower bound as well as the subtle differences that arise in the final guarantee.

Theorems 3–4 establish that an exponential dependence on the effective horizon $1/(1-\gamma)$ in the sample complexity is unavoidable under both value and policy learning. These bounds cover both risk-averse ($\beta > 0$) and risk-seeking ($\beta < 0$) agents, providing strong impossibility results for recursive ERM. Comparing with the sample complexity bounds of MB-ERM-QVI (Theorems 1–2), we observe a similar exponential dependence; however, a gap of order $\frac{1}{(1-\gamma)^2}e^{|\beta|/(1-\gamma)}$ remains. Closing this gap may require either a refined analysis of MB-ERM-QVI or more sophisticated algorithmic ideas. Nevertheless, these lower bounds confirm that risk-sensitive RL under ERM is fundamentally harder in the worst case than the risk-neutral setting, where minimax sample complexity scales polynomially with $1/(1-\gamma)$.

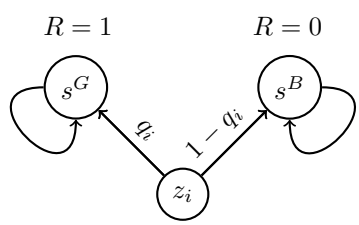


Figure 1: Hard-to-learn MDP construction

Proof sketch. The proofs are provided in Appendix F; here we sketch the main ideas for value learning (Theorem 3). The construction involves a class of hard-to-learn MDPs (Figure 1) with two absorbing states s^G and s^B, yielding rewards $R = 1$ and $R = 0$ under any action, respectively. All other state-action pairs z give zero reward and transition only to s^G or s^B with probability $P(s^G|z) = q$ and $P(s^B|z) = 1-q$, 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. It is key to choose them in a way 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 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 $z \neq z'$.

We note that, in the course of this analysis, we also correct a minor issue in Lemma 17 of [30]. Specifically, the issue arises in the derivation of a lower bound on the likelihood ratio between two Bernoulli distributions with means $p \geq \frac{1}{2}$ and $p+\alpha$ on a high probability event, for $p > \frac{1}{2}$. Additionally, we establish a corresponding bound for $p < \frac{1}{2}$, which may be of independent interest. For policy learning, a similar construction is used, augmented with a known action a_0 to facilitate the analysis.

Algorithmic intuition. The above construction also provides insight into how a model-based algorithm such as MB-ERM-QVI behaves on these instances. Since all rewards are zero except in the absorbing states, learning reduces to estimating the transition probability $q = P(s^G|z)$ for each state-action pair z. The algorithm therefore forms an empirical estimate $\hat{q}$ and computes Q-values based on this estimate. On the hard instances M_0 and M_1, the true probabilities differ only by a small amount ($q_0 = p$ vs. $q_1 = p+\alpha$). When the number of samples is limited, the empirical estimate $\widehat{q}$ will typically not be accurate enough to reliably distinguish between these two cases. As a result, the algorithm may construct a model that is statistically consistent with both M_0 and M_1, leading to Q-value estimates that are necessarily inaccurate for at least one of them.

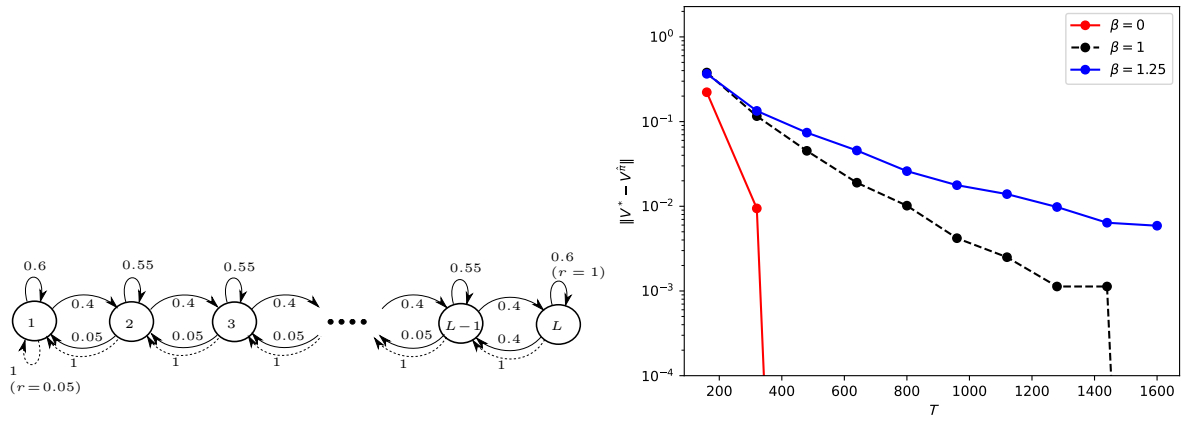
Remark 4. *The best lower bound in the risk-neutral setting is derived in [30] using a richer construction than above. However, with a risk-sensitive learning objective, the optimal Q-value function in that construction does not admit an analytical solution, which is crucial for tuning transition probabilities and deriving our bounds.*

Remark 5. *We note that the bound becomes vacuous for $|\beta| \leq (1-\gamma)\log(3)$. This is partly due to a final approximation introduced to make the bound more interpretable; importantly, this approximation does not affect the dependence on $|\beta|$ for large $|\beta|$. Even without this approximation, the bound still becomes vacuous as $\beta \to 0$. This behavior arises because, in this limit, $p \to 1$ or $p \to 0$ depending on the direction of the limit. Since our information-theoretic argument yields a sample complexity proportional to $p(1-p)$, the bound vanishes in this regime.*

8 Numerical experiments

We conduct a numerical experiments to showcase the empirical performance of MB-ERM-QVI. We consider a RiverSwim MDP [70] (Figure 2a) with 8 states with discount factor $\gamma = 0.95$. In this MDP, there two actions in each state, corresponding to moving 'left' or 'right'. The rewards are zero, except in two places: a reward of 0.05 (low reward) under 'left' in the left-most state ($s = 1$), and reward of 1 (high reward) under 'right' in the right-most state ($s = L$). The low reward is easy to access because of the actions with deterministic transitions. The agent has a risk-sensitive objective defined using recursive ERM. To showcase the performance of MB-ERM-QVI, we consider three different values of $\beta \in \{0, 1, 1.25\}$; we recall that $\beta = 0$ corresponds to the risk-neutral case. For each value of β, we generated datasets of increasing sizes $T \in \{160, 320, \ldots, 1600\}$ (multiples of $SA = 16$) by sampling each state-action pair in the MDP. The MB-ERM-QVI algorithm was run on each dataset to produce an output policy $\hat{\pi}$. This procedure was run for 1000 independent runs.

