On the Sample Complexity of Discounted Reinforcement Learning with Optimized Certainty Equivalents

Oliver Mortensen, Mohammad Sadegh Talebi

Keywords: Risk-averse Reinforcement Learning, Sample Complexity, Optimized Certainty Equivalent, Generative Model

Summary

We study risk-sensitive reinforcement learning in tabular discounted Markov Decision Processes (MDPs), assuming access to a generative model of the MDP. We consider a family of risk measures called the optimized certainty equivalents (OCEs), which includes important risk measures such as entropic risk, CVaR, and the mean-variance criterion. 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 OCEs. We provide an exact characterization of the utility functions u for which the corresponding recursive OCE objective is PAC-learnable. We establish that whenever u does not have full domain, i.e., $\mathrm{dom}(u) \neq \mathbb{R}$, the corresponding problem is not PAC-learnable. We present and analyze a simple model-based algorithm and derive PAC sample complexity bounds for both value and policy learning. Finally, we establish complementary lower bounds, demonstrating tightness with respect to the size SA of the state-action space. For a more restricted class of utilities, we derive lower bounds that make the dependence on the effective horizon, $1/(1-\gamma)$, explicit. In particular, for CVaR_τ, our lower bound scales as τ^{-2}, improving the dependence on τ by a factor of τ^{-1} over the best existing lower bound, although its dependence on $1/(1-\gamma)$ remains suboptimal.

Contribution(s)

1. We derive PAC bounds on the sample complexities of both value learning and policy learning when the iterative OCE is given with respect to a utility function with full domain
   Context: In RL with iterative risk-sensitive criterion so far sample complexity has only been studied for CVaR (Deng et al., 2025) and entropic risk (Mortensen & Talebi, 2026) respectively. The optimized certainty equivalents allows for a unified way of treating a large class of risk measures.
2. We classify which utility functions for which the iterative OCE criterion is learnable by demonstrating that when u does not have full domain there is a class of MDPs for which no algorithm can be (ε, δ)-correct on all its members.
   Context: While PAC-bounds have been obtained for CVaR and entropic risk, we know of no works that have discussed learnability, that is wether it is possible at all to obtain PAC-bounds for different risk criteria.
3. We provide lower bounds on the sample complexities that show our upper bounds are tight in $S, A, \frac{1}{\delta}$, and $\frac{1}{\varepsilon}$. Under more assumptions on u, we provide lower bounds for policy learning under CVaR_τ that enjoys a tighter dependence on τ than the one in Deng et al. (2025).
   Context: Deng et al. (2025) provide lower bounds on the sample complexity of policy learning with iterated CVaR and Mortensen & Talebi (2026) provide lower bounds for both policy and value sample complexity with iterated entropic risk.

On the Sample Complexity of Discounted Reinforcement Learning with Optimized Certainty Equivalents

Oliver Mortensen[1], **Mohammad Sadegh Talebi**[1]
olmo@di.ku.dk, sadegh.talebi@di.ku.dk

[1]**Department of Computer Science, University of Copenhagen, Denmark**

Abstract

We study risk-sensitive reinforcement learning in tabular discounted Markov Decision Processes (MDPs), assuming access to a generative model of the MDP. We consider a family of risk measures called the optimized certainty equivalents (OCEs), which includes important risk measures such as entropic risk, CVaR, and the mean-variance criterion. 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 OCEs. We provide an exact characterization of the utility functions u for which the corresponding recursive OCE objective is PAC-learnable. We establish that whenever u does not have full domain, i.e., $\mathrm{dom}(u) \neq \mathbb{R}$, the corresponding problem is not PAC-learnable. We present and analyze a simple model-based algorithm and derive PAC sample complexity bounds for both value and policy learning. Finally, we establish complementary lower bounds, demonstrating tightness with respect to the size SA of the state-action space. For a more restricted class of utilities, we derive lower bounds that make the dependence on the effective horizon, $1/(1-\gamma)$, explicit. In particular, for CVaR_τ, our lower bound scales as τ^{-2}, improving the dependence on τ by a factor of τ^{-1} over the best existing lower bound, although its dependence on $1/(1-\gamma)$ remains suboptimal.

1 Introduction

In reinforcement learning (RL), the standard objective is to maximize the expected return, defined as the (possibly discounted) sum of rewards (Sutton & Barto, 1998). However, because this objective is inherently *risk-neutral*, it may be inadequate for many high-stakes application domains, such as operations research (Delage & Mannor, 2010), healthcare (Ernst et al., 2006), and transportation (Kamran et al., 2020). Such applications must account for the variability of returns and the associated risks. One principled approach to addressing this limitation is to optimize a risk measure of the return distribution. Notable risk measures considered in the literature include mean-variance (Li & Ng, 2000), value-at-risk (VaR) (Dempster, 2002), Conditional VaR (CVaR) (Shapiro et al., 2021), entropic risk measure (ERM) (Howard & Matheson, 1972), and entropic VaR (EVaR) (Ahmadi-Javid, 2012), all of which have been applied in numerous application domains. Among these, CVaR and entropic risk have received great attention from the community for modeling risk-sensitivity in MDPs (Chow & Ghavamzadeh, 2014; Bisi et al., 2022; Bäuerle & Ott, 2011; Howard & Matheson, 1972; Borkar & Meyn, 2002; Hau et al., 2023b; Fei et al., 2020). However, much of the existing literature studies undiscounted problems, despite the relevance of discounted MDPs; see, e.g., Bäuerle & Rieder (2014); Hau et al. (2023b); Mortensen & Talebi (2026) for notable exceptions.

In this paper, we study risk-sensitive RL in discounted MDPs, assuming access to a generative model of the MDP, which is a simulator that generates samples from the true MDP for arbitrary state-action pairs. We consider a broad and important class of risk measures that includes those that

can be expressed as an Optimized Certainty Equivalent (OCE) (Ben-Tal & Teboulle, 2007). Formal definitions are deferred to Section 2, while Appendix A briefly reviews the relevant notions from the theory of risk measures. The class of OCE risk measures includes many important examples, such as CVaR, entropic risk, and the mean-variance criterion, each induced by a suitable choice of utility function.

We stress that in risk-sensitive RL, objectives can be broadly formulated in two ways: In the *non-recursive* (also called non-iterated or static) formulation, the risk measure is directly applied to the total return (e.g., $\sum_{t=0}^{\infty} \gamma^t r_t$ in the discounted case) (Borkar, 2002; Borkar & Meyn, 2002; Hau et al., 2023a). In contrast, in the *recursive* (also called iterated, nested, or dynamic) formulation, the risk measure is applied at every step t to the reward-to-go (Asienkiewicz & Jaśkiewicz, 2017; Bäuerle & Glauner, 2022; Deng et al., 2025; Mortensen & Talebi, 2026). The non-recursive approach may 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 approach may lead to a more cautious behavior by controlling risk at every step, which can be either desirable or overly conservative depending on the application (Deng et al., 2025; Xu et al., 2023; Wang et al., 2025). These formulations are therefore regarded as orthogonal modeling choices. From a technical standpoint, a key distinction is that non-recursive formulations do not generally admit Bellman-type optimality equations and may result in time-inconsistent optimal policies (see Jaquette (1976)), whereas recursive formulations preserve Bellman-type optimality structures. Motivated by these considerations, we study risk-sensitive discounted RL with objectives defined via the recursive OCE.

1.1 Main Contributions

We consider risk-sensitive RL in tabular discounted MDPs under recursive OCE in the generative setting. 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 (the *policy learning* problem) or an ε-close approximation of the optimal Q-value in the max-norm (the *value learning* problem), with probability exceeding $1 - \delta$.

We make the following contributions. We propose a model-based algorithm, called Model-Based OCE Value Iteration (MB-OCE-VI), and establish PAC-type bounds on its sample complexity for both value learning and policy learning (Theorem 1), under the assumption that the utility function u defining the OCE belongs to $\mathbb{U}_1$, which is essentially the set of utilities with a full domain —for a more precise definition and context, see Subsection 2.1. These bounds have an optimal dependence (up to log-factors) on the size of state-action space, SA, and hold simultaneously for all OCE objectives defined using utilities in $\mathbb{U}_1$.

We further show that the restriction to OCEs associated to utilities $u \in \mathbb{U}_1$ is necessary. We prove this claim by establishing impossibility results on the PAC sample complexity bound when OCE is defined by a utility function $u \notin \mathbb{U}_1$. We thus provide an exact characterization of OCE measures, which are PAC-learnable under value and policy learning.

Finally, we establish worst-case lower bounds on the sample complexity under OCEs for both value and policy learning problems. For each problem, we present two lower bounds. The first one (Theorems 4 and 6) is a general lower bound that holds for any OCE defined by a utility $u \in \mathbb{U}_1$. This general lower bound exhibits an optimal dependence (up to log-factors) on $S, A, \frac{1}{\delta}$, and ε, but a complicated dependence on the effective horizon, $1/(1-\gamma)$. The second lower bound (Theorems 5 and 7) makes the dependence on the effective horizon explicit, but holds for a sub-class of utilities in $\mathbb{U}_1$. Nevertheless, we show that this sub-class includes all strongly risk-averse coherent OCE risk measures that notably includes CVaR. For CVaR_τ, we show that the latter lower bound improves upon the best existing lower bound when τ is sufficiently small relative to $1/(1-\gamma)$. To the best of our knowledge, these results constitute the first upper and lower bounds on the sample complexity of recursive OCE in discounted MDPs, and are the first impossibility results for RL under OCEs.

1.2 Related Work

Risk-neutral discounted RL. There is a rich, and yet growing, literature on provably efficient RL in tabular discounted MDPs. In the generative setting, which we also consider, early work includes (Kakade, 2003; Kearns & Singh, 1998), followed by a line of work, notably including (Gheshlaghi Azar et al., 2013; Sidford et al., 2018; Wang, 2020; Li et al., 2024; Jin et al., 2024). Gheshlaghi Azar et al. (2013) provide the first minimax-optimal sample complexity bounds of $\widetilde{\mathcal{O}}\big(\frac{SA}{\varepsilon^2(1-\gamma)^3}\big)$ for both value and policy learning, accompanied by a lower bound of $\widetilde{\Omega}\big(\frac{SA}{\varepsilon^2(1-\gamma)^3}\big)$ for value learning. Model-free methods are analyzed in a more recent line of work such as (Sidford et al., 2018; Wang, 2020; Jin et al., 2024). Recently, Li et al. (2024) derive an optimal bound valid for the entire ε-range, using model-based methods constructed from the empirical MDP but with reward perturbations or conservative planning. Existing optimal sample-complexity analyses rely on techniques that crucially exploit the additive structure of the return. As a result, these techniques generally do not extend to risk-sensitive objectives.

Risk-sensitive RL. A substantial literature studies decision making under risk measures in bandit and RL settings. For bandits, we refer to (Maillard, 2013; Khajonchotpanya et al., 2021). For RL, representative examples include work on CVaR (Deng et al., 2025; Du et al., 2023; Chen et al., 2024; Lam et al., 2022), entropic risk (Borkar & Meyn, 2002; Moharrami et al., 2025; Marthe et al., 2023; 2025; Hau et al., 2023b), and mean-variance risk (Sood et al., 2023; Huang et al., 2022; La & Ghavamzadeh, 2013). In particular, under recursive CVaR, Deng et al. (2025) analyze sample complexity in the generative setting and provide a lower bound. Under recursive ERM, most of the literature studies online episodic RL in the regret setting (e.g., Fei et al. (2020; 2021a;b); Hu & Leung (2023); Liang & Luo (2024)). For the discounted setting, Mortensen & Talebi (2026) provide PAC upper and lower bounds in the generative setting.

Another line of work develops algorithms for entire families of risk measures. Two notable families studied in this context are coherent risk measures and OCEs. For coherent risk measures, we refer to Petrik & Subramanian (2012); Tamar et al. (2015); Lam et al. (2022); Zhao et al. (2024). For OCEs, notable existing results include Wang et al. (2025); Xu et al. (2023); Rigter et al. (2023); Lee et al. (2025). These works either do not provide provably sample-efficient learning guarantees or do not study discounted MDPs. Among the closest works, Rigter et al. (2023) consider offline RL in discounted MDPs under recursive OCE but do not provide sample-complexity guarantees.

