
A Reductions Approach to Risk-Sensitive Reinforcement Learning with Optimized Certainty Equivalents

Kaiwen Wang^{1,2} Dawen Liang² Nathan Kallus^{1,2} Wen Sun¹

Abstract

We study risk-sensitive RL where the goal is learn a history-dependent policy that optimizes some risk measure of cumulative rewards. We consider a family of risks called the optimized certainty equivalents (OCE), which captures important risk measures such as conditional value-at-risk (CVaR), entropic risk and Markowitz’s mean-variance. In this setting, we propose two meta-algorithms: one grounded in optimism and another based on policy gradients, both of which can leverage the broad suite of risk-neutral RL algorithms in an augmented Markov Decision Process (MDP). Via a reductions approach, we leverage theory for risk-neutral RL to establish novel OCE bounds in complex, rich-observation MDPs. For the optimism-based algorithm, we prove bounds that generalize prior results in CVaR RL and that provide the first risk-sensitive bounds for exogenous block MDPs. For the gradient-based algorithm, we establish both monotone improvement and global convergence guarantees under a discrete reward assumption. Finally, we empirically show that our algorithms learn the optimal history-dependent policy in a proof-of-concept MDP, where all Markovian policies provably fail.

1. Introduction

In reinforcement learning (RL), the classical objective by which we measure how good is a policy is the expected cumulative rewards along the trajectory (Sutton, 2004; Sutton and Barto, 2018). However, the mean reward of a choice (the choice of policy in RL being an example of such a choice) is a *risk-neutral* objective that is often observed to be inconsistent with human preferences, causing Allais Paradoxes (Allais, 1990; Kahneman and Tversky, 2013) and calling for

alternative *risk-sensitive* objectives (Yaari, 1987; Bowling et al., 2023). This is especially true in safety-critical settings (Artzner et al., 1999; Coronato et al., 2020; Wang et al., 2023b). In risk-sensitive RL (RSRL), the objective is some risk measure of cumulative rewards under a policy rather than the mean cumulative rewards (Howard and Matheson, 1972; Artzner et al., 1999).

In this paper, we propose a framework for solving RSRL with any static risk measure that can be expressed as an optimized certainty equivalent (OCE) (Ben-Tal and Teboulle, 2007). Specifically, for a utility function $u : \mathbb{R} \rightarrow [-\infty, \infty)$, the static OCE of a policy π is defined as

$$\text{OCE}_u(\pi) := \max_{b \in \mathbb{R}} \{b + \mathbb{E}_\pi[u(\sum_{h=1}^H r_h - b)]\}, \quad (1)$$

where the expectation is w.r.t. the random cumulative reward $\sum_{h=1}^H r_h$ under trajectories from π . The OCE captures several important risks for different choices of u . For example, with the hinge utility $u(t) = \min(t/\tau, 0)$, OCE_u becomes CVaR at level $\tau \in (0, 1]$, which measures the mean outcome among the worst τ -fraction of cases (Rockafellar and Uryasev, 2000). Moreover, OCE with the quadratic utility recovers Markowitz’s mean-variance (Markowitz, 1952) and the exponential utility recovers entropic risk (Föllmer and Schied, 2011). We present a primer on OCE and more examples in App. B. We remark that another risk-sensitive objective is the iterated risk, which captures risk at each time step and is orthogonal to our static risk objective (as we explain in related works Sec. 1.1).

A key technical challenge in OCE RL is that optimal policies are in general non-Markovian (*i.e.*, history-dependent), which renders this more complex than risk-neutral RL where a Markovian optimal policy always exists. In prior works, a promising solution is to augment the MDP with a new state that tracks the cumulative rewards so far, which has been studied for CVaR and spectral risks (Bäuerle and Ott, 2011; Bäuerle and Glauner, 2021; Bastani et al., 2022). In this paper, we show that the augmented MDP (AugMDP) can be generalized with an OCE-based reward function to solve OCE RL with Markovian augmented policies (Sec. 2.1). We then propose two meta-algorithms that reduce OCE RL to risk-neutral RL by using the AugMDP as a bridge.