Figure 2 depicts $\|V^* - V^{\hat{\pi}}\|$ averaged over the 1000 runs for the three values of β. The true value of $\hat{\pi}$ is computed using a value iteration procedure based on (7). In this figure, one may consider a particular value of ε to observe the number of samples needs to output an ε-optimal policy. It is evident that this number of samples increases as β increases. Furthermore, it is evident that even a small increase in β (from 1 to 1.25) leads to a large increase in the number of samples to learn an ε-optimal policy. We also note that when the output policy $\widehat{\pi}$ matches the optimal policy for all the runs, the curve would hit the horizontal axis. These results showcases that learning a near-optimal policy under ERM could be substantially more demanding in terms of data even in such a rather simple MDP, and the data efficiency may be severely impacted as the risk parameter β increases.



(a) (b)

Figure 2: (a) The L-state RiverSwim MDP [70]; (b) Policy learning error $\|V^* - V^{\hat{\pi}}\|$ under MB-ERM-QVI.

9 Concluding Remarks

We studied the sample complexities of value learning and policy learning in finite discounted MDPs, where the agent exhibits recursive risk preferences modeled via the ERM and has access to a generative model. The generative model setting is commonly used in theoretical RL as it provides a clean framework in which the statistical difficulty of the problem can be isolated and precisely characterized. In particular, it removes effects due to exploration and data collection, allowing us to focus on the intrinsic difficulty induced by the risk-sensitive objective. We introduced a model-based algorithm, MB-ERM-QVI, and derived PAC-type bounds on its sample complexity. In addition, we established sample complexity lower bounds for both policy and value learning. These lower bounds reveal that an exponential dependence on the horizon $1/(1-\gamma)$ is unavoidable in the worst case, demonstrating that this setting is fundamentally harder than the risk-neutral case. To our knowledge, these constitute the first upper and lower bounds for this setting. The bounds are tight in S, A, δ, and ε, but gaps remain in the dependence on $1/(1-\gamma)$.

Closing these remaining gaps leads to interesting directions for future work. Firstly, our lower bound is valid for $|\beta| > (1-\gamma)\log(3)$, and it would therefore be interesting to derive a lower bound that would uncover the risk-neutral lower bound as β tends to zero. It is not immediately clear whether this requires improving the upper bound, the lower bound, or both. Designing hard MDP instances for lower bounds is subtle: they should be difficult to learn, yet still tractable enough to compute value functions; this is a challenge amplified in the risk-sensitive setting. The plug-in estimator used in MB-ERM-QVI is provably optimal in the risk-neutral case, raising the natural question of whether this optimality extends to ERM. Another promising direction is the development of model-free algorithms for this setting and the analysis of their statistical efficiency.

It would also be valuable to extend the study to MDPs with function approximation (e.g., [78]), as well as more complex RL settings, including offline RL [61], where data is collected under a fixed (but unknown) behavior policy, and online RL [70, 45], where the agent's learned policy directly impacts the data collection process. A representative next step is offline RL, where, under a model-based approach, plug-in estimators combined with pessimism-style corrections are often used (e.g., [48]). Some technical tools developed in this paper may be of independent interest and could potentially be useful in analyzing related offline RL settings.

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

[6] Nicole Bäuerle and Anna Jaśkiewicz. "Markov decision processes with risk-sensitive criteria: An overview". In: *Mathematical Methods of Operations Research* 99.1 (2024), pp. 141–178.

[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] Nicole Bäuerle and Ulrich Rieder. "More risk-sensitive Markov decision processes". In: *Mathematics of Operations Research* 39.1 (2014), pp. 105–120.

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

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

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

[12] Vivek S Borkar. "A sensitivity formula for risk-sensitive cost and the actor–critic algorithm". In: *Systems & Control Letters* 44.5 (2001), pp. 339–346.

[13] Vivek S Borkar. "Q-learning for risk-sensitive control". In: *Mathematics of operations research* 27.2 (2002), pp. 294–311.

[14] 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.

[15] 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.

[16] 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.

[17] Yin-Lam Chow and Marco Pavone. "A framework for time-consistent, risk-averse model predictive control: Theory and algorithms". In: *2014 American Control Conference*. IEEE. 2014, pp. 4204–4211.

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

[19] Yinlam Chow et al. "Risk-constrained reinforcement learning with percentile risk criteria". In: *Journal of Machine Learning Research* 18.167 (2018), pp. 1–51.

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

[21] Michael Alan Howarth Dempster. *Risk management: value at risk and beyond*. Cambridge University Press, 2002.

[22] Zilong Deng, Simon Khan, and Shaofeng Zou. "Near-Optimal Sample Complexity for Iterated CVaR Reinforcement Learning with a Generative Model". In: *The 28th International Conference on Artificial Intelligence and Statistics*. 2025.

[23] 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.

[24] 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.

[25] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. "Risk-sensitive reinforcement learning with function approximation: A debiasing approach". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3198–3207.

[26] Yingjie Fei et al. "Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 20436–20446.

[27] 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.

[28] Hans Föllmer and Alexander Schied. "Convex and coherent risk measures". In: *Encyclopedia of Quantitative Finance* (2010), pp. 355–363.

[29] Deep Kumar Ganguly et al. "Risk-Seeking Reinforcement Learning via Multi-Timescale EVaR Optimization". In: *Transactions on Machine Learning Research* ().

[30] 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.

[31] 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.

[32] Jia Lin Hau et al. "On dynamic programming decompositions of static risk measures in Markov decision processes". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 51734–51757.

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

[34] 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.

[35] Yilie Huang, Yanwei Jia, and Xunyu Zhou. "Achieving mean–variance efficiency by continuous-time reinforcement learning". In: *Proceedings of the Third ACM International Conference on AI in Finance*. 2022, pp. 377–385.

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

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

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

[39] Danial Kamran et al. "Risk-aware high-level decisions for automated driving at occluded intersections with reinforcement learning". In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2020, pp. 1205–1212.

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

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

[42] 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.