2 Setting: RL under Recursive OCE

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 $\|\cdot\| := \|\cdot\|_\infty$. We let $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ denote the space of essentially bounded random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

2.1 Optimized Certainty Equivalents

For risk-averse agents it is natural to be able to rank different random variables based on risk measures. Ben-Tal & Teboulle (2007) propose the optimised certainty equivalent of a class of utility functions $\mathbb{U}_0$, which we define shortly. This approach not only provides a direct link between microeconomic foundations and risk but also form a unifying framework as risk measures derived as OCEs are guaranteed to be convex and include several well-known risk measures such as entropic risk, the mean-variance criterion, and conditional value-at-risk.

Let $u : \mathbb{R} \to [-\infty, \infty)$ be a closed, proper concave, non-decreasing function satisfying that $u(0) = 0$ and $1 \in \partial u(0)$, where ∂u denotes the superdifferential of u. Further, define $\mathrm{dom}(u) := \{t \in \mathbb{R} | u(t) > -\infty\}$. The collection of all such functions are denoted by $\mathbb{U}_0$. A subset of great interest in this paper is the collection of finite utility functions, i.e., those where $\mathrm{dom}(u) = \mathbb{R}$, which we

denote by $\mathbb{U}_1$; that is, $\mathbb{U}_1 = \{u \in \mathbb{U}_0 | \operatorname{dom}(u) = \mathbb{R}\}$. We further define the subclass of strongly risk-averse utility functions $\mathbb{U}_1^< := \{u \in \mathbb{U}_1 | \forall t \neq 0 : u(t) < t\}$.

The optimized certainty equivalent (OCE) of u is the map $\mathrm{OCE}^u : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ defined as

$$\mathrm{OCE}^u(X) = \sup_{\eta \in \mathbb{R}} (\eta + \mathbb{E}[u(X - \eta)]) .$$

The notion of OCE was first introduced by Ben-Tal & Teboulle (2007), where they also show that the negative of the OCE is a convex risk measure and that $\mathrm{OCE}^u(X) \leq \mathbb{E}[X]$. For brevity, we will often use the notation $\rho := \mathrm{OCE}^u$ to refer to the OCE associated to the utility function u. The OCE ρ admits the following properties: (i) $\rho(0) = 0$ (normalization), (ii) $X \leq Y$ implies $\rho(X) \leq \rho(Y)$ (monotonicity), and (iii) $\rho(c) = c$ for $c \in \mathbb{R}$ (consistency).

Let $\mathcal{S}$ be a finite set with size $S := |\mathcal{S}|$, and X be a random variable with support $\{v(s)\}_{s \in \mathcal{S}}$ with probabilities given by $\mathbb{P}(X = v(s)) = p(s)$. We introduce the shorthand $\rho_p(v(s)) := \rho(X)$.

2.2 Discounted Markov Decision Processes and Recursive OCE Objectives

A finite, discounted infinite-horizon Markov decision process (MDP) is a 5-tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, 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 deterministic reward function, and $\gamma \in (0, 1)$ is the discount factor. We write $P(s'|s, a)$ to denote the transition probability to state $s' \in \mathcal{S}$ under pair (s, a), and often use $P_{s,a}(s')$ as a corresponding shorthand notation. Finally, a stationary deterministic policy is a map $\pi : \mathcal{S} \to \mathcal{A}$.

The agents interaction with the MDP M is as follows. At initialization 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 $a_t \in \mathcal{A}$. 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 ρ. In the classical setting, the value of a policy π is defined as discounted sum of rewards collected under π. However, the classical objective fails to capture the inherent risk coming from the stochastic transitions and thus the uncertainty about the rewards collected on a trajectory. Under the recursive OCE criterion, the agent's objective is defined using the iteration of OCEs (Bäuerle & Jaśkiewicz, 2024). The value function V^π and the state-action value function (or Q-value) Q^π of a policy π are defined informally as

$$\begin{aligned} V^\pi(s_0) &= r_0 + \gamma \rho_{s_0, \pi(s_0)} \Big(r_1 + \gamma \rho_{s_1, \pi(s_1)} \big(r_2 + \gamma \rho_{s_2, \pi(s_2)}(\ldots) \big) \Big) , \\ Q^\pi(s_0, a) &= R(s_0, a) + \gamma \rho_{s_0, a} \Big(r_1 + \gamma \rho_{s_1, \pi(s_1)} \Big(r_2 + \gamma \rho_{s_2, \pi(s_2)}(\ldots) \Big) \Big) , \quad \forall a \in \mathcal{A}, \end{aligned}$$

where we used the convention that $\rho_{s,a} := \rho_{P_{s,a}}$ and $r_t := R(s_t, \pi(s_t))$. For all (s, a), let $V^*(s) := \sup_\pi V^\pi(s)$ and $Q^*(s, a) := \sup_\pi Q^\pi(s, a)$ denote the optimal value (at state s) and Q-value (at state-action (s, a)), respectively, where sup is taken over the set of all possible policies. Any policy π^* satisfying $V^{\pi^*}(s) = V^*(s)$ for all $s \in \mathcal{S}$ is called optimal, and any policy π obeying $V^\pi(s) \geq V^*(s) - \varepsilon$ for a given $\varepsilon > 0$ is called ε-optimal. As established in Bäuerle & Jaśkiewicz (2024), there exists a stationary deterministic optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$ that achieves $V^*(s)$ for all states s simultaneously, which is shown to satisfy the optimal Bellman equations:

$$\begin{aligned} V^*(s) &= \max_{a \in \mathcal{A}} \big(R(s, a) + \gamma \rho_{s,a}(V^*(s')) \big), && \forall s \in \mathcal{S}. \\ Q^*(s, a) &= R(s, a) + \gamma \rho_{s,a}\big(\max_{a' \in \mathcal{A}} Q^*(s', a') \big), && \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \end{aligned}$$

We introduce the optimal Bellman operator $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined for $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by

$$(\mathcal{T}f)(s,a) := R(s,a) + \gamma\rho_{s,a}\big(\max_{a'\in\mathcal{A}} f(s',a')\big), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

It is evident that Q^* is the unique fixed-point of $\mathcal{T}$: $Q^* = \mathcal{T}Q^*$.

Since $\mathcal{T}$ is a γ-contraction (Lemma 4 in Appendix B), it follows that Q^* can be efficiently approximated arbitrarily well by a value-iteration-type algorithm (see Algorithm 2).

2.3 Learning Performance

Under a given OCE risk measure ρ applied recursively, we consider RL algorithms that aim to find an ε-optimal policy or an ε-optimal value function for input $\varepsilon > 0$. This is done by assuming access to a generative model (or simulator) of the MDP, which 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 $Q_T^{\mathcal{U}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ using T samples, whereas the second outputs a policy $\pi_T^{\mathcal{U}} : \mathcal{S} \to \mathcal{A}$.

We evaluate the quality of an algorithm that outputs a Q-value by $\|Q^* - Q_T^{\mathcal{U}}\|$, and that outputs a policy by $\|V^* - V^{\pi_T^{\mathcal{U}}}\|$. 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) as formalized below, which we borrow from the literature on risk-neutral RL:

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 $\mathbb{P}(\|Q^* - Q^{\mathcal{U}}\| \le \varepsilon) \ge 1-\delta$ 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 $\mathbb{P}(\|V^* - V^{\pi^{\mathcal{U}}}\| \le \varepsilon) \ge 1-\delta$ 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*.

3 Model-based Utility Value Iteration

We now present a simple model-based algorithm, called Model-Based OCE Value Iteration (MB-OCE-VI), for value and policy learning settings with the RL objective defined using recursive OCEs, assuming access to a generative model of the MDP.

We introduce some necessary notations. Let $\widehat{P}$ denote the plug-in estimator of the transition function P, built using N independent samples from each state-action pair of the MDP. More precisely, 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 pair $(s,a) \in Z$. Further, let $\widehat{M} = (\mathcal{S},\mathcal{A},R,\widehat{P},\gamma)$ be the corresponding empirical MDP built using $\widehat{P}$ as described in 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: OCE-VI (for empirical MDP)

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

Initialization: Set $Q(s,a) = \frac{1}{2(1-\gamma)}$ for all (s,a)
for $j = 0,1,\ldots,k-1$ **do**
  for *all* $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
    $Q_{j+1}(s,a) = R(s,a) + \gamma\rho_{s,a}(\max_{a'} Q(s',a'))$
  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

We are now ready to introduce OCE-VI whose pseudocode is provided as Algorithm 2. It is a value iteration type algorithm that extends the classical value iteration for risk-neutral objectives to those defined recursively using the OCE ρ applied to the empirical MDP $\widehat{M}$. OCE-VI extends the ERM-QVI algorithm of Mortensen & Talebi (2026) from entropic risk to general OCEs.

Equipped with these, we introduce the MB-OCE-VI algorithm. For any input $\varepsilon > 0$, it first collects $T = NSA$ samples, which is done by making N calls to the generative model, and then computes $\widehat{P}$. Then, it runs value iteration updates, which returns a policy π_k and a Q-value estimate Q_k. Finally, one can set $k = \log\left(\frac{1}{2\varepsilon(1-\gamma)}\right)/\log(1/\gamma)$ (see Lemma 5 in Appendix B).

4 Sample Complexity Analysis of MB-OCE-VI

In this section, we present PAC bounds on the sample complexity of MB-OCE-VI for both value and policy learning. The bounds hold under the assumption that the OCE is defined by a utility $u \in \mathbb{U}_1$; we refer to Section 2.1 for the definition of $\mathbb{U}_1$. As it turns out, this restriction is necessary since the set $\mathbb{U}_1$ will be shown to be precisely the utilities that guarantee continuity of value-functions between similar MDPs on the same state-action space.

We introduce some necessary notations. For a given utility function u, u'_+ denotes the right derivative of u. Further, $\gamma \in (0,1)$, we define $H_\gamma := \frac{1}{1-\gamma}$.

Theorem 1. *Assume $u \in \mathbb{U}_1$. If the total number T of calls to the generative model satisfies*

$$T \geq 32\frac{\gamma^2 SA[u(-H_\gamma)]^2}{\varepsilon^2(1-\gamma)^2}\log\left(\frac{8\gamma SAu'_+(-H_\gamma)}{\delta\varepsilon(1-\gamma)^2}\right),$$

then it holds that $\mathbb{P}(\|Q^* - Q_k\| \geq \varepsilon) \leq \delta$. Furthermore, if

$$T \geq 128\frac{\gamma^4 SA[u(-H_\gamma)]^2}{\varepsilon^2(1-\gamma)^4}\log\left(\frac{16\gamma^2 SAu'_+(-H_\gamma)}{\delta\varepsilon(1-\gamma)^3}\right),$$

then it holds that $\mathbb{P}(\|V^* - V^{\pi_k}\| > \varepsilon) \leq \delta$.

Before presenting the proof, we compare in Table 1 the policy learning sample complexity upper bounds for several representative risk measures.

4.1 Comparison with Existing Sample Complexity Bounds

Here, we provide a comparison between the sample complexity of policy learning in Theorem 1 to best existing bounds for some concrete cases of OCE in the recursive setting.

CVaR. For CVaR with parameter $\tau \in (0,1)$, Theorem 1 gives a policy learning sample complexity of $\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^6\tau^2}\right)$. For recursive CVaR, the best existing bound is due to Deng et al. (2025), which scales as $\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^4\tau^2}\right)$. This shows that our general analysis leaves a gap of $1/(1-\gamma)^2$. The analysis in Deng et al. (2025) exploits the specific properties of CVaR that cannot be generalized to a generic OCE. We also mention that there is no specialized result for value learning for CVaR, to the best of our knowledge.