First, in Sec. 3, we propose a meta-algorithm based opti-

¹Cornell Tech ²Netflix Research. Correspondence to: Kaiwen Wang <kaiwenw.github.io>. Work done as Netflix intern.

mistic RL oracles and we prove that its OCE regret can be bounded by the RL oracle’s regret within the AugMDP up to a factor depending on u (Thm. 3.2). We show that this approach can be instantiated with many RL oracles including UCB-VI for tabular MDPs (Azar et al., 2017) and Rep-UCB (Uehara et al., 2022) for low-rank MDPs, generalizing prior results in CVaR RL (Wang et al., 2023a; Zhao et al., 2024) to the OCE setting. While prior works in RSRL only considered model-based algorithms, we show that our approach is also amenable to the model-free optimistic RL oracle called GOLF (Jin et al., 2021). This establishes the first risk-sensitive RL bounds for exogenous block MDPs (Efroni et al., 2022; Xie et al., 2023).

Next, in Sec. 4, we propose a gradient-based meta-algorithm that uses policy gradient oracles and prove two types of guarantees: (1) convergence to the optimal OCE policy (Thm. 4.2), and (2) monotone improvement in a risk lower bound (Thm. 4.4), a novel property in RSRL theory. We instantiate our method with the natural policy gradient (NPG) oracle (Kakade, 2001) and prove these two guarantees under the standard conditions of policy optimization (Agarwal et al., 2021), plus a discrete reward assumption. Finally, our numerical simulation validates that our meta-algorithms can indeed learn the optimal OCE policy in a simple MDP where all Markovian policies (e.g., Dabney et al., 2018; Lim and Malik, 2022) must have sub-optimal performance. This validates our thesis that history-dependent, non-Markovian policies are required for OCE optimality.

In sum, our main contributions are the following:

1. We propose an optimistic meta-algorithm for OCE RL that generalizes all prior works in CVaR to the more general OCE setting. We also apply GOLF as a novel oracle and prove the first risk-sensitive bounds in the challenging exogenous block MDP. (Sec. 3)
2. We propose a gradient-based meta-algorithm for OCE RL that enjoys *both* local improvement and global convergence. These are the first finite-sample bounds for gradient-based RSRL algorithms. (Sec. 4)
3. In a numerical simulation, we show that our meta-algorithms indeed optimally solve OCE RL in a proof-of-concept MDP where Markovian policies provably have sub-optimal performance. (Sec. 5)

1.1. Related Works

From the theoretical side, prior works with risk-sensitive regret or PAC bounds are based on optimistic, model-based oracles for CVaR or entropic risk. For CVaR, Bastani et al. (2022); Wang et al. (2023a) proposed UCB-based algorithms for tabular MDPs by using the AugMDP proposed by Bäuerle and Ott (2011). Zhao et al. (2024) then extended this approach to low-rank MDPs using the Rep-UCB oracle

(Uehara et al., 2022) in the AugMDP. For entropic risk, Fei et al. (2020; 2021) proposed UCB-based algorithms for tabular MDPs based on exponential Bellman equations. These prior works are limited to model-based optimistic oracles and largely do not extend beyond tabular MDPs, while our work provides a general framework for applying both model-free and model-based oracles to rich-observation OCE RL, including the challenging exogenous block MDP (Efroni et al., 2022). Moreover, these prior works are limited to CVaR or entropic risk, while our work provides a unifying framework for the broad class of OCE risk measures.