[43] Prashanth La and Mohammad Ghavamzadeh. "Actor-critic algorithms for risk-sensitive MDPs". In: *Advances in neural information processing systems* 26 (2013).

[44] 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.

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

[46] Duan Li and Wan-Lung Ng. "Optimal dynamic portfolio selection: Multiperiod mean-variance formulation". In: *Mathematical finance* 10.3 (2000), pp. 387–406.

[47] 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.

[48] Gen Li et al. "Breaking the sample size barrier in model-based reinforcement learning with a generative model". In: *Operations Research* 72.1 (2024), pp. 203–221.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*. Vol. 1. Oxford University Press, 1995.

[55] Oliver Mihatsch and Ralph Neuneier. "Risk-sensitive reinforcement learning". In: *Machine learning* 49.2 (2002), pp. 267–290.

[56] Mehrdad Moharrami et al. "A policy gradient algorithm for the risk-sensitive exponential cost mdp". In: *Mathematics of operations research* 50.1 (2025), pp. 431–458.

[57] Xinyi Ni and Lifeng Lai. "Risk-sensitive reinforcement learning via Entropic-VaR optimization". In: *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2022, pp. 953–959.

[58] Xinyi Ni, Guanlin Liu, and Lifeng Lai. "Risk-sensitive reward-free reinforcement learning with CVaR". In: *Forty-first International Conference on Machine Learning*. 2024.

[59] Marek Petrik and Dharmashankar Subramanian. "An approximate solution method for large risk-averse markov decision processes". In: *Conference on Uncertainty in Artificial Intelligence*. 2012.

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

[61] 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.

[62] Marc Rigter, Bruno Lacerda, and Nick Hawes. "One risk to rule them all: A risk-sensitive perspective on model-based offline reinforcement learning". In: *Advances in neural information processing systems* 36 (2023), pp. 77520–77545.

[63] Utsav Sadana, Erick Delage, and Angelos Georghiou. "Mitigating optimistic bias in entropic risk estimation and optimization with an application to insurance". In: *arXiv preprint arXiv:2409.19926* (2024).

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

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

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

[67] 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.

[68] Satinder P Singh and Richard C Yee. "An upper bound on the loss from approximate optimal-value functions". In: *Machine Learning* 16.3 (1994), pp. 227–233.

[69] Srijan Sood et al. "Deep reinforcement learning for optimal portfolio allocation: A comparative study with mean-variance optimization". In: *FinPlan* 2023.2023 (2023), p. 21.

[70] 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.

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

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

[73] Kaiwen Wang et al. "A Reductions Approach to Risk-Sensitive Reinforcement Learning with Optimized Certainty Equivalents". In: *Forty-second International Conference on Machine Learning*. 2025.

[74] 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.

[75] Wenhao Xu, Xuefeng Gao, and Xuedong He. "Regret bounds for Markov decision processes with recursive optimized certainty equivalents". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 38400–38427.

[76] 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).

[77] Yujie Zhao et al. "RA-PbRL: Provably efficient risk-aware preference-based reinforcement learning". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 60835–60871.

[78] 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.

A Risk Measures

In this section, we give a very brief introduction to risk measures. In the actuarial and mathematical finance literature, working with both losses and reward is common. A good introduction is [28], which like us uses the rewards formulation. Due to this ambiguity in the literature, we here collect some precise definitions for the reward setting, and then list some of the most important risk measures.

To begin with, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a background probability space, and $\mathcal{M}$ some convex cone of random variables defined on the background space. That is, for any $X, Y \in \mathcal{M}$ and $\lambda > 0$, it holds that $X + Y \in \mathcal{M}$ and $\lambda X \in \mathcal{M}$.

Definition 2 (Risk measure)**.** *A functional $\psi : \mathcal{M} \to \mathbb{R}$ is said to be a* risk measure *if it satisfies the following properties:*

$$\begin{aligned} &\psi(0) = 0, && \textit{(Normalization)} \\ &\textit{if } X \leq Y \textit{ then } \psi(X) \geq \psi(Y), && \textit{(Monotonicity)} \\ &\psi(X + c) = \psi(X) - c, \quad \forall c \in \mathbb{R}. && \textit{(Translation invariance)} \end{aligned}$$

If, in addition, ψ satisfies the properties

$$\begin{aligned} &\psi(cX) = c\psi(X), \quad \forall c > 0, && \textit{(Positive homogeneity)} \\ &\psi(X + Y) \leq \psi(X) + \psi(Y), && \textit{(Sub-additivity)} \end{aligned}$$

it is called a coherent risk measure*. A weaker notion is* convex risk measure*, which is one obeying*

$$\psi(\lambda X + (1 - \lambda)Y) \leq \lambda\psi(X) + (1 - \lambda)\psi(Y), \quad \forall \lambda \in [0, 1]\,. \qquad \textit{(Convexity)}$$

Finally, a risk-measure ψ is called law-invariant *if $\psi(X)$ only depends on the distribution of X under $\mathbb{P}$.*

We now mention some examples of risk measures.

Entropic Risk Measure (ERM). The risk measure given by

$$\mathrm{ERM}_\beta(X) = \frac{1}{\beta}\log\left(\mathbb{E}[e^{-\beta X}]\right)$$

is known as the entropic risk measure (ERM) with parameter $\beta \neq 0$. Notably, ERM is not coherent (see, e.g., [28]) as it is does not satisfy the positive homogeneity property. Letting $\beta \to 0$ one recovers the expectation $\mathbb{E}[X]$, and letting $\beta \to \infty$ yields the essential infimum risk measure.

Value-at-Risk (VaR). The risk measure given by

$$\mathrm{VaR}_\alpha(X) := q_\alpha(X) := \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}$$

is called the Value-at-Risk (VaR) at level $\alpha \in (0,1)$. VaR is in general not sub-additive, and hence also not coherent.

Conditional Value-at-Risk (CVaR). The risk measure given by

$$\mathrm{CVaR}_\alpha(X) := \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)du$$

is known as the Conditional Value-at-Risk (CVaR), or sometimes as the expected shortfall (ES). It is known to be a coherent risk-measure.

All the examples so far are evidently law-invariant.