Entropic risk. For the entropic risk with parameter $\beta > 0$, the best available sample complexities for value learning and policy learning are reported in Mortensen & Talebi (2026). For value learning, Theorem 1 gives a bound of $\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3\beta}\left(e^{\beta/(1-\gamma)}-1\right)^2\right)$, which is worse by a factor of $1/(1-\gamma)$ compared to the bound in Mortensen & Talebi (2026), which scales as $\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2\beta^2}\left(e^{\beta/(1-\gamma)}-1\right)^2\right)$. For policy learning, the resulting bound from Theorem 1 scales as $\widetilde{\mathcal{O}}\left(\frac{SA}{\varepsilon^2(1-\gamma)^5\beta}\left(e^{\beta/(1-\gamma)}-1\right)^2\right)$, which is again off by a factor of $1/(1-\gamma)$ compared to the corresponding bound in Mortensen

Name	OCE(X)	Utility $u(t)$	Sample Complexity (Policy Learning)
Entropic, $\beta > 0$	$-\frac{1}{\beta}\log(\mathbb{E}[e^{-\beta X}])$	$\frac{1}{\beta}(1-e^{-\beta t})$	$\widetilde{\mathcal{O}}\Big(\frac{SA}{\varepsilon^2(1-\gamma)^5\beta}\big(e^{\beta/(1-\gamma)}-1\big)^2\Big)$
CVaR, $\tau \in (0,1)$	$\text{CVaR}_\tau(X)$	$\big[\frac{t}{\tau}\big]^-$	$\widetilde{\mathcal{O}}\Big(\frac{SA}{\varepsilon^2(1-\gamma)^6\tau^2}\Big)$
Mean-variance	$\mathbb{E}[X] - \frac{1}{2}\text{Var}(X)$	$\begin{cases} t-\frac{1}{2}t^2 & t \le 1 \\ \frac{1}{2} & t > 1 \end{cases}$	$\widetilde{\mathcal{O}}\Big(\frac{SA}{\varepsilon^2(1-\gamma)^8}\Big)$
Essential Infimum	$\text{Essinf}(X)$	$\begin{cases} 0 & t \ge 0 \\ -\infty & t < 0 \end{cases}$	∞

Table 1: Sample complexity bounds implied by Theorem 1 for selected risk measures.

& Talebi (2026). We note that the bounds in Mortensen & Talebi (2026) uses some proof elements that are specifically tailored to the entropic measure, and cannot be applied for generic OCEs.

Remark 1. *We show in Proposition 2 that* $u'_+(-\frac{\gamma}{1-\gamma}) \le |u(-H_\gamma)|$ *for piecewise differentiable* $u \in \mathbb{U}_1$. *Further, since* $|u(-H_\gamma)| \ge H_\gamma$, *the dominating term for the effective horizon* H_γ *in the bound is* $|u(-H_\gamma)|$. *For the entropic risk, this is exponentially larger than* H_γ, *while for CVaR and the mean-variance criterion, it contributes with polynomial factors of* H_γ.

4.2 Proof of Theorem 1

In this section, we prove Theorem 1. First, we present a lemma, proven in Appendix C, that characterizes the smoothness of Q-values under the OCE defined by a utility $u \in \mathbb{U}_1$, when the transition function is perturbed. This result could be of interest, beyond the considered RL setting.

Lemma 1. *Let* $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ *and* $\widetilde{M} = (\mathcal{S}, \mathcal{A}, \widetilde{P}, R, \gamma)$ *be two MDPs that only differ in their transition function. Let* π *be a fixed stationary policy, and let* $u \in \mathbb{U}_1$ *be a utility function. Then,*

$$\|Q^\pi - \widetilde{Q}^\pi\| \le \frac{\gamma}{1-\gamma}\max_{s,a}\sup_{\eta\in[0,H_\gamma]}\left|\sum_{s'\in\mathcal{S}}[P_{s,a}(s') - \widetilde{P}_{s,a}(s')]u(V^\pi(s')-\eta)\right|.$$

Proof of Theorem 1. Let $\varepsilon > 0$. Let Q_k be the Q-function output of OCE-VI after k iterations, and let $\widehat{Q}^*$ denote the optimal Q-value in $\widehat{M}$. Note that $\widehat{Q}^{\pi^*}$ is the Q-value in the empirical MDP of the optimal policy of the true MDP. Using a standard decomposition which uses $\widehat{Q}^* \ge \widehat{Q}^{\pi^*}$, we have for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned} Q_k(s,a) &= Q^*(s,a) + Q_k(s,a) - \widehat{Q}^*(s,a) + \widehat{Q}^*(s,a) - Q^*(s,a) \\ &\ge Q^*(s,a) + Q_k(s,a) - \widehat{Q}^*(s,a) + \widehat{Q}^{\pi^*}(s,a) - Q^*(s,a) \\ &\ge Q^*(s,a) - \|Q_k - \widehat{Q}^*\| - \|\widehat{Q}^{\pi^*} - Q^*\|. \end{aligned} \tag{1}$$

Therefore, to establish ε-value-correctness it suffices to ensure $\|\widehat{Q}^{\pi^*} - Q^*\| \le \varepsilon/2$ and $\|Q_k - \widehat{Q}^*\| \le \varepsilon/2$. By Lemma 5 in the appendix, we can have $\|Q_k - \widehat{Q}^*\| < \varepsilon/2$ by picking $k \ge \log(\frac{1}{(1-\gamma)\varepsilon})/\log(1/\gamma)$.

To control $\|\widehat{Q}^{\pi^*} - Q^*\|$, we apply Lemma 1 with $\widetilde{M} = \widehat{M}$ and $\pi = \pi^*$, which yields

$$\|Q^* - \widehat{Q}^{\pi^*}\| \le \frac{\gamma}{1-\gamma}\max_{s,a}\sup_{\eta\in[0,H_\gamma]}\left|\sum_{s'}[P_{s,a}(s') - \widehat{P}_{s,a}(s')]u(V^*(s')-\eta)\right|.$$

For a fixed (s,a), let η^* be an arbitrary but fixed optimizer of

$$\sup_{\eta\in[0,H_\gamma]}\left|\sum_{s'\in\mathcal{S}}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V^*(s')-\eta)\right|.$$

Since $\eta^\star$ is data-dependent, a direct application of Hoeffding's inequality is not allowed. To handle this, we discretize the interval $[0, H_\gamma]$. Let D denote the corresponding discretized set built using uniform discretization. The following lemma controls the introduced error, which establishes that u is sufficiently regular to get a handle on the number of discretization points needed:

Lemma 2. *Let $u \in \mathbb{U}_1$, and define $\bar{\eta} = \min_{\eta \in D} u'_+(-H_\gamma)|\eta^* - \eta|$. If $|D| \geq \frac{u'_+(-H_\gamma)}{\varepsilon(1-\gamma)}$, then*

$$\max_{s,a} \left| \sum_{s'} (P_{s,a}(s') - \widehat{P}_{s,a}(s')) u(V^*(s') - \eta^*) \right| \leq \max_{s,a} \left| \sum_{s'} (P_{s,a}(s') - \widehat{P}_{s,a}(s')) u(V^*(s') - \bar{\eta}) \right| + \frac{\varepsilon}{2}.$$

Lemma 2 implies that

$$\begin{aligned} &\mathbb{P}\left(\max_{s,a} \sup_{\eta \in [0, H_\gamma]} \sum_{s' \in \mathcal{S}} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] u(V^*(s') - \eta) > \varepsilon \right) \\ &\qquad \leq \mathbb{P}\left(\exists \bar{\eta} \in D : \max_{s,a} \sum_{s' \in \mathcal{S}} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] u(V^*(s') - \bar{\eta}) > \frac{\varepsilon}{2} \right). \end{aligned} \tag{2}$$

An application of Hoeffding's inequality (Lemma 6 in the appendix) and taking a union bound over D and all state-action pairs, it follows that the right-hand side of (2) will be smaller than δ if any state-action pair is sampled $N = \frac{8}{\varepsilon^2}[u(-H_\gamma)]^2 \log\left(\frac{4SAu'_+(-H_\gamma)}{\delta\varepsilon(1-\gamma)}\right)$ times. After adjusting ε appropriately, it follows that if

$$T \geq 32 \frac{\gamma^2 SA[u(-H_\gamma)]^2}{\varepsilon^2(1-\gamma)^2} \log\left(\frac{8\gamma SAu'_+(-H_\gamma)}{\delta\varepsilon(1-\gamma)^2} \right),$$

then $\mathbb{P}(\|\widehat{Q}^{\pi^*} - Q^*\| \geq \frac{\varepsilon}{2}) < \delta$, showing the first part of the theorem.

To prove the second result, we use the following lemma, which is proven in the appendix:

Lemma 3. *Let $\varepsilon > 0$. Let $\overline{V} \in \mathbb{R}^S$ be a value function obeying $\|V^* - \overline{V}\| < \varepsilon$, 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, $\|V^* - V^{\pi_G}\| \leq \frac{2\gamma}{1-\gamma}\varepsilon$.*

Applying Lemma 3 with $\overline{V} = \widehat{V}^{\pi_k}$, 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, the true value of π_k satisfies, with probability at least $1 - \delta$,

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

which after properly adjusting ε yields the announced sample complexity for policy learning. □

5 Impossibility Results

In this section, we present some results for PAC-learnability under OCEs. Specifically, we establish that for essentially all utility functions $u \in \mathbb{U}_0$ for which $\operatorname{dom}(u) \neq \mathbb{R}$, it is impossible to obtain PAC bounds for the corresponding learning problems under OCE^u. This is done by constructing a parametric family of simple MDPs, for which the value functions are not continuous in the parameter. By making this parameter sufficiently small, we can thus have two MDPs with a large gap in value functions, while the number of samples it takes to distinguish them can be made arbitrarily large.

We will have to require that u not coincide with the identity for $u \geq 0$. As formalized in Proposition 1, the only utilities this assumption rules out all lead to the same risk measure, namely the expectation.

Theorem 2 (PAC-learning impossibility, value learning)**.** *Let $u \in \mathbb{U}_0$ be a utility function for which (i) $\mathrm{dom}(u) \neq \mathbb{R}$ and (ii) there exists $t_0 > 0$ such that $u(t_0) < t_0$. Then, there exists a class of MDPs $\mathbb{M}$ with S states, A actions, and discount factor γ such that no value learning algorithm $\mathcal{U}$ can be (ε, δ)-correct on $\mathbb{M}$.*

Theorem 3 (PAC-learning impossibility, policy learning)**.** *Let $u \in \mathbb{U}_0$ be a utility function for which (i) $\mathrm{dom}(u) \neq \mathbb{R}$ and (ii) there exists $t_0 > 0$ such that $u(t_0) < t_0$. Then, there exists a class of MDPs $\mathbb{M}$ with S states, A actions, and discount factor γ such that no policy learning algorithm $\mathcal{U}$ can be (ε, δ)-correct on $\mathbb{M}$.*

Both theorems are proven in Appendix D. The following result, proven in Appendix F, shows that the assumption in Theorems 2–3 only excludes the utilities that correspond to $\mathrm{OCE}^u \equiv \mathbb{E}$.

Proposition 1. *Let $u \in \mathbb{U}_0$ be any utility function for which $u(t) = t$ for all $t \geq 0$. Then, $\mathrm{OCE}^u(X) = \mathbb{E}[X]$.*

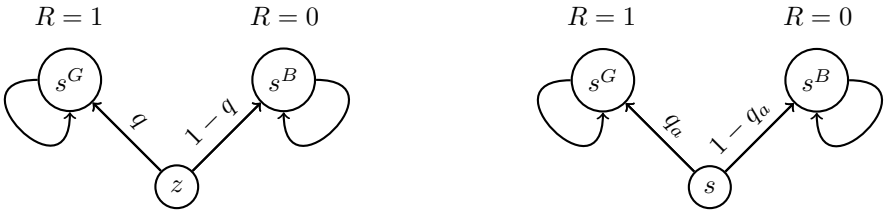
In summary, the upper bounds and the above impossibility results now classify exactly the class of utility functions for which the worst-case sample complexities are finite. These are precisely the utility functions in $\mathbb{U}_1$, i.e., the ones with full domain $\mathrm{dom}(u) = \mathbb{R}$, apart from the special utilities u for which $\mathrm{OCE}^u = \mathbb{E}$, which coincide with the risk-neutral setting.