Chen et al. (2024) derives regret bounds for static Lipschitz risks under function approximation which is more related to our setting. However, their optimistic algorithm requires bounded eluder dimension or witness rank, which does not easily apply to the challenging exogenous block MDP setting. Indeed, the Bellman eluder dimension can grow with the size of the exogenous state space which is exponentially large or infinite (Xie et al., 2023). In contrast, we employ a coverability argument in the augmented MDP to obtain the first PAC bound for exogenous block MDP in risk-sensitive RL. Moreover, our work also studies other methods such as policy gradients, which are complementary to optimistic algorithms.

There are also several prior works on gradient-based RSRL algorithms that prove *local and asymptotic convergence* (Chow and Ghavamzadeh, 2014; Tamar et al., 2015a;b; Chow et al., 2018; Greenberg et al., 2022). These guarantees are not ideal since (1) any local optima can be much worse than the global optima and (2) asymptotics do not quantify the rate of convergence which could be very slow. In contrast, we fill this gap by proving non-asymptotic *finite-sample* bounds for the *global convergence* of gradient-based algorithms, by leveraging recent advances in risk-neutral policy gradient theory (Agarwal et al., 2021; Xiao, 2022). Finally, we also prove monotone improvement which is a novel property for RSRL to the best of our knowledge.

From the empirical side, a popular framework for RSRL is distributional RL (DistRL), which has demonstrated state-of-the-art results in both online (Bellemare et al., 2017; Dabney et al., 2018; Keramati et al., 2020; Ma et al., 2020) and offline RL (Urpí et al., 2021; Ma et al., 2021). At a high level, these algorithms learn the reward-to-go distribution and then greedily act to optimize the risk of the learned distribution at each step. Unfortunately, this can diverge for some MDPs since it is not consistent with any Bellman equations. To address this, Lim and Malik (2022) proposed to apply DistRL in an AugMDP, but their algorithm still relies on the *strong assumption* that the optimal policy is Markov, which is false in our counterexample (Sec. 5).

Another solution is Trajectory Q-Learning (TQL) (Zhou et al., 2023) which learns history-dependent policies and dis-

tributional Q -functions. While learning history-dependent policies ensures that TQL converges to the optimal RSRL policy, the learning process can be quite costly in sample complexity, computation and memory since it essentially *ignores the Markov property of MDPs*. Instead, our approach searches over the minimal sufficient policy set, the Markovian policies in the AugMDP, to optimally solve OCE RL while also maintaining statistical and computational efficiency. Finally, TQL requires DistRL to learn the reward-to-go distribution, while our framework supports both regression-based and distribution-based methods for learning the AugMDP value function.

Iterated risk (a.k.a. dynamic risk). Another risk-sensitive RL objective is the iterated risk $\rho(r_1 + \rho(r_2 + \rho(r_3 + \dots)))$, where ρ is some risk measure. The iterated risk, *a.k.a.* dynamic risk, captures risk of the rewards-to-go at *each time step* and has been studied by many works (Ruszczyński, 2010; Du et al., 2023; Lam et al., 2023; Xu et al., 2023; Rigter et al., 2023; Liang and Luo, 2024). Dynamic risk also has close relations to distributionally robust MDPs, where an adversary can perturb the transition kernel at each time step (Wiesemann et al., 2013; Kumar et al., 2023; Bennett et al., 2024). The main benefit of iterated risk is that there exist Bellman equations and an optimal Markovian policy like in the classic RL setting. However, iterated risk may be overly risky in certain cases and overly conservative in other cases, and is also less interpretable due to the nested structure (Lim and Malik, 2022). In contrast, our static risk $\rho(r_1 + r_2 + r_3 + \dots)$ models the trajectory-level risk rather than at the step-level and is more interpretable. We highlight that static and dynamic risk are orthogonal objectives and the objective is ultimately the user’s choice.