It is worth highlighting that the actual functional ρ used to rank random variables is the *negative* of the ERM-risk measure $\mathrm{ERM}_\beta(X)$, introduced above, with the interpretation being that a lower quantity of risk is preferable. More formally, we consider the functional $\rho : \mathcal{M} \to \mathbb{R}$ given by $\rho(X) := -\mathrm{ERM}_\beta(X)$, featuring the following properties:

$$\begin{aligned}
&\rho(0) = 0, && \text{(Normalization)}\\
&\text{if } X \leq Y \text{ then } \rho(X) \leq \rho(Y), && \text{(Monotonicity)}\\
&\rho(X + c) = \rho(X) + c. && \text{(Translation invariance)}
\end{aligned}$$

It is common in the literature to overload notation and also refer to ρ as the ERM and we will do so and henceforth we will no longer work directly with risk measures, but only with this specific functional ρ. It follows immediately from the normalization and translation invariance that $\rho(c) = c$ for any $c \in \mathbb{R}$.

We will often use the short-hand notation $\rho_{s,a}(V(s'))$ as ρ applied to the random variable X that takes on the values $\{V(s')\}_{s' \in S}$ with probabilities $\mathbb{P}(X = V(s')) = P(s'|s,a)$.

B Bellman Optimality and Bellman Recursions

In this section, we properly define the state-value functions and state-action value functions of any possibly history-dependent policy π and show that the problem of finding an optimal policy can be achieved by a stationary policy and that the value functions satisfy Bellman recursions when the value functions are defined iteratively with respect to the ERM. Several similar results exist in the literature, e.g., [4] and [6] that also cover the case of $\beta > 0$. These results are derived under more general assumptions on $\mathcal{S}$ and $\mathcal{A}$. These general assumptions are trivially satisfied when $\mathcal{S}$ and $\mathcal{A}$ are finite but their proofs require assumptions on the functionals to ensure the existence of a stationary optimal policy usually by invoking a measurable selection theorem. We avoid this complication by only considering finite $\mathcal{S}$ and $\mathcal{A}$ and we in turn also give the first proof for state-action value functions and not just for value-functions, which is needed as we consider the problem of learning.

Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$ be a finite MDP with ρ being the ERM, and $R(s,a) \in [0,1]$ a deterministic reward function. Let $D = \mathcal{S} \times \mathcal{A}$, $H_1 = \mathcal{S}$ and $H_k = D^{k-1} \times \mathcal{S}$ for $k \geq 2$ the set of all possible histories up to

stage k. A policy $\pi = (\pi_k)_{k\in\mathbb{N}}$ is a sequence of maps $\pi_k : H_k \to \mathcal{A}$. We denote the set of all policies by Π, and identify the set of all stationary policies with the set of measurable maps F from $\mathcal{S}$ to $\mathcal{A}$, which is simply the set of all maps from $\mathcal{S}$ to $\mathcal{A}$ since all maps between finite sets that are both equipped with the discrete topology are measurable with respect to the induced Borel σ-algebras. Let $B(H_k)$ be the set of maps $V_k : H_k \to \mathbb{R}$ equipped with the supremum norm, and let $\pi = (\pi_k)_{k\in\mathbb{N}}$ be any policy. For any $V_{k+1} \in B(H_{k+1})$ and $h_k \in H_k$, we denote by $\rho_{h_k,\pi_k}(V_{k+1})$ the functional ρ applied to the random variable concentrated on the set $\{V_{k+1}(h_k, \pi_k(h_k), s')\}_{s'\in S}$ with $\mathbb{P}(s_{k+1} = s') = \mathbb{P}(s'|s_k, \pi_k(h_k))$. By monotonicity of ρ, we get that $\rho_{h_k,\pi_k}(V_{k+1}) \leq \|V_{k+1}\|$.

Next, we define the operators $L_{\pi_k} : B(H_{k+1}) \to B(H_k)$ by

$$(L_{\pi_k} V_{k+1})(h_k) = L_{\pi_k,V_{k+1}}(h_k) := R(s_k, \pi_k(h_k)) + \gamma\rho_{h_k,\pi_k}(V_{k+1}).$$

Similarly, we define $L_a : B(H_{k+1}) \to B(H_k)$ by

$$(L_a V_{k+1})(h_k) = L_{a,V_{k+1}}(h_k) := R(s_k, a) + \gamma\rho_{s_k,a}(V_{k+1}),$$

with $\rho_{s_k,a}$ defined analogously as ρ_{h_k,π_k} as above. By the basic properties of risk-measures, it follows directly that $0 \leq L_{\pi_k V_{k+1}}(h_k) \leq 1 + \gamma\|V_{k+1}\|$ and similarly for L_a.

For any initial state $s_0 = s$, we define the N-step discounted utility as

$$J_N(s, a, \pi) := (L_a \circ L_{\pi_2} \circ \cdots \circ L_{\pi_N})\mathbf{0}(s)$$

where $\mathbf{0}(h_k) = 0$ for all $h_k \in H_k$ and all $k \in \mathbb{N}$.

By monotonicity of ρ, it holds that the sequence $(J_N(s, a, \pi))_{N\in\mathbb{N}}$ is non-decreasing and bounded in the interval $[0, \frac{1}{1-\gamma}]$ for any $s, a, \pi \in \mathcal{S} \times \mathcal{A} \times \Pi$, and so the limit

$$J(s, a, \pi) := \lim_{N\to\infty} J_N(s, a, \pi)$$

exists for any state s, any action a, and any policy π.

The agent wishes to find $J^*(s, a) = \sup_{\pi\in\Pi} J(s, a, \pi)$ and an optimal policy π^* attaining $J^*(s, a)$, namely, $J(s, a, \pi^*) = J^*(s, a)$.

Theorem 5. *There exist a unique non-negative function $Q \in B(\mathcal{S}\times\mathcal{A})$ (non-negative map from $\mathcal{S}\times\mathcal{A} \to \infty$ equipped with sup-norm) and a stationary decision rule $f^* : \mathcal{S} \to \mathcal{A}$ such that*

$$\begin{aligned} Q(s, a) &= R(s, a) + \gamma\rho_{s,a}(\max_{a'} Q(s', a')), \\ &= R(s, a) + \gamma\rho_{s,a}(Q(s', f^*(s'))). \end{aligned}$$

Moreover, $Q(s, a) = J^*(s, a) = J(s, a, f^*)$ meaning that f^* is an optimal stationary policy.