Remark 2. *Inspecting the class of MDPs used in our construction, we note that it is straightforward to extend the same impossibility results to the case of RL problems with objectives defined using non-recursive OCEs (see Chapter 5 in Bäuerle & Jaśkiewicz (2024)). This is due to the fact that the construction we considered allows for direct computation of the value functions in this case.*

6 Lower Bounds

In this section, we provide sample complexity lower bounds for both policy and value learning. For each problem, we present two lower bounds. The first one (Theorems 4 and 6) is a general lower bound valid for all utilities $u \in \mathbb{U}_1$. This lower bound has an optimal dependence (up to log-factors) on $S, A, \frac{1}{\delta}, \varepsilon$, but a complicated dependence on u and H_γ. The second lower bound (Theorems 5 and 7) enjoys a more interpretable dependence on H_γ, but holds for a sub-class of utilities in $\mathbb{U}_1$. The sub-class includes all strongly risk-averse coherent OCE risk measures including CVaR.

The MDP constructions used in the proofs are similar to Mortensen & Talebi (2026), where only the entropic risk is considered. The main challenge with general OCE is to get tight lower bounds on the difference in value functions for the MDP building blocks (shown in Figure 1) with similar transitions.



(a) Hard-to-learn MDP construction building block for value learning

(b) Hard-to-learn MDP construction building block for policy learning

Figure 1: Lower bound building blocks for value learning (a) and policy learning (b)

The idea behind the proofs is to first obtain a tight lower bound on the difference in value functions for two instances with almost similar transition probabilities to ensure only the optimal policy is ε-good for policy learning and that an output $Q^{\mathcal{U}}$ cannot be ε-good on both instances for value learning. Then one has to lower bound the number of samples needed for the learner to correctly identify which instance the samples are from, which is harder the more similar the instances are.

Finally, one has to combine the building blocks to get the dependence on S and A and then perform some parameter tuning to make the learning as hard as possible.

6.1 Value Learning Lower Bounds

Theorem 4. *Let $u \in \mathbb{U}_1$ be a utility function. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$, constants $c_1, c_2 > 0$, and a strictly positive function $\Phi : \mathbb{U}_1 \times [1, \infty) \to (0, \infty)$ such that for any RL algorithm $\mathcal{U}$ that outputs a Q-value $Q^{\mathcal{U}}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in \big(0, \bar{\varepsilon}\big)$, it holds: if the total number T of samples satisfies*

$$T \leq \frac{SA\Phi(u, H_\gamma)}{c_1 \varepsilon^2} \log\Big(\frac{SA}{c_2 \delta}\Big),$$

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 5. *Let $u \in \mathbb{U}_1$ be a utility function for which $u'_+(0) < 1 < u'_-(0)$. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$ and 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 any $\varepsilon \in \big(0, \bar{\varepsilon}\big)$, the following holds: if the total number T of samples satisfies*

$$T \leq \frac{SA}{c_1 \varepsilon^2} \log\Big(\frac{SA}{c_2 \delta}\Big) |u(-H_\gamma)|^2 \left(1 - \left[\frac{1 - u'_+(0)}{u'_+(-H_\gamma) - u'_+(0)}\right]^2\right),$$

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

6.2 Policy Learning Lower Bounds

Theorem 6. *Let $u \in \mathbb{U}_1$ be a utility function. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$, a strictly positive function $\Phi : \mathbb{U}_1 \times [1, \infty) \to (0, \infty)$, and constants $c_1, c_2 > 0$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $\pi_{\mathcal{U}}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in (0, \bar{\varepsilon})$, we have: if the total number T of samples satisfies*

$$T \leq \frac{SA\Phi(u, H_\gamma)}{c_1 \varepsilon^2} \log\Big(\frac{S}{c_2 \delta}\Big),$$

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

Theorem 7. *Let $u \in \mathbb{U}_1$ be a utility function for which $u'_+(0) < 1 < u'_-(0)$. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$ and constants $c_1, c_2 > 0$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $\pi_{\mathcal{U}}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in \big(0, \bar{\varepsilon}\big)$, the following holds: if the total number T of samples satisfies*

$$T \leq \frac{SA}{c_1 \varepsilon^2} \log\Big(\frac{S}{c_2 \delta}\Big) |u(-H_\gamma)|^2 \left(1 - \left[\frac{1 - u'_+(0)}{u'_+(-H_\gamma) - u'_+(0)}\right]^2\right),$$

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

6.3 Discussion of Lower Bounds

The presented lower bounds establish the first sample complexity lower bounds for the family of OCEs, to our best knowledge. For specific OCE measures, there are two relevant lower bounds in the literature: the one in Deng et al. (2025) derived for CVaR (policy learning), and the one in Mortensen & Talebi (2026) for the entropic risk (value and policy learning). The lower bound in Mortensen & Talebi (2026) has an exponential dependence on H_γ. The dependence on H_γ in our lower bound is not straightforward, which makes the comparison difficult. However, it could be that our bounds are not as sharp as those in Mortensen & Talebi (2026), but they hold for a much larger set of problems. A comparison with the bound in Deng et al. (2025) is provided later.

Theorems 4 and 6 establish that our upper bounds from Theorem 1 are tight in $S, A, \frac{1}{\varepsilon}$, and $\frac{1}{\delta}$, but with a dependence on H_γ that is not interpretable.

While the assumption that $u'_+(0) < 1 < u'_-(0)$ excludes entropic risk and the mean-variance criterion, it is inclusive enough to include all finite strongly risk-averse coherent risk measures, which by Theorem 3.1 of Ben-Tal & Teboulle (2007) are exactly the ones for which $u = \begin{cases} \lambda_1 t & t \geq 0 \\ \lambda_2 t & t < 0 \end{cases}$ for some λ_1 and λ_2 with $0 \leq \lambda_1 < 1 < \lambda_2$. Furthermore, for any fixed finite strongly risk-averse coherent risk measure, we obtain a lower bound scaling as $\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2}\right)$. The risk measure CVaR_τ is obtained by taking $\lambda_1 = 0$ and $\lambda_2 = \frac{1}{\tau}$, yielding the following lower bound:

Corollary 1. *For CVaR_τ, there exist constants $c_1, c_2 > 0$ and $\bar{\varepsilon}$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $\pi_\mathcal{U}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in (0, \bar{\varepsilon})$, if the total number T of samples satisfies*

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

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

Corollary 1 implies a sample complexity lower bound of $\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2(1-\gamma)^2\tau^2}\right)$ for policy learning under CVaR_τ. This bound can be compared with the lower bound in Deng et al. (2025), scaling as $\widetilde{\Omega}\left(\frac{(1-\gamma\tau)SA}{\varepsilon^2(1-\gamma)^4\tau}\right)$, which to our knowledge constitutes the best existing lower bound for policy learning under CVaR. While the two bounds have the same dependence on SA and $\log(1/\delta)$, they differ in terms of dependence on τ and the effective horizon H_γ. As $\tau \leq 1$, our lower bound has a stronger dependence on τ, while that of Deng et al. (2025) has a better dependence on the effective horizon. So neither bound dominates the other uniformly. The comparison depends on the relative scaling of τ and H_γ. If τ is sufficiently small (e.g., relative to H_γ^2), the τ-dependence dominates and our lower bound has better overall scaling. But for large horizon relative to τ, the lower bound of Deng et al. (2025) becomes larger.

7 Concluding Remarks

We studied the PAC-learnability and sample complexity of recursive OCE reinforcement learning in finite discounted MDPs with a generative model. We introduced a simple model-based algorithm MB-OCE-VI and derived PAC bounds on the sample complexity when the utility function defining the OCE has full domain. We then show that the assumption of full domain is needed to obtain PAC bounds by proving that it is impossible to obtain worst-case PAC guarantees when the domain is not full. We therefore obtain a complete characterization of the utility functions for which PAC learning is possible. We also derive lower bounds for all utilities with full domain that demonstrate the tightness of our upper bounds in $S, A, \frac{1}{\varepsilon}$, and $\frac{1}{\delta}$, but with a complicated dependence on the effective horizon, $1/(1-\gamma)$. Finally, for a more restricted class of utilities that includes all strongly risk-averse risk measures, we give tighter lower bounds and show that they can outperform the best existing lower bound for CVaR_τ in the regime where τ is very small.

While we solve the learnability problem, we leave open some gaps in the effective horizon. Closing these gaps remains an interesting direction for future work. It is currently unclear whether the remaining gap is primarily due to the upper bounds or the lower bounds, although we conjecture that both can be improved. Improving the upper bounds, however, seems to require new analytical techniques as the MB-OCE-VI algorithm when $\text{OCE}^u = \mathbb{E}$ is shown to be optimal in the classical setting. Improving the lower bounds might require both new analytical techniques and a new hard-to-learn MDP construction. Finally, we also believe it would be interesting to consider more complex RL settings such as offline RL (Rashidinejad et al., 2021), where data is collected under an unknown but fixed behavior policy.

Acknowledgments

The authors would like to acknowledge the support from Independent Research Fund Denmark under grant number 1026-00397B. Mohammad Sadegh Talebi was partially supported by Innovation Fund Denmark under grant number 1063-00031B.

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012.

Hubert Asienkiewicz and Anna Jaśkiewicz. A note on a new class of recursive utilities in Markov decision processes. *Applicationes Mathematicae*, 44:149–161, 2017.

Nicole Bäuerle and Alexander Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022.

Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: An overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Lorenzo Bisi, Davide Santambrogio, Federico Sandrelli, Andrea Tirinzoni, Brian D Ziebart, and Marcello Restelli. Risk-averse policy optimization via risk-neutral policy optimization. *Artificial Intelligence*, 311:103765, 2022.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2): 294–311, 2002.

Vivek S Borkar and Sean P Meyn. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

Yu Chen, Yihan Du, Pihe Hu, Siwei Wang, Desheng Wu, and Longbo Huang. Provably efficient iterated CVaR reinforcement learning with function approximation and human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. *Advances in neural information processing systems*, 27, 2014.

Erick Delage and Shie Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

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

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.

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.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal STI strategies for HIV: A reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE, 2006.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021a.

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

Hans Föllmer and Alexander Schied. Convex and coherent risk measures. *Encyclopedia of Quantitative Finance*, pp. 355–363, 2010.

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

Jia Lin Hau, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On dynamic programming decompositions of static risk measures in Markov decision processes. *Advances in Neural Information Processing Systems*, 36:51734–51757, 2023a.

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

Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

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*, pp. 5411–5437. PMLR, 2023.

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*, pp. 377–385, 2022.

Stratton C Jaquette. A utility criterion for Markov decision processes. *Management Science*, 23(1): 43–49, 1976.

Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. *Advances in Neural Information Processing Systems*, 37:117481–117508, 2024.

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

Danial Kamran, Carlos Fernandez Lopez, Martin Lauer, and Christoph Stiller. Risk-aware high-level decisions for automated driving at occluded intersections with reinforcement learning. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1205–1212. IEEE, 2020.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in Neural Information Processing Systems*, 11, 1998.

Najakorn Khajonchotpanya, Yilin Xue, and Napat Rujeerapaiboon. A revised approach for risk-averse multi-armed bandits under CVaR criterion. *Operations Research Letters*, 49(4):465–472, 2021.

Prashanth La and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. *Advances in Neural Information Processing Systems*, 26, 2013.

Thanh Lam, Arun Verma, Bryan Kian Hsiang Low, and Patrick Jaillet. Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2022.

Jane H Lee, Baturay Saglam, Spyridon Pougkakiotis, Amin Karbasi, and Dionysis Kalogerias. Risk-averse constrained reinforcement learning with optimized certainty equivalents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Duan Li and Wan-Lung Ng. Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, 10(3):387–406, 2000.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

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*, pp. 1774–1782. PMLR, 2024.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pp. 218–233. Springer Berlin Heidelberg, 2013.

Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in Markov decision processes. *Advances in Neural Information Processing Systems*, 36:56488–56507, 2023.

Alexandre Marthe, Samuel Bounan, Aurélien Garivier, and Claire Vernade. Efficient risk-sensitive planning via entropic risk measures. *arXiv preprint arXiv:2502.20423*, 2025.

Mehrdad Moharrami, Yashaswini Murthy, Arghyadip Roy, and Rayadurgam Srikant. A policy gradient algorithm for the risk-sensitive exponential cost MDP. *Mathematics of Operations Research*, 50(1):431–458, 2025.