2. Preliminaries

We consider an MDP with state space \mathcal{S} , action space \mathcal{A} , horizon H , transition kernels $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ and conditional reward distributions $R_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta([0, 1])$, where $\Delta(\mathcal{S})$ denotes the set of distributions on the set \mathcal{S} . A history-dependent policy $\pi = (\pi_h)_{h \in [H]}$ interacts with the MDP in the following way: start from an initial state s_1 , then, for each step $h = 1, 2, \dots, H$, sample an action $a_h \sim \pi_h(s_h, \mathcal{H}_h)$, collect a reward $r_h \sim R_h(s_h, a_h)$, observe the next state $s_{h+1} \sim P_h(s_h, a_h)$. Here, $\mathcal{H}_h = \{(s_i, a_i, r_i)\}_{i \in [h-1]}$ denotes the history so far. We denote the cumulative reward distribution under π as $Z(\pi) := \sum_{h=1}^H r_h$, where randomness arises from both the MDP and the policy π . We also assume that $Z(\pi) \in [0, 1]$ almost surely for all policies (Jiang and Agarwal, 2018). The goal of OCE RL is to learn a history-dependent policy that maximizes the OCE objective:

$$\text{OCE}_u^* := \max_{\pi \in \Pi_{\text{HD}}} \text{OCE}_u(\pi), \quad (2)$$

where Π_{HD} denotes the set of history-dependent policies.

We focus on the online setting, where at each round $k = 1, 2, \dots, K$, the learner outputs a (history-dependent) policy $\pi^k \in \Pi_{\text{HD}}$. The learner’s goal is to minimize the total sub-optimality across K rounds:

$$\text{Reg}_{\text{OCE}}(K) := \sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k).$$

Probably approximately correct (PAC) bounds are upper bounds on $\text{Reg}_{\text{OCE}}(K)$ that hold with high probability over the randomness of the learner. In particular, PAC bounds imply that at least one policy has high OCE, *i.e.*, $\min_{k \in [K]} \{\text{OCE}_u^* - \text{OCE}_u(\pi^k)\} \leq \frac{1}{K} \text{Reg}_{\text{OCE}}(K)$. If the learner only executes policy π^k at round k , then $\text{Reg}_{\text{OCE}}(K)$ is also called a *regret* bound.

2.1. Augmented MDP for OCE

The key challenge in RSRL with static risk is that the optimal policies are history-dependent, as there are no Bellman-like equations. A promising solution from prior works is to augment the MDP with a scalar state, which has been studied for CVaR and spectral risks (Bäuerle and Ott, 2011; Bäuerle and Glauner, 2021) – in this augmented state space, there does exist a Markovian optimal policy, bypassing the need to maintain history-dependent policies. In this section, we describe the augmented MDP (AugMDP) and extend it to the OCE setting, generalizing prior constructions.

The AugMDP has states $(s_h, b_h) \in \mathcal{S}_{\text{aug}} := \mathcal{S} \times [-1, 1]$, where s_h is the original MDP state and b_h is a new scalar state called the budget, which intuitively tracks the cumulative rewards so far. The initial budget $b_1 \in [0, 1]$ is chosen by the learning algorithm and the budget transitions via $b_{h+1} = b_h - r_h$ where r_h is reward at step h . The reward of the AugMDP is defined as $r_{\text{aug}}^h(s, b, a) = 0$ for $h < H$ and $r_{\text{aug}}^H(s, b, a) = \mathbb{E}_{r \sim R_H(s, a)}[u(r - b)]$, where $V_u^{\max} = \max_{c \in [-1, 1]} |u(c)|$ measures the scale of r_{aug} . The augmented reward encodes the utility u and is critical for the following optimality theorem.