Proof. We start by proving existence of Q. Let $L : B(\mathcal{S} \times \mathcal{A}) \to B(\mathcal{S} \times \mathcal{A})$ denote the operator given by

$$LQ(s, a) := R(s, a) + \gamma\rho_{s,a}(\max_{a'} Q(s', a')).$$

Let $Q_1, Q_2 \in B(\mathcal{S} \times \mathcal{A})$. We then have for all (s, a) that

$$\begin{aligned} LQ_1(s, a) - LQ_2(s, a) &= \gamma[\rho_{s,a}(\max_{a'} Q_1(s', a')) - \rho_{s,a}(\max_{a'} Q_2(s', a'))] \\ &= \gamma[\rho_{s,a}(\max_{a'} Q_1(s', a')) - \rho_{s,a}(\max_{a'} Q_1(s', a') - \max_{a'} Q_1(s', a') + \max_{a'} Q_2(s', a'))] \\ &\leq \gamma[\rho_{s,a}(\max_{a'} Q_1(s', a')) - \rho_{s,a}(\max_{a'} Q_1(s', a') + \max_{a'}\{Q_2(s', a') - Q_1(s', a')\})] \\ &\leq \gamma[\rho_{s,a}(\max_{a'} Q_1(s', a')) - \rho_{s,a}(\max_{a'} Q_1(s', a') + \|Q_1 - Q_2\|)] \\ &= \gamma\|Q_1 - Q_2\|. \end{aligned}$$

We start by showing that $L : B(\mathcal{S} \times \mathcal{A}) \to B(\mathcal{S} \times \mathcal{A})$, that is, it takes non-negative functions and returns non-negative functions. By normalization and monotonicity of ρ, we have for any non-negative $Q \in B(\mathcal{S} \times \mathcal{A})$ that

$$LQ(s,a) = R(s,a) + \gamma \rho_{s,a}(\max_{a'} Q(s,a)) \geq 0 + \gamma \rho_{s,a}(0) = 0.$$

By a completely similar argument, we have $LQ_2(s,a) - LQ_1(s,a) \leq \gamma \|Q_1 - Q_2\|$, so that L is a contraction, and since for $\mathcal{S} \times \mathcal{A}$ we can identify $B(\mathcal{S} \times \mathcal{A})$ with the closed subset of the complete metric space$(\mathbb{R}^{S \times A}, \|\cdot\|)$ that consists of vectors with non-negative coordinates. Since this subspace is closed, it is also a complete metric space and the existence of Q then follows from the Banach fixed-point theorem.

Since there are only finitely many states and actions, we can pick a stationary decision rule where $f^*(s)$ is an arbitrary element of $\operatorname{argmax}_a Q(s,a)$.

Let V be the function given by $V(s) := Q(s, f^*(s))$ for all s. We then observe that

$$V(s) \geq R(s,a) + \gamma \rho_{s,a}(V(s'))$$

for every $s \in \mathcal{S}$. Let $\pi = (\pi_k)_k \in \mathbb{N}$ be any policy in Π. The above inequality then shows that for any history $h_k, k \in \mathbb{N}$ we have that $V(s_k) \geq L_{\pi_k} V(h_k)$ and furthermore we note that $Q(s_1, a) = L_a V(h_1)$. This implies for any $N \in \mathbb{N}$ that

$$Q(s,a) \geq (L_a \circ L_{\pi_2} \circ \cdots \circ L_{\pi_N}) V(s) \geq (L_a \circ L_{\pi_2} \circ \cdots \circ L_{\pi_N}) \mathbf{0}(s) = J_N(s,a,\pi)\,,$$

where we have used that $Q(s,a) \geq 0$. Finally taking the limit we find that $Q(s,a) \geq J(s,a,\pi)$.

Finally, we aim to show that $Q(s,a) \leq J(s,a,f^*)$. By induction, we wish to show that $V(s) \leq J_N(s, f^*(s), f^*) + \gamma^N \|V\|$ for all $N \in \mathbb{N}$. For the induction step, we start by noting that $J_1(s, f^*(s), f^*) = R(s, f^*(s))$ and so

$$\begin{aligned} V(s) &= R(s, f^*(s)) + \gamma \rho_{s, f^*(s)}(V(s')) \\ &\leq R(s, f^*(s)) + \gamma \rho_{s, f^*(s)}(\|V\|) \\ &= R(s, f^*(s)) + \gamma \|V\| \\ &= J_1(s, f^*(s), f^*) + \gamma \|V\|\,, \end{aligned}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. For the induction step, we assume that $V(s) \leq J_N(s, f^*(s), f^*) + \gamma^N \|V\|$. By using that $V(s) = L_{f^*} V(s)$ and that L is monotone, we see that

$$\begin{aligned} V(s) &= L_{f^*} V(s) \\ &\leq L_{f^*}(J_N(s, f^*(s), f^*) + \gamma^N \|V\|) \\ &= \big(R(\cdot, f^*(\cdot)) + \gamma \rho_{\cdot, f^*(\cdot)}(J_N(\cdot, f^*(\cdot), f^*) + \gamma^N \|V\|\big)(s) \\ &= J_{N+1}(s, f^*(s), f^*) + \gamma^{N+1} \|V\|\,, \end{aligned}$$

from which taking the limit as $N \to \infty$, we get that $V(s) \leq J(s, f^*(s), f^*)$.

Finally, since

$$Q(s,a) = L_a V(s) \leq L_a J(s, f^*(s), f^*) = J(s,a,f^*)\,,$$

the conclusion holds.

Since this shows that an optimal stationary policy exists, it will suffice to consider only stationary policies and one can by completely analogous arguments show that for any stationary policy π, there exists a non-negative map $Q^\pi \in B(\mathcal{S} \times \mathcal{A})$ such that $Q^\pi(s,a) = J(s,a,\pi)$, that is, Q^π satisfies the Bellman recursion:

$$Q^\pi(s,a) = R(s,a) + \gamma \rho_{s,a}(Q^\pi(s', \pi(s'))\,,$$

and similarly for state-value functions $V^\pi(s) := Q^\pi(s, \pi(s))$. □

We also remark that in the proof, we see directly that $Q(s,a) \in [0, \frac{1}{1-\gamma}]$ for all (s,a).

C 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-ERM-QVI, whose proof follows very similar lines to the proof of Lemma 3 in [1].

Lemma 6. *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 $s \in \mathcal{S}$,

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

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

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

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 7. *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 same technique after some elementary calculations.