Oliver Mortensen and M. Sadegh Talebi. Recursive entropic risk optimization in discounted MDPs: Sample complexity bounds with a generative model. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Marek Petrik and Dharmashankar Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

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

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

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. 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*, pp. 770–787, 2018.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Srijan Sood, Kassiani Papasotiriou, Marius Vaiciulis, and Tucker Balch. Deep reinforcement learning for optimal portfolio allocation: A comparative study with mean-variance optimization. In *Proceedings of the Workshop on Financial Planning and Scheduling (FinPlan)*, pp. 28–35. AAAI Press, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press Cambridge, 1998.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in Neural Information Processing Systems*, 28, 2015.

Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. In *Forty-second International Conference on Machine Learning*, 2025.

Mengdi Wang. Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

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

Yujie Zhao, Jose E Escamill, Weyl Lu, and Huazheng Wang. RA-PbRL: Provably efficient risk-aware preference-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37:60835–60871, 2024.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Risk Measures

In this section, we briefly introduce risk measures. See, e.g., Föllmer & Schied (2010) for a good reference that like us model stochastic outcomes as rewards. We here collect some precise definitions for the reward setting and list some important examples.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a background probability space, and $\mathcal{M}$ a 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 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., Föllmer & Schied (2010)) as it 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

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

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

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

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

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.

Mean-variance criterion The risk measure $\text{MV}(X) := \frac{1}{2}\text{Var}(X) - \mathbb{E}[X]$ is called the mean-variance criterion risk measure and is one way to explicitly account for variance when modeling risk.

Essential infimum. The functional $-\text{Essinf}(X)$ is a risk measure. It can reasonably be said to be the most risk-averse risk measure as it only takes the worst-case outcome of a random variable into consideration and ignores all other distributional aspects of the random variable. It can be obtained as $\lim_{\beta\to\infty} \text{ERM}_\beta(X)$ and $\lim_{\tau\to 0} \text{VaR}_\tau(X)$

B Convergence of OCE-VI

Lemma 4. *The operator $\mathcal{T}$ is a γ-contraction with respect to $\|\cdot\|_\infty$.*

Proof. 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

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

Then, by monotonicity and consistency of ρ, it holds that

$$\begin{aligned}
\|Q' - W'\| &= Q'(s, a) - W'(s, a) \\
&= \gamma\Big(\rho_{s,a}(V(s')) - \rho_{s,a}(X(s'))\Big) \\
&= \gamma\Big(\rho_{s,a}(X(s') + V(s') - X(s')) - \rho_{s,a}(X(s'))\Big) \\
&\leq \gamma\Big(\rho_{s,a}(X(s') + \|V - X\|) - \rho_{s,a}(X(s'))\Big) \\
&= \gamma\|V - X\| \\
&\leq \gamma\|Q - W\|.
\end{aligned}$$

□

From this fact we get the following convergence guarantee on OCE-VI (Algorithm 2):

Lemma 5. *If $k \geq \log\left(\frac{1}{2\varepsilon(1-\gamma)}\right)/\log(1/\gamma)$, then $\|Q^* - Q_k\| \leq \varepsilon$.*

Proof. Since $\mathcal{T}$ is a γ-contraction, we have that

$$\|Q_k - Q^*\| = \|\mathcal{T}Q_{k-1} - \mathcal{T}Q^*\| \leq \gamma\|Q_{k-1} - Q^*\|\,,$$

from which it follows that $\|Q_0 - Q^*\| \leq \gamma^k\|Q_0 - Q^*\|$. Furthermore, by the definition of Q_0, it holds that $\|Q_0 - Q^*\| \leq \frac{1}{2(1-\gamma)}$. Solving $\frac{\gamma^k}{2(1-\gamma)} \leq \varepsilon$ for γ yields the announced result. □

C Missing Lemmas for Upper Bounds

C.1 Proof of Lemma 3

We first prove a result that bounds the value of a greedy policy by the value-function for which the policy is greedy. The result is a generalization of Singh & Yee (1994) and Mortensen & Talebi (2026) to general risk measures. We use the notation $\rho_{s,a}(V(s'))$ as shorthand for ρ applied to the categorical random variable X which takes values in the set $\{V(s')\}_{s'\in\mathcal{S}}$ with probabilities given by $\mathbb{P}(X = V(s')) = P(s'|s, a)$.

Lemma 3. *Let $\varepsilon > 0$. Let $\overline{V} \in \mathbb{R}^S$ be a value function obeying $\|V^* - \overline{V}\| < \varepsilon$, 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, $\|V^* - V^{\pi_G}\| \leq \frac{2\gamma}{1-\gamma}\varepsilon$.*

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 broken arbitrarily. Since π^G is greedy with respect to $\overline{V}$, we have that

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

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

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

Since ρ is monotone and translation invariant, it follows that

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

and by a similar argument

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

yielding the inequality

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

Combining the previous inequalities, we finally see that

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

from which the result follows. □

C.2 Proof of Lemma 1

Lemma 1. *Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and $\widetilde{M} = (\mathcal{S}, \mathcal{A}, \widetilde{P}, R, \gamma)$ be two MDPs that only differ in their transition function. Let π be a fixed stationary policy, and let $u \in \mathbb{U}_1$ be a utility function. Then,*

$$\|Q^\pi - \widetilde{Q}^\pi\| \leq \frac{\gamma}{1-\gamma}\max_{s,a}\sup_{\eta\in[0,H_\gamma]}\left|\sum_{s'\in\mathcal{S}}[P_{s,a}(s') - \widetilde{P}_{s,a}(s')]u(V^\pi(s') - \eta)\right|.$$

In the following proof, we use the following convention. We write $[V|P]$ for the random variable taking values given by the vector V with probability distribution P.

Proof. We suppress π from the notation since it is fixed throughout. Let (s,a) be a state-action pair such that $\|Q - \widetilde{Q}\| = |Q(s,a) - \widetilde{Q}(s,a)|$. We consider two cases.

Case 1: $Q(s,a) \geq \widetilde{Q}(s,a)$**.** From monotonicity and consistency of ρ, it follows that

$$\begin{aligned}\rho([\widetilde{V}|\widetilde{P}_{s,a}]) &= \rho([V + \widetilde{V} - V|\widetilde{P}_{s,a}]) \\ &\geq \rho([V - \|V - \widetilde{V}\||\widetilde{P}_{s,a}]) \\ &= -\|V - \widetilde{V}\| + \rho([V|\widetilde{P}_{s,a}]),\end{aligned}$$

which along with the Bellman recursion implies

$$\begin{aligned}\|Q - \widetilde{Q}\| &= R(s,a) + \gamma\rho([V|P_{s,a}]) - R(s,a) - \gamma\rho([\widetilde{V}|\widetilde{P}_{s,a}]) \\ &\leq \gamma\|V - \widetilde{V}\| + \gamma\big(\rho([V|P_{s,a}]) - \rho([V|\widetilde{P}_{s,a}])\big).\end{aligned}$$

Hence,

$$\|Q - \widetilde{Q}\| \leq \frac{\gamma}{1-\gamma}\big(\rho([V|P_{s,a}]) - \rho([V|\widetilde{P}_{s,a}])\big).$$

Using that the OCE is given by the solution to an optimization problem and using Proposition 2.1 in Ben-Tal & Teboulle (2007) showing that it suffices to optimize over the interval $\eta \in [0, H_\gamma]$, we then have

$$\begin{aligned}\rho([V|P_{s,a}(\cdot)]) - \rho([V|\widetilde{P}_{s,a}]) &= \sup_{\eta\in[0,H_\gamma]}\Big(\eta + \sum_{s'\in\mathcal{S}} P_{s,a}(s')[u(V(s') - \eta)]\Big) \\ &\quad - \sup_{\eta\in[0,H_\gamma]}\Big(\eta + \sum_{s'\in\mathcal{S}} \widetilde{P}_{s,a}(s')[u(V(s') - \eta)]\Big) \\ &\leq \sup_{\eta\in[0,H_\gamma]} \sum_{s'\in\mathcal{S}} [P_{s,a}(s') - \widetilde{P}_{s,a}(s')]u(V(s') - \eta) \\ &\leq \sup_{\eta\in[0,H_\gamma]} \Big|\sum_{s'\in\mathcal{S}} [P_{s,a}(s') - \widetilde{P}_{s,a}(s')]u(V(s') - \eta)\Big|,\end{aligned}$$

thus concluding this case.

Case 2: $Q(s,a) < \widetilde{Q}(s,a)$**.** The proof for this case is similar to that of Case 1, but instead uses the fact that

$$\rho([\widetilde{V}|\widetilde{P}_{s,a}]) = \rho([\widetilde{V} - V + V|\widetilde{P}_{s,a}]) \leq \|V - \widetilde{V}\| + \rho([V|\widetilde{P}_{s,a}]),$$

from which we obtain

$$\begin{aligned}\|Q - \widetilde{Q}\| &= \gamma\big(\rho([\widetilde{V}|\widetilde{P}_{s,a}]) - \rho([V|P_{s,a}])\big) \\ &\leq \gamma\|V - \widetilde{V}\| + \gamma\big(\rho([V|\widetilde{P}_{s,a}]) - \rho([V|P_{s,a}])\big).\end{aligned}$$

The proof of this case follows by noting that

$$\begin{aligned}\rho([V|\widetilde{P}_{s,a}]) - \rho([V|P_{s,a}]) &\leq \sup_{\eta\in[0,H_\gamma]} \sum_{s'\in\mathcal{S}} [\widetilde{P}_{s,a}(s') - P_{s,a}(s')]u(V(s') - \eta) \\ &\leq \sup_{\eta\in[0,H_\gamma]} \Big|\sum_{s'\in\mathcal{S}} [P_{s,a}(s') - \widetilde{P}(s')]u(V(s') - \eta)\Big|.\end{aligned}$$

□

C.3 Concentration via Hoeffding's Inequality

The next result is an application of Hoeffding's inequality to establish a bound on the number of samples needed for the expression inside the supremum in Lemma 1 for a fixed η to concentrate. Let $\widehat{P}_{s,a}(s') = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{X_n = s'\}}$, where X_n is sampled i.i.d. from $\{1, \ldots, S\}$ with probabilities according to $P_{s,a}$.

Lemma 6 (Hoeffding bound)**.** *Fix $\eta \in [0, H_\gamma]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let $V \in \mathbb{R}^S$. If the number N of samples from the state-action pair (s,a) satisfies $N \geq \frac{2}{\varepsilon^2}\big[u(-H_\gamma)\big]^2 \log(2/\delta)$, then with probability exceeding $1-\delta$,*

$$\left| \sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] u(V(s') - \eta) \right| \leq \varepsilon \,.$$

Proof. We first observe that for the random variable $\sum_{s' \in \mathcal{S}} \mathbb{1}_{\{X_n = s'\}} u(V(s') - \eta)$, we have that

$$\begin{aligned} \mathbb{E}\Big[\sum_{s'} \mathbb{1}_{\{X_n = s'\}} u(V(s') - \eta) \Big] &= \sum_{s'} \mathbb{E}\big[\mathbb{1}_{\{X_n = s'\}}\big] u(V(s') - \eta) \\ &= \sum_{s'} P_{s,a}(s') u(V(s') - \eta) \end{aligned}$$

and that it is bounded in $[u(-H_\gamma), u(H_\gamma)] \subset [u(-H_\gamma), -u(-H_\gamma)]$. Also, since

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

we have by Hoeffding's inequality that

$$\mathbb{P}\bigg(\left| \sum_{s'} [P_{s,a}(s') - \widehat{P}_{s,a}(s')] u(V(s') - \eta) \right| \geq \varepsilon \bigg) \leq 2 \exp\bigg(- \frac{2N\varepsilon^2}{\big(-2u(-H_\gamma)\big)^2} \bigg).$$

The lemma follows by observing that this bound is smaller than δ if $N \geq \frac{2}{\varepsilon^2}\big[u(-H_\gamma)\big]^2 \log(2/\delta)$. □

C.4 Proof of Lemma 2

The proof of Lemma 2 follows from the following two lemmas (Lemma 7 and Lemma 8).