Theorem 2.1. *There exists an initial budget $b_1^* \in [0, 1]$ s.t. the optimal risk-neutral π_{aug}^* in the AugMDP with initial budget b_1^* achieves optimal OCE in the original MDP.*

Since the optimal risk-neutral policy π_{aug}^* is Markovian and deterministic, this theorem crucially mitigates the need to maintain history-dependent policies. We note that $(\pi_{\text{aug}}^*, b_1^*)$ is a special history-dependent policy in the original MDP. Thus, Thm. 2.1 proves that the set of (π_{aug}, b_1) , where π_{aug} is a Markovian augmented policy and b_1 is an initial budget, is a special class of history-dependent policies that is sufficient for OCE optimality. Moreover, a natural implementation of Thm. 2.1 is to solve risk-neutral RL in the AugMDP, which is the key idea behind our algorithms. We remark that our AugMDP construction generalizes the CVaR AugMDP from (Bäuerle and Ott, 2011) to any OCE risk.

For the AugMDP, we also define its quality function $Q_{\text{aug}}^{\pi,h}(s_h, b_h, a_h) = \mathbb{E}_{\pi}[u(\sum_{t=h}^H r_t - b_h) \mid s_h, a_h, b_h]$ and value function $V_{\text{aug}}^{\pi,h}(s_h, b_h) = \mathbb{E}_{\pi}[Q_{\text{aug}}^{\pi,h}(s_h, b_h, a_h) \mid s_h, b_h]$. Also, let $V_{\text{aug}}^*(s_h, b_h) = \max_{\pi \in \Pi_{\text{HD}}} V_{\text{aug}}^{\pi,h}(s_h, b_h)$ be the optimal value function of the AugMDP.

Proof of Thm. 2.1. The main step is to write OCE_u^* in terms of the optimal value of the AugMDP:

$$\begin{aligned} \text{OCE}_u^* &\stackrel{(i)}{=} \max_{\pi \in \Pi_{\text{HD}}} \max_{b \in [0,1]} \{b + \mathbb{E}[u(Z(\pi) - b)]\} \\ &= \max_{b \in [0,1]} \{b + \max_{\pi \in \Pi_{\text{HD}}} \mathbb{E}[u(Z(\pi) - b)]\} \\ &\stackrel{(ii)}{=} \max_{b \in [0,1]} \{b + \max_{\pi \in \Pi_{\text{HD}}} V_{\text{aug}}^{\pi,1}(s_1, b)\} \\ &\stackrel{(iii)}{=} \max_{b \in [0,1]} \{b + V_{\text{aug}}^*(s_1, b)\}. \end{aligned} \quad (3)$$

Step (i) uses OCE's definition in Eq. (1) and $Z(\pi) \in [0, 1]$ w.p. 1, so the \max_b can be taken over $[0, 1]$. Step (ii) uses the definition of b_h and r_{aug} to deduce that $\mathbb{E}_{\pi}[u(\sum_h r_h - b_1)] = \mathbb{E}_{\pi}[u(r_H - b_H) \mid b_1] = \mathbb{E}_{\pi}[r_{\text{aug}}^H(s_H, b_H, a_H) \mid b_1] = V_{\text{aug}}^{\pi,1}(s_1, b_1)$ for any π and b_1 . Step (iii) uses the classical RL result that the optimal risk-neutral policy π_{aug}^* achieves the maximum value V_{aug}^* over history-dependent policies (Puterman, 2014). Thus, letting $b_1^* = \arg \max_{b \in [0,1]} \{b + V_{\text{aug}}^*(s_1, b)\}$, we have:

$$\begin{aligned} \text{OCE}_u(\pi_{\text{aug}}^*, b_1^*) &= \max_{b \in [0,1]} \{b + \mathbb{E}_{\pi_{\text{aug}}^*, b_1^*}[u(Z(\pi_{\text{aug}}^*) - b)]\} \\ &\geq b_1^* + \mathbb{E}_{\pi_{\text{aug}}^*, b_1^*}[u(Z(\pi_{\text{aug}}^*) - b_1^*)] \\ &= b_1^* + V_{\text{aug}}^*(s_1, b_1^*) = \text{OCE}_u^*. \end{aligned}$$

This proves that $(\pi_{\text{aug}}^*, b_1^*)$ achieves the optimal OCE_u^* . \square

Remark: the AugMDP is easy to simulate. The transition function for the augmented state b_h is simply $b_{h+1} = b_h - r_h$, which is *known and deterministic*. This ensures that simulating the AugMDP is *computationally efficient*. This is also important for proving bounds in the AugMDP. For example, model-based algorithms do not need to learn the AugMDP transitions (since it is known), so they generally inherit the same bounds as in the original MDP.