Let $f(x) = \log(1-x)$ 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 8. *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. □

D Analysis of MB-ERM-QVI: Missing Proofs

D.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 $\beta > 0$, we then have

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

and the lemma follows. The proof for the case of $\beta < 0$ follows very similar lines, and is thus omitted. □

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

To show the other claim, we start by noting that $\|V^{\pi_k} - V^*\| \leq \|Q^{\pi_k} - Q^*\|$. Furthermore, 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,

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

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

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

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

D.3 Proof of Lemma 3

Proof. There are four cases to consider, corresponding to the combinations of $\beta > 0$ or $\beta < 0$ and whether, at the state–action pair (s, a) attaining the maximum, $Q_1(s, a) > Q_2(s, a)$ or $Q_2(s, a) > Q_1(s, a)$.

Case 1: $\beta > 0$ and $Q_1(s, a) > Q_2(s, a)$. We have

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

Rearranging the terms yields the asserted result:

$$\|Q_1 - Q_2\| \leq \frac{\gamma}{1-\gamma} \frac{e^{\beta/(1-\gamma)}}{\beta} \Big| \sum_{s'} [P_2(s'|s,a) - P_1(s'|s,a)] e^{-\beta V_1(s')} \Big| .$$

Case 2: $\beta > 0$ and $Q_1(s, a) < Q_2(s, a)$. The proof is very similar to Case 1, but the extension $V_2(s) = V_1(s) + V_2(s) - V_1(s)$ is now done in the numerator instead:

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

which again yields

$$\|Q_1 - Q_2\| \leq \frac{\gamma}{1-\gamma} \frac{e^{\beta/(1-\gamma)}}{\beta} \Big| \sum_{s'} [P_2(s'|s,a) - P_1(s'|s,a)] e^{-\beta V_1(s')} \Big| .$$

Case 3: $\beta < 0$ **and** $Q_1(s,a) > Q_2(s,a)$**.** We have

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

which implies

$$\|Q_1 - Q_2\| \leq \frac{\gamma}{1-\gamma} \frac{e^{|\beta|/(1-\gamma)}}{|\beta|} \Big|\sum_{s'} [P_2(s'|s,a) - P_1(s'|s,a)]e^{|\beta|[V_1(s') - \frac{1}{1-\gamma}]}\Big|.$$

Case 4: $\beta < 0$ **and** $Q_2(s,a) > Q_1(s,a)$**.** The proof of this case is similar to the other three cases and is omitted. □

D.4 Proof of Lemma 4

Let N denote the number of calls to the generative model on each state-action pair such that the total number of calls is SAN. Let $\widehat{P}(s'|s,a)$ denote the plug-in estimator obtained from N samples of $s' \sim P_{s,a}$, that is $\widehat{P}(s'|s,a) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{X_n = s'\}}$, where X_n taking values in $\mathcal{S}$ according to $P_{s,a}$.

Lemma 4. *Let π be any fixed policy and $\tau > 0$. If $N > \frac{1}{2\tau^2}\left(1 - e^{-|\beta|/(1-\gamma)}\right)^2 \log(2SA/\delta)$, then it holds that*

$$(i) \qquad \max_{s,a} \Big|\sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')]e^{-\beta V^\pi(s')}\Big| < \tau\,, \quad \textit{with probability} \geq 1-\delta, \quad \beta > 0;$$

$$(ii) \qquad \max_{s,a} \Big|\sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')]e^{-|\beta|(V^\pi(s') - \frac{1}{1-\gamma})}\Big| < \tau \quad \textit{with probability} \geq 1-\delta, \quad \beta < 0.$$

Proof. We only prove the first claim —i.e., the case of $\beta > 0$— as the other case is proven using completely similar lines. We note that for the random variable $\sum_{s'} \mathbb{1}_{\{X_n = s'\}} e^{-\beta V^\pi(s')}$, we have that

$$\begin{aligned}
\mathbb{E}\Big[\sum_{s'} \mathbb{1}_{\{X_n = s'\}} e^{-\beta V^\pi(s')}\Big] &= \sum_{s'} \mathbb{E}\big[\mathbb{1}_{\{X_n = s'\}}\big] e^{-\beta V^\pi(s')} \\
&= \sum_{s'} P(s'|s,a) e^{-\beta V^\pi(s')}
\end{aligned}$$

and that it is bounded in $[e^{-\beta/(1-\gamma)}, 1]$. Also, since

$$\sum_{s'} \widehat{P}(s'|s,a) e^{-\beta V^\pi(s')} = \frac{1}{N} \sum_{n=1}^{N} \sum_{s'} \mathbb{1}_{\{X_n = s'\}} e^{-\beta V^\pi(s')},$$

it follows directly from Hoeffding's inequality that

$$\mathbb{P}\Big(\Big|\sum_{s'}[P(s'|s,a)-\widehat{P}(s'|s,a)]e^{-\beta V^{\pi}(s')}\Big|\geq\varepsilon\Big)\leq 2\exp\Big(-\frac{2N\varepsilon^2}{\big(1-e^{-\beta/(1-\gamma)}\big)^2}\Big).$$

Thus, by picking $N=\frac{1}{2\varepsilon^2}\big(1-e^{-\beta/(1-\gamma)}\big)^2\log(2SA/\delta)$ and a union bound,

$$\mathbb{P}\Big(\max_{s,a}|\sum_{s'}[P(s'|s,a)-\widehat{P}(s'|s,a)]e^{-\beta V^{\pi}(s')}|\geq\varepsilon\Big)\leq\delta\,.$$

□

E Proof of Lemma 5

Next we prove a result that bounds the quality of a greedy policy with respect to the quality of the value-function for which the policy is greedy. The result is a generalization of [68] from the expectation to that of ERM and the derivation follow the same lines. Throughout, we use the notation $\rho_{s,a}(V(s'))$ as shorthand notation for ρ applied to the categorical random variable X with support $\{V(s')\}_{s'\in\mathcal{S}}$ where $\mathbb{P}(X=V(s'))=P(s'|s,a)$.