Lemma 7. *Let $V \in \mathbb{R}^S$. For $\eta^*, \bar{\eta} \in [0, H_\gamma]$, it holds that*

$$\begin{aligned} &\max_{s,a} \left| \sum_{s'} (P_{s,a}(s') - \widehat{P}_{s,a}(s')) u(V(s') - \eta^*) \right| \\ &\qquad \leq \max_{s,a} \left| \sum_{s'} (P_{s,a}(s') - \widehat{P}_{s,a}(s')) u(V(s') - \bar{\eta}) \right| + u'_+(-H_\gamma) |\eta^* - \bar{\eta}| \,. \end{aligned}$$

Proof. Since u is increasing and concave, we have for $x, y \in [-H_\gamma, H_\gamma]$ where $x \leq y$ that

$$u(y) \leq u(x) + u'_+(x)(y - x) \leq u(x) + u'_+(-H_\gamma)(y - x),$$

and so

$$0 \leq u(y) - u(x) \leq u'_+(-H_\gamma)(y - x).$$

Similarly, if $y \leq x$, it holds that

$$0 \leq u(x) - u(y) \leq u'_+(-H_\gamma)(x - y).$$

Now, consider a pair (s,a). For any $s' \in \mathcal{S}$, plugging in $x = V(s') - \eta^*$ and $y = V(s') - \bar{\eta}$, we have

$$|u(V(s') - \eta^*) - u(V(s') - \bar{\eta})| \leq u'_+(-H_\gamma) |\eta^* - \bar{\eta}| \,.$$

Next we notice that

$$\sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \eta^*) = \sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \bar{\eta}) + \sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))\big[u(V(s') - \eta^*) - u(V(s') - \bar{\eta})\big].$$

Since

$$-u'_+(-H_\gamma)|\eta^* - \bar{\eta}| \leq \sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))\big[u(V(s') - \eta^*) - u(V(s') - \bar{\eta})\big] \leq u'_+(-H_\gamma)|\eta^* - \bar{\eta}|\,,$$

we obtain that

$$\left|\sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \eta^*) - \sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \bar{\eta})\right| \leq u'_+(-H_\gamma)|\eta^* - \bar{\eta}|\,.$$

Finally, by the reverse triangle inequality we have

$$\begin{aligned}
&\left||\sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \eta^*)| - |\sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \bar{\eta})|\right| \\
&\quad\leq \left|\sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \eta^*) - \sum_{s'}(P_{s,a}(s') - \widehat{P}_{s,a}(s'))u(V(s') - \bar{\eta})\right| \\
&\quad\leq u'_+(-H_\gamma)|\eta^* - \bar{\eta}|.
\end{aligned}$$

The lemma follows by taking maximum over (s, a) on both sides. □

Lemma 8. *If the set D of discretization points satisfies $|D| \geq \frac{u'_+(-H_\gamma)}{\varepsilon(1-\gamma)}$, it holds that*

$$\min_{\bar{\eta}\in D} u'_+(-H_\gamma)|\eta^* - \bar{\eta}| < \frac{\varepsilon}{2}\,.$$

Proof. For any interval $[0, L]$ of length L that is discretized uniformly by $|D|$ points, the interval length between any two discretization points is $I = \frac{L}{|D|+1}$. Hence, for any point $x \in [0, L]$, its distance to its nearest discretization point is $\frac{I}{2} = \frac{L}{2(|D|+1)}$. To ensure that the said distance is less than $\xi > 0$, solving for $|D|$ shows that it suffices that $|D| \geq \frac{L}{2\xi}$. Plugging in $L = H_\gamma$ and $\xi = \frac{\varepsilon}{2u'_+(-H_\gamma)}$, we obtain the announced results. □

D Proofs of Impossibility Results

D.1 Proof of Theorem 3

Theorem 2 (PAC-learning impossibility, value learning)**.** *Let $u \in \mathbb{U}_0$ be a utility function for which (i) $\mathrm{dom}(u) \neq \mathbb{R}$ and (ii) there exists $t_0 > 0$ such that $u(t_0) < t_0$. Then, there exists a class of MDPs $\mathbb{M}$ with S states, A actions, and discount factor γ such that no value learning algorithm $\mathcal{U}$ can be (ε, δ)-correct on $\mathbb{M}$.*

Proof. We consider a class $\mathbb{M}$ of MDPs with a single action a and three states s_0, s_G and s_B, where s_G and s_B are absorbing and $R(s_G, a) = 1$ and $R(s_B, a) = R(s_0, a) = 0$. Finally, from s_0 it is possible to transition to s_G with probability p, and to s_B with probability $1 - p$. The MDPs in $\mathbb{M}$ differ only in their value of p. Hence, we can parametrize them by p and write M_p to represent the MDP in which transition probability to s_G is p. Let Q_1 and Q_p denote the Q-values of M_1 and M_p, respectively. We have

$$Q_1(s_0, a) = \frac{\gamma}{1-\gamma}, \qquad Q_p(s_0, a) = \gamma \sup_{\eta\in[0,H_\gamma]} \Big\{\eta + pu(H_\gamma - \eta) + (1-p)u(-\eta)\Big\}\,.$$

By picking γ large enough, we get that $H_\gamma - \xi > t_0$, for some $t_0 > 0$, where $\xi =: -\inf \mathrm{dom}(u)$. Since for any $\eta > \xi$, we have that $\eta + u(H_\gamma - \eta) + (1-p)u(-\eta) = -\infty$, it holds that

$$\sup_{\eta \in [0, H_\gamma]} \Big\{ \eta + pu(H_\gamma - \eta) + (1-p)u(-\eta) \Big\} = \sup_{\eta \in [0, \xi]} \Big\{ \eta + pu(H_\gamma - \eta) + (1-p)u(-\eta) \Big\}.$$

Moreover, since

$$\begin{aligned} \sup_{\eta \in [0, \xi]} \Big\{ \eta + pu(H_\gamma - \eta) + (1-p)u(-\eta) \Big\} &\leq \sup_{\eta \in [0, \xi]} \Big\{ \eta + pu(H_\gamma - \eta) \Big\} \\ &= \xi + pu(H_\gamma - \xi) < H_\gamma, \end{aligned}$$

it follows that $Q_1(s_0, a) - Q_p(s_0, a) \geq \gamma(H_\gamma - \xi) =: \Delta > 0$ for all $p < 1$. Introduce

$$\mathcal{E}_p := \{|Q_p^*(s_0, a) - Q^{\mathcal{U}}(s_0, a)| \leq \varepsilon\}, \qquad \mathcal{E}_1 := \{|Q_1^*(s_0, a) - Q^{\mathcal{U}}(s_0, a)| \leq \varepsilon\}.$$

By picking $\varepsilon < \Delta/2$, the output $Q^{\mathcal{U}}$ of any (ε, δ)-correct algorithm $\mathcal{U}$ satisfies that $\mathcal{E}_1 \cap \mathcal{E}_p = \emptyset$, which implies that $Q^{\mathcal{U}}$ can never be ε-correct on both MDPs simultaneously.

The next argument relies on a change-of-measure between the one induced by the two MDPs. Since under $\mathbb{P}_1$ the only possible sequence of transitions from trying action a in s_0 is to observe a transition to s_G every time, the probability of its occurrence under $\mathbb{P}_p$ is p^N, where N is the number of samples. Assuming $\delta < \frac{1}{2}$, we then have for any (ε, δ)-correct algorithm $\mathcal{U}$ that

$$\mathbb{P}_p(\mathcal{E}_p^c) \geq \mathbb{P}_p(\mathcal{E}_1) = \mathbb{E}_p[\mathbb{1}_{\mathcal{E}_1}] = p^N \mathbb{E}_1[\mathbb{1}_{\mathcal{E}_1}] = p^N \mathbb{P}_1(\mathcal{E}_1) \geq p^N(1-\delta) \geq \frac{1}{2} p^N \,.$$

Solving the equation $\frac{1}{2}p^N \geq \delta$ for N, we find that if $N \leq \frac{\log(1/2\delta)}{\log(1/p)}$, then any algorithm that is (ε, δ)-correct on M_1 cannot also be (ε, δ)-correct on M_p. Since this expression tends to infinity as $p \to 1$, the result follows. □

D.2 Proof of Theorem 3

Theorem 3 (PAC-learning impossibility, policy learning)**.** *Let $u \in \mathbb{U}_0$ be a utility function for which (i) $\mathrm{dom}(u) \neq \mathbb{R}$ and (ii) there exists $t_0 > 0$ such that $u(t_0) < t_0$. Then, there exists a class of MDPs $\mathbb{M}$ with S states, A actions, and discount factor γ such that no policy learning algorithm $\mathcal{U}$ can be (ε, δ)-correct on $\mathbb{M}$.*

Proof. We consider a class of MDPs with 3 states s_0, s^G, s^B and 2 actions a_1 and a_2. We will index the MDPs in $\mathbb{M}$ by $\{1, 2\} \times (0, 1)$ and write M_i^p for $i \in \{1, 2\}$ and $p \in (0, 1)$. The state s^G is absorbing under any action and yields a reward of 1 under any action. The state s^B is absorbing under any action and yields zero reward under any action. The state s_0 yields zero reward under any action and under MDP $M_j^p \in \mathbb{M}$ we have $\mathbb{P}_{j,p}(s^G | s_0, a_j) = p$, $\mathbb{P}_{j,p}(s^B | s_0, a_j) = 1 - p$ and $\mathbb{P}_{j,p}(s^G | s_0, a_{-j}) = 1$ where a_{-j} denotes the other action that is not a_j.

Using the exact same argument as in Theorem 2, by picking γ so large that $H_\gamma > \xi =: -\inf \mathrm{dom}(u)$, we get for any member $M \in \mathbb{M}$ the following lower bound on the value-gap: $V^*(s_0) - V^{a_{-j}}(s_0) > \gamma(H_\gamma - \xi) =: \Delta > 0$ between the optimal action a^* and the other action a_{-j}. Picking $\varepsilon < \Delta$ only the optimal policies are ε-good.

The next part of the argument is a likelihood-ratio type argument showing that, as p approaches 1, the number of samples required to distinguish between the two MDPs becomes arbitrarily large. Let $\mathcal{U}$ be an algorithm that picks an action a based n_1 tries of action a_1 and n_2 tries of action a_2, where the total number of samples is $N = n_1 + n_2$. Let $L_i(m_1, m_2, n_1, n_2)$ denote the probability of observing m_1 successes by trying a_1 for n_1 times, and m_2 successes by trying a_2 for n_2 times under hypothesis H_i. It is clear to see that

$$L_1(m_1, m_2, n_1, n_2) = p^{m_1} \mathbb{1}_{\{m_2 = n_2\}}, \qquad L_2(m_1, m_2, n_1, n_2) = p^{m_2} \mathbb{1}_{\{m_1 = n_1\}}$$

We want to evaluate this likelihood-ratio on the event that all tries turn out to be successes, that is on the event $\mathcal{E} := \{m_1 = n_1 \text{ and } m_2 = n_2\}$, where clearly $\mathbb{P}_{1,p}(\mathcal{E}) = p^{n_1}$. Assuming $\delta < \frac{1}{2}$, we observe that

$$\mathbb{P}_{1,p}(B_2) = \mathbb{E}_1[\mathbb{1}_{B_2}] = \mathbb{E}_2\left[\frac{L_1(n_1, n_2, n_1, n_2)}{L_2(n_1, n_2, n_1, n_2)}\mathbb{1}_{\mathcal{E}}\mathbb{1}_{B_2}\right] \geq p^{n_1}\mathbb{P}_{2,p}(B_2) > \frac{1}{2}p^{n_1} \geq \frac{1}{2}p^N.$$

Solving $\frac{1}{2}p^N = \delta$, we find that if $N \leq \frac{\log(1/2\delta)}{\log(1/p)}$, the algorithm that is (ε, δ)-correct on M_2^p cannot also be (ε, δ)-correct on M_1^p. Finally, by taking the limit $p \to 1$ the result follows. □

E Lower Bounds

In this section, we provide two template constructions for lower bounds for value and policy learning, and then prove the lower bounds.

E.1 Hard-to-Learn MDP Constructions

First, we describe the template constructions for the hard-to-learn MDPs for value learning and policy learning. The constructions and proof techniques are borrowed from (Gheshlaghi Azar et al., 2013) and (Mortensen & Talebi, 2026).

E.1.1 Value Learning

For a lower bound, we construct the following class of MDPs with $S' := S + 2$ states and A actions where the states are labeled $s_1, , \ldots, s_S, s^G, s^B$ and the actions are labeled $a_1, \ldots, a_A$. The states s^G and s^B are absorbing under any actions, and $R(s^G, a) = 1$ and $R(s^B, a) = 0$ for all $a \in A$. For the states $s \in \{s_1, \ldots, s_S\}$, we have $R(s, a) = 0$ for all $a \in \mathcal{A}$. We have SA state-action pair combinations from $\{s_1, \ldots, s_S\} \times \mathcal{A} =: Z$ 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. Let us denote the collection of all such MDPs by $\mathbb{M}$.

In this class of MDPs, we will usually for each $z_i \in Z$ consider two MDPs: M_0 where $q_i = p$, and M_1 where $q_i = p + \alpha$, with $p \in (\frac{1}{2}, 1)$ and $\alpha \in (0, \frac{1-p}{5})$. We denote the corresponding optimal value functions by Q_0^* and Q_1^*.

We let $\mathbb{E}_0$ and $\mathbb{P}_0$ denote, respectively, the expectation and probability operators under H_0, and $\mathbb{E}_1$ and $\mathbb{P}_1$ are similarly defined for H_1.

Theorem 8. *Consider $\gamma \geq \frac{1}{2}$, $\delta < \frac{1}{4}$, and $u \in \mathbb{U}_1$. If there exists $p \in (\frac{1}{2}, 1)$ and $\Delta > 0$ such that for any $a \in (0, \frac{1-p}{5})$ and any $z_i \in Z$, it holds that*

$$Q_1^*(z) - Q_0^*(z) \geq \gamma\alpha\Delta,$$

then there exist $\bar{\varepsilon}(u, H_\gamma)$ and constants $c_1, c_2 > 0$ such that if the total number T of samples is less than

$$T \leq \frac{1}{c_1}\frac{SA\Delta^2 p(1-p)}{\varepsilon^2}\log\left(\frac{SA}{c_2\delta}\right),$$

for any algorithm $\mathcal{U}$ and any $\varepsilon \in (0, \bar{\varepsilon})$, there exists some MDP $M_i \in \mathbb{M}$ where $\mathbb{P}(\|Q_m^* - Q^{\mathcal{U}}\| > \varepsilon) > \delta$.

Proof. By assumption, we can pick $p \in (\frac{1}{2}, 1)$ and $\Delta > 0$ such that for $z \in Z$ it holds that

$$Q_1^*(z) - Q_0^*(z) \geq \gamma\alpha\Delta \geq \frac{\alpha\Delta}{2}.$$

Hence, for any $\varepsilon < \frac{\Delta(1-p)}{20}$, we can pick $\alpha = \frac{4\varepsilon}{\Delta}$ to ensure

$$Q_1^*(z) - Q_0^*(z) \geq 2\varepsilon$$

and so no output $Q^{\mathcal{U}}$ can be ε-close to both Q_1^* and Q_2^* simultaneously and therefore the two sets $B_1 := \{|Q_1^*(z) - Q^{\mathcal{U}}(z)| \leq \varepsilon\}$ and $B_0 := \{|Q_0^*(z) - Q^{\mathcal{U}}(z)| \leq \varepsilon\}$ are disjoint.

Let t be the number of times the algorithm tries z. 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 to s^G in the t trials. Next, we introduce

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

and the event

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

By (Gheshlaghi Azar et al., 2013, Lemma 16), we have $\mathbb{P}_0(\mathcal{E}) > \frac{3}{4}$ and thus $\mathbb{P}_0(B_0 \cap \mathcal{E}) > \frac{1}{2}$. Now by (Mortensen & Talebi, 2026, 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 with $\frac{\theta}{8} > \delta$, we find

$$t < \frac{p(1-p)}{32\alpha^2}\log\Big(\frac{1}{8\delta}\Big) = \frac{\Delta^2 p(1-p)}{512\varepsilon^2}\log\Big(\frac{1}{8\delta}\Big).$$

Furthermore, since $B_0 \subset B_1^c$, we conclude that if the algorithm $\mathcal{U}$ tries the pair z less than

$$\widetilde{T}(\varepsilon,\delta) := \frac{\Delta^2 p(1-p)}{512\varepsilon^2}\log\Big(\frac{1}{8\delta}\Big)$$

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

Let $n := SA$. If the total number of transition samples is less than $\frac{n}{2}\widetilde{T}(\varepsilon,\delta)$ there has to be at least $n/2$ state-action pairs z_i that have been tried at most $\widetilde{T}(\varepsilon,\delta)$ times. We assume that these pairs are $\{z_i\}_{i=1}^{n/2}$ without loss of generality.

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 suffices to only consider algorithms that output an estimate of $Q_{T_i}^{\mathcal{U}}$ based on samples from z_i, since any other samples cannot possibly yield information on $Q^*(z_i)$.

By defining the events $\Lambda_i := \{|Q_{M_1}^*(z_i) - Q_{T_i}^{\mathcal{U}}(z_i)| > \varepsilon\}$, we therefore 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 in the first equality, and the second equality follows from independence. It now follows directly that

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

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

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

By setting $\delta' = \delta\frac{n}{4}$, substituting back S', and using $\frac{S}{2} \leq S-2$ for $S\geq 2$, we have

$$T = \frac{SA\Delta^2 p(1-p)}{2048\varepsilon^2}\log\Big(\frac{SA}{64\delta}\Big)$$

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

E.1.2 Policy Learning

The class of MDPs we consider has $S+2$ states labelled $s_1,\dots,s_S,s^G,s^B$ and $A+1$ actions labelled $a_0,a_1,\dots,a_A$. The state s^G is absorbing and yields a reward of $R=1$ under all actions. The state s^B is also absorbing and yields a reward of $R=0$ under all actions. All other states $s_1,\dots,s_S$ yields zero reward under all actions and can only tansition to either s^G or s^B with probabilities depending on the action and the MDP.

For each state $s_i \in \{s_1,\dots,s_S\}$ we consider the $a+1$ hypotheses H_0^i and H_l^i for $l\neq 0$ defined as

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

where $p\in(\frac{1}{2},1)$ and $\alpha\in(0,\frac{1-p}{10})$ and $q(s_i,a) := \mathbb{P}(s^G|s_i,a)$.

We use $V_l^j(s_i)$ to denote the value function of state s_i in any of the MDPs where a_l is the optimal action in state s_i under any policy for which $\pi(s)=a_j$. From the construction it is clear that this value function in s_i does not depend on the entire policy but only the action taken in s_i.

Note that under H_l^i the optimal action is a_l with the second best option being a_0 and all remaining actions being even worse in the sense that $V_l^*(s_i) \geq V_l^0(s_i) \geq V_l^j(s_i)$ where V^* is the optimal value-function and V^l is the value-function under any policy where action a_l is taken in state s_i.

Theorem 9. *Let $\gamma\geq\frac{1}{2}$, $\delta<\frac{1}{4}$ and $u\in\mathbb{U}_1$ be given. Furthermore, assume that there exist $p\in(\frac{1}{2},1)$ and $\Delta>0$ such for every $\alpha\in(0,\frac{1-p}{10})$, every $l\in\{0,1,\dots,A\}$, and $s_i\in\{s_1,\dots,s_S\}$, it holds that*

$$V_0^*(s_i) - V_0^l(s_i) \geq \gamma\alpha\Delta, \qquad V_l^*(s_i) - V_l^0(s_i) \geq \gamma\alpha\Delta.$$

Then there exist constants $c_1,c_2>0$ such that if the total number of samples is less than

$$T \leq \frac{1}{c_1}\frac{SA\Delta^2 p(1-p)}{\varepsilon^2}\log\Big(\frac{S}{c_2\delta}\Big),$$

then for any algorithm $\mathcal{U}$ and any $\varepsilon \in (0, \frac{\Delta(1-p)}{10})$, there exists some MDP $M_i \in \mathbb{M}$ such that $\mathbb{P}(\|V_m^* - V^{\pi_\mathcal{U}}\| > \varepsilon) > \delta$.

Proof. If we choose $\alpha = \frac{2\varepsilon}{\Delta}$ for any $\varepsilon < \frac{\Delta(1-p)}{20}$ any suboptimal action is ε-bad.

Now 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 it follows 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 chooses 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^G in t trials.

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

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

From the Chernoff-Hoeffding bound and as shown in Gheshlaghi Azar et al. (2013), we have that $\mathbb{P}_0(\mathcal{E}) > \frac{3}{4}$, and thus, $\mathbb{P}_0(B \cap \mathcal{E}) > \frac{1}{2}$. From (Mortensen & Talebi, 2026, Theorem 6), we get that

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

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

$$t < \widetilde{T}(\varepsilon, \delta) := \frac{\Delta^2 p(1-p)}{128\varepsilon^2} \log\left(\frac{1}{8\delta}\right),$$

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 total number of transition samples is less than $\frac{S}{2}\widetilde{T}(\varepsilon, \delta)$, there has to be at least $\frac{S}{2}$ states in the set $\{s_i\}_{i=1}^S$ for which at least one action (apart from a_0) has been tried at most $\widetilde{T}(\varepsilon, \delta)$ times. We might without loss of generality 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 $i \leq S/2$. By the structure of the MDPs in $\mathbb{M}$ it is suffices to only consider algorithms that outputs an estimate of $\pi_{T_i}^\mathcal{U}$ based on samples from s_i since any other samples yields no information on $\pi^*(s_i)$.

Let us define the events $\Lambda_i := \{|V_{M_1}^*(s_i) - V^{\pi_{T_i}^\mathcal{U}}(s_i)| > \varepsilon\}$ 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

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

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

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

when $\frac{\delta S}{2} \leq 1$. By setting $\delta' = \delta\frac{S}{4}$, substituting back S' and A', and assuming that $S \geq 4$ and $A \geq 2$, we conclude that if the number of samples is smaller than

$$T = \frac{SA\Delta^2 p(1-p)}{1024} \log\Big(\frac{S}{64\delta}\Big)$$

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

E.2 Proofs of Lower Bounds

In this section, we give the proofs of the lower bounds using the constructions described above.

E.2.1 Value Learning Lower Bounds

Theorem 4. *Let $u \in \mathbb{U}_1$ be a utility function. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$, constants $c_1, c_2 > 0$, and a strictly positive function $\Phi : \mathbb{U}_1 \times [1,\infty) \to (0,\infty)$ such that for any RL algorithm $\mathcal{U}$ that outputs a Q-value $Q^{\mathcal{U}}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in \big(0, \bar{\varepsilon}\big)$, it holds: if the total number T of samples satisfies*

$$T \leq \frac{SA\Phi(u, H_\gamma)}{c_1\varepsilon^2} \log\Big(\frac{SA}{c_2\delta}\Big),$$

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

Proof. On the small MDP type sketched in Figure 1, we will give a lower bound on the optimal Q-functions on z_i for two different parameter choices of q, namely $q_0 = p$ and $q_1 = p + \alpha$.