Lemma 5. *Let $\alpha>0$. Let $\overline{V}\in\mathbb{R}^S$ be a value function obeying $\|V^*-\overline{V}\|<\alpha$, and $\pi^G:=\operatorname{argmax}_a[R(s,a)+\gamma\rho_{s,a}(\overline{V}(s'))]$ be a greedy policy with respect to $\overline{V}$. Then, $\|V^*-V^{\pi_G}\|\leq\frac{2\gamma}{1-\gamma}\alpha$.*

Proof. Let $\bar{s}$ be a state such that $\|V^*-V^G\|=V^*(\bar{s})-V^G(\bar{s})$, where $V^G:=V^{\pi_G}$. We then consider the two actions $a^*:=\pi^*(\bar{s})$ and $a^G:=\pi^G(\bar{s})$; ties can be breaking arbitrarily. Since π^G is greedy with respect to V^G, we have that

$$R(\bar{s},a^*)+\gamma\rho_{\bar{s},a^*}(\widehat{V}(s'))\leq R(\bar{s},a^G)+\gamma\rho_{\bar{s},a^G}(\widehat{V}(s'))\,.$$

By assumption, it holds for any $s\in\mathcal{S}$ that

$$V^*(s)-\varepsilon\leq\widehat{V}(s)\leq V^*(s)+\varepsilon.$$

By monotonicity and translation invariance of ρ, we thus get

$$\begin{aligned}R(\bar{s},a^*)+\gamma\rho_{\bar{s},a^*}(\widehat{V}(s'))&\geq R(\bar{s},a^*)+\gamma\rho_{\bar{s},a^*}(V^*(s')-\varepsilon)\\&=R(\bar{s},a^*)+\gamma\rho_{\bar{s},a^*}(V^*(s'))-\gamma\varepsilon\,,\end{aligned}$$

and similarly we have

$$R(\bar{s},a^G)+\gamma\rho_{\bar{s},a^G}(\widehat{V}(s'))\leq R(\bar{s},a^G)+\gamma\rho_{\bar{s},a^G}(V^*(s'))+\gamma\varepsilon,$$

which collectively imply

$$R(\bar{s},a^*)-R(\bar{s},a^G)\leq 2\gamma\varepsilon+\gamma\big(\rho_{\bar{s},a^G}(V^*(s'))-\rho_{\bar{s},a^*}(V^*(s')\big)\,.$$

Finally, we obtain

$$\begin{aligned}V^*(\bar{s})-V^G(\bar{s})&=R(\bar{s},a^*)-R(\bar{s},a^G)+\gamma\rho_{\bar{s},a^*}(V^*(s'))-\gamma\rho_{\bar{s},a^G}(V^G(s'))\\&\leq 2\gamma\varepsilon+\gamma\rho_{\bar{s},a^G}(V^*(s'))-\gamma\rho_{\bar{s},a^*}(V^*(s')+\gamma\rho_{\bar{s},a^*}(V^*(s'))-\gamma\rho_{\bar{s},a^G}(V^G(s'))\\&=2\gamma\varepsilon+\gamma\big(\rho_{\bar{s},a^G}(V^*(s')-\rho_{\bar{s},a^G}(V^G(s'))\big)\\&=2\gamma\varepsilon+\gamma\|V^*-V^G\|,\end{aligned}$$

from which the result follows. □

F Proofs of Lower Bounds

F.1 Lower Bound on Bernoulli Likelihood Ratio

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 [30]; however, we stress that our treatment fixes an error in the proof, which however requires slightly stronger assumptions than those imposed in [30]. In addition, while the result in [30] 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

$$L_m(w) := \mathbb{P}_m(W = w) \tag{13}$$

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

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

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

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 7) 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

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

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

$$\frac{L_1(W)}{L_2(W)}\mathbb{1}_{\mathcal{E}} \geq \left(\frac{2\theta}{c_2}\right)^{\frac{2\sqrt{2}}{\sqrt{c_1}}+\frac{2(1-p)}{pc_1}+\frac{2}{c_1}}\mathbb{1}_{\mathcal{E}},$$

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

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

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 $c_1 = 32$:

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

□

F.2 Lower Bound for Q-value Learning

For a lower bound we construct the following class of MDPs with $S' := S + 2$ states and A actions where the first states are labelled $S_1, , ..., s_S, s^G, s^B$ and the actions are labelled $a_1, ..., a_A$. The states s^G and s^B are absorbing under any actions and $R(s^G, a) = 1$ for all j and $R(s^B, a) = 0$ for all $a \in \mathcal{A}$. For the states $s \in \{s_1, ..., s_S\}$, we have that $R(s, a) = 0$ for all $a \in \mathcal{A}$. We have SA state-action pair combinations from $\{s_1, ..., s_S\} \times \mathcal{A} =: Z$ on which we assume some ordering allowing us to write $z_i, i \in [SA]$. Finally for all state-action pairs $z_i \in [SA]$ we have $P(s^G|z_i) = q_i$ and $P(s^B|z_i) = 1 - q_i$ for some $q_i \in [0, 1]$. The structure of this class of MDPs allows us to get lower bounds on the samples needed to learn the Q-value of each state-action pair z_i and then use the fact that samples used to learn the Q-values for different state-action pairs bring no information on each other to get the final bound.

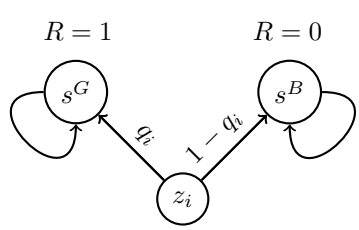


Figure 3: Dynamics and rewards of the hard-to-learn MDP class

For any state-action pair we can explicitly calculate the state-action value-functions

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

Denote the collection of all such MDPs by $\mathbb{M}$.

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

We use M_0 to denote an MDP where H_0^i holds 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 $Q^*_{M_1}(z_i) - Q^*_{M_0}(z_i) > 2\varepsilon$, 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

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

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

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

In particular, this means that the events $B_0 := \{|Q^*_{M_0}(z_i) - Q^{\mathcal{U}}_t(z_i)| \leq \varepsilon\}$ and $B_1 := \{|Q^*_{M_1}(z_i) - Q^{\mathcal{U}}_t(z_i)| \leq \varepsilon\}$ are disjoint events. Let t be the number of times the algorithm tries z_i. 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 to s^G in the t trials. We then define $\tilde{k}, \tilde{p}$ and θ by

$$\theta := \exp\Big(-\frac{32\alpha^2 t}{p(1-p)}\Big), \qquad \tilde{p} = \max\{p, 1-p\}, \qquad \tilde{k} := \begin{cases} k & \text{if } p \geq \frac{1}{2} \\ t-k & \text{if } p < \frac{1}{2} \end{cases}$$

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 [30] and thus $\mathbb{P}_0(B_0 \cap \mathcal{E}) > \frac{1}{2}$. Now by Theorem 6, we get that

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

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 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 $B_0 \subset B_1^c$.

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

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

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

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

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

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

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

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

when $\delta\frac{n}{2} \leq 1$ by Lemma 8. By setting $\delta' = \delta\frac{n}{4}$ and substituting back S' we obtain the result. This shows that if the number of samples is smaller than

$$
T = \frac{(S'-2)A}{4096}\frac{\gamma^2}{\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}} - 3}{|\beta|^2}\log\Big(\frac{(S'-2)A}{32\delta}\Big) \tag{14}
$$

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

F.3 Lower Bound for Policy Learning

For a lower bound we construct the following class of MDPs with $S' := S + 2$ states and $A' := A + 1$ actions where the first states are labelled $s_1, \ldots, s_S, s^G, s^B$ and the actions are labelled $a_0, a_1, ..., a_A$. The states s^G and s^B are absorbing under any actions and $R(s^G, a) = 1$ for all j and $R(s^B, a) = 0$ for all $a \in \mathcal{A}$. For the states $s \in \{s_1, ..., s_S\}$, we have that $R(s, a) = 0$ for all $a \in \mathcal{A}$.

From the state s_i with probabilities that depend on the action taken the agent will then end up in either a good state s^G which is absorbing and yields the maximal unit reward under all actions or in the bad state s^B which is also absorbing but which yields no reward under any action. The different MDPs thus differ only in their transition probabilities in the choice states s_i.

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

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

where p and α are given by

$$\begin{aligned} p &= \begin{cases} 1 - e^{-\beta \frac{1}{1-\gamma}} & \beta > 0, \\ e^{-|\beta| \frac{1}{1-\gamma}} & \beta < 0, \end{cases} \\ \alpha &= \frac{5|\beta|}{\gamma} \frac{\varepsilon}{e^{|\beta| \frac{1}{1-\gamma}} - 1}\,, \end{aligned}$$

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 $\{s_i, s^G, s^B\}$. 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 $V^*(s_i) - V^0(s_i)$ where V^0 is the value by choosing $a = 0$. We now show that any non-optimal action is ε-bad on s_i.

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

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

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

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

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

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

Now having 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 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 $B = \{\pi^{\mathcal{U}}(s_i) = a_0\}$ 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

$$\mathcal{E} = \left\{\tilde{p}t - \tilde{k} \leq \sqrt{2p(1-p)\log(\frac{8}{2\theta})}\right\}. \tag{15}$$

From the Chernoff-Hoeffding bound and as shown in [30], we have that $\mathbb{P}_0(\mathcal{E}) > \frac{3}{4}$, and thus, $\mathbb{P}_0(B \cap \mathcal{E}) > \frac{1}{2}$. From Theorem 6, we get that

$$\mathbb{P}_1(B) \geq \mathbb{P}_1(B \cap \mathcal{E}) = \mathbb{E}_1[\mathbb{1}_B \mathbb{1}_{\mathcal{E}}] \geq \mathbb{E}_0\left[\frac{L_1(W)}{L_0(W)}\mathbb{1}_{\mathcal{E}}\mathbb{1}_B\right] \geq \mathbb{E}_0\left[\frac{\theta}{4}\mathbb{1}_{\mathcal{E}}\mathbb{1}_B\right] = \frac{\theta}{4}\mathbb{P}_0(\mathcal{E} \cap B) \geq \frac{\theta}{8}\,. \tag{16}$$

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

$$t < \widetilde{T}(\varepsilon,\delta) := \frac{1}{800}\log(\frac{1}{8\delta})\frac{\gamma^2}{\varepsilon^2} \cdot \frac{e^{|\beta|\frac{1}{1-\gamma}} - 3}{|\beta|^2} \tag{17}$$

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 $H_l^i, l = 1, 2, \ldots, A$, it follows that the algorithm needs at least $\widetilde{T}(\varepsilon,\delta) := A\widetilde{T}(\varepsilon,\delta)$ samples to be (ε,δ)-correct on the state s_i.

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

If the number of total transition samples is less than $\frac{S}{2}\widetilde{T}(\varepsilon,\delta)$, then there must be at least $\frac{S}{2}$ states in the set $\{s_i\}_{i=1}^{S}$ for which some action (apart from a_0) has been tried no more than $\widetilde{T}(\varepsilon,\delta)$ times. Without loss of generality, we might assume that these are the states $\{s_i\}_{i=1}^{S/2}$ and that it is action a_1 that has been tried out at most $\widetilde{T}(\varepsilon,\delta)$ times in each of these states.

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

Let us define the events $\Lambda_i := \{|V^*_{M_1}(s_i) - V^{\pi^{\mathcal{U}}_{T_i}}(s_i)| > \varepsilon\}$ for $i = 1, \ldots, S$. Then, 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

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

where the first line follows from the law of total probability, and the second line from independence. We now have directly that

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

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

$$\begin{aligned}
\mathbb{P}(\|V^* - V^{\pi^{\mathcal{U}}_T}\| > \varepsilon) &\geq \mathbb{P}\Big(\bigcup_{1\leq i\leq S/2} \Lambda(z)\Big) \\
&= 1 - \mathbb{P}\Big(\bigcap_{1\leq i\leq S/2} \Lambda_i^c\Big) \\
&\geq 1 - \mathbb{P}\Big(\{\Lambda_i^c\}_{1\leq i\leq S/2}\,\Big|\,\{T_{z_i} \leq \widetilde{T}(\varepsilon,\delta)\}_{1\leq i\leq S/2}\Big) \\
&\geq 1 - (1-\delta)^{S/2} \\
&\geq \frac{\delta S}{4},
\end{aligned}$$

when $\frac{\delta S}{2} \leq 1$ by Lemma 8. By setting $\delta' = \delta\frac{S}{4}$ and substituting back S' and A', we obtain the result. This shows that if the number of samples is smaller than

$$T = \frac{(S'-2)(A'-1)}{1600}\log(\frac{S'-2}{32\delta})\frac{\gamma^2}{\varepsilon^2} \cdot \frac{e^{|\beta|\frac{1}{1-\gamma}} - 3}{|\beta|^2}$$

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