Clearly, for any Q, we have that

$$Q^*(z) = \gamma \max_{\eta \in [0, H_\gamma]} \{\eta + qu(H_\gamma - \eta) + (1-q)u(-\eta)\}.$$

We use η_1 and η_0 to denote the respective maximizers of q_1 and q_0. We then have that

$$\begin{aligned} &Q_1^*(z) - Q_0^*(z) \\ &\quad = \gamma\big(\eta_1 + (p+\alpha)u(H_\gamma - \eta_1 + (1-p-\alpha)u(-\eta_1) - \eta_0 - pu(H_\gamma - \eta_0) - (1-p)u(-\eta_0)\big) \\ &\quad \geq \gamma\alpha\big(u(H_\gamma - \eta_0) - u(-\eta_0)\big), \end{aligned}$$

with $\Delta = \big(u(H_\gamma - \eta_0) - u(-\eta_0)\big)$. By Theorem 11, there exists $p \in (\frac{1}{2}, 1)$ such that $\Delta := u(H_\gamma - \eta_0) - u(\eta_0) > 0$. Thus, the lemma follows by Theorem 8 with $\Phi(u, H_\gamma) = \Delta^2 p(1-p)$. □

Theorem 5. *Let $u \in \mathbb{U}_1$ be a utility function for which $u'_+(0) < 1 < u'_-(0)$. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$ and 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 any $\varepsilon \in \big(0, \bar{\varepsilon}\big)$, the following holds: if the total number T of samples satisfies*

$$T \leq \frac{SA}{c_1 \varepsilon^2} \log\Big(\frac{SA}{c_2 \delta}\Big) |u(-H_\gamma)|^2 \left(1 - \left[\frac{1 - u'_+(0)}{u'_+(-H_\gamma) - u'_+(0)}\right]^2\right),$$

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

Proof. The idea of the proof is similar to that of Theorem 4 with the same MDP construction. Recall that

$$Q_1^*(z) - Q_0^*(z) \geq \gamma\alpha\big(u\big(H_\gamma - \eta_0\big) - u(-\eta_0)\big).$$

By the assumption on u, we have by Theorem 10, we can pick

$$p = \frac{1}{2} + \frac{1}{2} \frac{1 - u'_+(0)}{u'_+(-H_\gamma) - u'_+(0)} > \frac{1}{2}$$

to ensure that $\eta_0 = H_\gamma$ and so that $\Delta = u(H_\gamma - \eta_0) - u(-\eta_0) = -u(-H_\gamma)$. The result then follows from Theorem 8. □

E.2.2 Policy Learning Lower Bounds

Theorem 6. *Let $u \in \mathbb{U}_1$ be a utility function. There exist $\bar{\varepsilon}(u, H_\gamma) > 0$, a strictly positive function $\Phi : \mathbb{U}_1 \times [1, \infty) \to (0, \infty)$, and constants $c_1, c_2 > 0$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $\pi_{\mathcal{U}}$, any $\delta \in (0, \frac{1}{4})$, and any $\varepsilon \in (0, \bar{\varepsilon})$, we have: if the total number T of samples satisfies*

$$T \leq \frac{SA\Phi(u, H_\gamma)}{c_1 \varepsilon^2} \log\Big(\frac{S}{c_2 \delta}\Big),$$

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

Proof. We have for any s and $l \neq 0$

$$\begin{aligned} &V_0^*(s) - V_0^l(s) \\ &\quad = \gamma\big(\eta_1 + (p+\alpha)u(H_\gamma - \eta_1) + (1-p-\alpha)u(-\eta_1) - \eta_0 - pu(H_\gamma - \eta_0) - (1-p)u(-\eta_0)\big) \\ &\quad \geq \gamma\alpha\big(u(H_\gamma - \eta_0) - u(-\eta_0)\big), \end{aligned}$$

where η_1 is an optimizer of $\max_{\eta\in[0,H_\gamma]}\{\eta+(p+\alpha)u(H_\gamma-\eta)-(1-p-\alpha))u(-\eta)\}$ and η_0 is an optimizer of $\max_{\eta\in[0,H_\gamma]}\{\eta+pu(H_\gamma-\eta)-(1-p)u(-\eta)\}$. Similarly for all $l=1\dots,A$,

$$V_l^*(s)-V_l^0(s)\geq\gamma\alpha\big(u(H_\gamma-\eta_1)-u(-\eta_1)\big).$$

By Theorem 10, there exists $\bar{p}_0<1$ so that for all $p>\bar{p}_0$ it holds that $u(H_\gamma-\eta_0)-u(-\eta_0)>0$ and $\bar{p}_1<1$ so that for $p+\alpha>\bar{p}_1$ it holds that $u(H_\gamma-\eta_1)-u(-\eta_1)>0$. By picking p large enough and ε sufficiently small then α is also sufficiently small so that $p+2\alpha<1$ and both $u(H_\gamma-\eta_0)-u(-\eta_0)>0$ and $u(H_\gamma-\eta_1)-u(-\eta_1)>0$. Plugging in

$$\Delta=\min\{u(H_\gamma-\eta_1)-u(-\eta_1),u(H_\gamma-\eta_0)-u(-\eta_0)\},$$

the results follows from Theorem 9 with $\Phi(u,H_\gamma)=\Delta^2p(1-p)$. □

Theorem 7. *Let $u\in\mathbb{U}_1$ be a utility function for which $u'_+(0)<1<u'_-(0)$. There exist $\bar{\varepsilon}(u,H_\gamma)>0$ and constants $c_1,c_2>0$ such that for any RL algorithm $\mathcal{U}$ that outputs a policy $\pi_\mathcal{U}$, any $\delta\in(0,\frac{1}{4})$, and any $\varepsilon\in\big(0,\bar{\varepsilon}\big)$, the following holds: if the total number T of samples satisfies*

$$T\leq\frac{SA}{c_1\varepsilon^2}\log\Big(\frac{S}{c_2\delta}\Big)|u(-H_\gamma)|^2\bigg(1-\bigg[\frac{1-u'_+(0)}{u'_+(-H_\gamma)-u'_+(0)}\bigg]^2\bigg),$$

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

Proof. The idea of the proof is similar to that of Theorem 6. Recall that

$$V_0^*(s)-V_0^l(s)\geq\gamma\alpha\big(u(H_\gamma-\eta_0)-u(-\eta_0)\big),\qquad V_l^*(s)-V_l^0(s)\geq\gamma\alpha\big(u(H_\gamma-\eta_1)-u(-\eta_1)\big).$$

By Theorem 10, if both p and $p+\alpha\geq\max\{\frac{1}{2},1-\frac{1-u'_+(0)}{u'_+(-H_\gamma)-u'_+(0)}\}$, then the above right-hand sides are both equal to $\Delta=|u(-H_\gamma)|$. For sufficiently small ε, we get that α is sufficiently small so that $p=\frac{1}{2}+\frac{1}{2}\frac{1-u'_+(0)}{u'_+(-H_\gamma)-u'_+(0)}$ ensures this. Plugging in for p and Δ, the result follows. □

F Auxiliary Results

Proposition 1. *Let $u\in\mathbb{U}_0$ be any utility function for which $u(t)=t$ for all $t\geq0$. Then, $\mathrm{OCE}^u(X)=\mathbb{E}[X]$.*

Proof. Let u be any utility function such that the assumption in the proposition holds. Since any OCE satisfies that $\mathrm{OCE}^u(X)\leq\mathbb{E}[X]$, we need to show that $\mathrm{OCE}^u(X)\geq\mathbb{E}[X]$. By picking $\eta=\mathrm{Essinf}(X)$, we get that $u(X-\eta)=X-\eta$ and so

$$\eta+\mathbb{E}[u(X-\eta)]=\eta+\mathbb{E}[X-\eta]=\mathbb{E}[X].$$

Thus, by taking the supremum over all η, the result follows. □

Proposition 2. *For any piecewise differentiable utility $u\in\mathbb{U}_1$, it holds that $-u(-H_\gamma)\geq u'_+(-\frac{\gamma}{1-\gamma})$.*

Proof. Let $u\in\mathbb{U}_1$, and assume that u is differentiable. Since u is increasing and concave, we have for $t\geq1$ that

$$u(-t)=u(0)+\int_0^{-t}u'(x)dx=-\int_{-t}^0u'(x)dx\leq-\int_{-t}^{-t+1}u'(x)dx\leq-u'(-t+1).$$

Plugging in $t=H_\gamma$, we have that $-u(-H_\gamma)\geq u'(-\frac{\gamma}{1-\gamma})$. The result then follows from partitioning the integral over the different subdomains on which u is differentiable. □

Theorem 10. *Let $x > 0$ and X be a random variable with $\mathbb{P}(X = x) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Let $u \in \mathbb{U}_1^<$ be a strongly risk-averse utility function. If*

$$p \geq 1 - \frac{1 - u'_+(0)}{u'_+(-x) - u'_+(0)},$$

then $\eta^* = x$ is a solution to

$$\sup_{\eta \in [0,x]} \{\eta + pu(x - \eta) + (1 - p)u(-\eta)\}.$$

Proof. By definition, $\eta^* = x$ is a solution if and only if for all $\eta \in [0, x]$, it holds that

$$\eta + pu(x - \eta) + (1 - p)u(-\eta) \leq x + (1 - p)u(-x),$$

where trivially the inequality holds for $\eta = x$, and where for $\eta < x$ the inequality is equivalent to

$$1 - p \leq \frac{x - \eta - u(x - \eta)}{-u(-x) + u(-\eta) - u(x - \eta)} = \frac{1 - \frac{u(x-\eta)}{x-\eta}}{\frac{-u(-x)+u(-\eta)}{x-\eta} - \frac{u(x-\eta)}{x-\eta}}.$$

Since the upper bound above is valid for $\eta \in [0, x)$, we find a lower bound on the right-hand side that holds over $\eta \in [0, x)$. To this effect, first note that

$$\frac{-u(-x) + u(-\eta)}{x - \eta} = \frac{u(-x) - u(-\eta)}{-x - (-\eta)} \leq u'_+(-x),$$

and since for any $b > 1$, the map $t \mapsto \frac{1-t}{b-t}$ is decreasing on $t \in [0, 1]$, and $\frac{u(x-\eta)}{x-\eta} \leq u'_+(0)$ by concavity of u, we then have

$$\begin{aligned} \frac{1 - \frac{u(x-\eta)}{x-\eta}}{\frac{-u(-x)+u(-\eta)}{x-\eta} - \frac{u(x-\eta)}{x-\eta}} &\geq \frac{1 - \frac{u(x-\eta)}{x-\eta}}{u'_+(-x) - \frac{u(x-\eta)}{x-\eta}} \\ &\geq \frac{1 - u'_+(0)}{u'_+(-x) - u'_+(0)}, \end{aligned}$$

thus proving the lemma. □

Theorem 11. *Let $x > 0$ and $u \in \mathbb{U}_1$. Then, there exists some $\bar{p} \in (\frac{1}{2}, 1)$ such that for all $p \in (\bar{p}, 1)$,*

$$u(x - \eta^*) - u(-\eta^*) > 0,$$

where η^* is the solution to

$$\max_{\eta \in [0,x]} \{\eta + pu(x - \eta) + (1 - p)u(-\eta)\}.$$

Proof. We first note that $u(x - \eta) \geq 0$ and $-u(-\eta) \geq 0$ for all $\eta \in [0, x]$ and that $u(x - \eta) - u(-\eta) = 0$ if and only if both $\eta = 0$ and $u(t) = 0$ for all $t \geq 0$.

If $u(t)$ is not identically zero on $t \geq 0$, then $u(t) > 0$ for all $t > 0$. Hence, $u(x-\eta) = 0$ implies that $\eta = x$. However, since $-u(-x) \geq x > 0$, we can in this case conclude that $u(x-\eta^*)-u(-\eta^*) > 0$ for all $p \in (0, 1)$.

Now we treat the case where $u(t) = 0$ for all $t \geq 0$, which we partition into two cases: $u(t) = t$ for $t \in [-x, 0]$ (Case (i)) and $u(-x) < -x$ (Case (ii)).

Case (i). We have that $\eta + pu(x - \eta) + (1 - p)u(-\eta) = p\eta$, which is clearly maximized by $\eta^* = x$ and so, $u(x - \eta^*) - u(-\eta^*) = -u(-x) > 0$ for all $p \in (0, 1)$.

Case (ii). We note that $\eta^* = 0$ is a solution if and only if for all $\eta \in [0, x]$, it holds that $\eta + (1 - p)u(-\eta) \leq 0$ or equivalently $p \leq 1 - \frac{\eta}{-u(-\eta)}$. But since $u(-x) < -x$, the inequality is violated if $p > 1 + \frac{-x}{-u(-x)}$, and since $\frac{-x}{-u(-x)} > -1$, we can pick

$$\bar{p} = \frac{1 + \frac{x}{-u(-x)}}{2} \in \left(\tfrac{1}{2}, 1\right),$$

ensuring that $\eta = 0$ is not a solution (as $\eta = x$ violates the inequality). Thus, $u(x - \eta^*) - u(-\eta^*) > 0$. □