3. Meta-Algorithm with Optimism

In this section, we propose a meta-algorithm for OCE RL based on optimistic RL oracles. Our framework generalizes existing works in CVaR RL to more OCE risk measures and to more complex MDPs such as exogenous block MDPs. At a high level, the meta-algorithm implements the intuition from Thm. 2.1 and has two parts: (1) we apply an RL oracle in the AugMDP to learn π_{aug}^* , and (2) then select the right initial budget b_1 . We start by formally defining optimistic oracles, which are RL algorithms that can explore using optimistic value functions:

Definition 3.1 (Optimistic oracle). At each round $k = 1, 2, \dots, K$, the oracle OPTALG first outputs an optimistic value function $\hat{V}_{1,k}(\cdot)$, then sees an initial state $s_{1,k}$, and finally outputs a policy π^k which is rolled-out from $s_{1,k}$ to collect data. Two conditions must be satisfied:

1. (Optimism) $\hat{V}_{1,k}(s_1) \geq V_1^*(s_1) - \varepsilon_k^{\text{opt}}$ for all initial states s_1 , where $\varepsilon_k^{\text{opt}}$ are slack variables;
2. (Bounded Regret) $\sum_{k=1}^K (\hat{V}_{1,k}(s_{1,k}) - V_1^{\pi^k}(s_{1,k})) \leq V^{\max} \text{Reg}_{\text{Opt}}(K)$, where V^{\max} is the scale of cumulative rewards and $\text{Reg}_{\text{Opt}}(K)$ bounds the total sub-optimality gap across K rounds. For this to be useful, $\text{Reg}_{\text{Opt}}(K)$ should be sublinear in K , i.e., $\mathcal{O}(\sqrt{K})$.

We remark that the conditions of OPTALG can be satisfied in the Augmented MDP by modifying standard optimistic algorithms such as UCB-VI, Rep-UCB and GOLF. For model-based oracles (e.g., UCB-VI, Rep-UCB), the augmented transition kernel for b is known and deterministic and thus introduces no extra statistical or computational complexity for model learning in the augmented MDP. For model-free oracles (e.g., GOLF), we need to be more careful and redefine important concepts such as completeness and coverability in the augmented MDP. We provide detailed algorithms and proofs for satisfying OPTALG in App. D.

Given an oracle OPTALG, our optimistic meta-algorithm (Alg. 1) proceeds as follows: at each round $k = 1, 2, \dots, K$, we query OPTALG in the AugMDP and obtain an optimistic value function $\hat{V}_{1,k}(\cdot)$. Then, we compute the initial budget by solving:

$$\hat{b}_k \leftarrow \arg \max_{b_1 \in [0,1]} \{b_1 + \hat{V}_{1,k}(s_1, b_1)\}. \quad (4)$$

Finally, we give the initial state (s_1, \hat{b}_k) to OPTALG, receive policy π^k , and proceed to the next round. We now state our main result for Alg. 1.

Theorem 3.2. Assuming that OPTALG satisfies Def. 3.1, running Alg. 1 for K rounds ensures that:

$$\sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k, \hat{b}_k) \leq V_u^{\max} \text{Reg}_{\text{Opt}}(K) + \sum_{k=1}^K \varepsilon_k^{\text{opt}}.$$

This theorem is a deterministic statement which bounds Alg. 1's OCE regret by the oracle's AugMDP regret, up to a scaling of V_u^{\max} . Here, V_u^{\max} intuitively measures the statistical hardness of estimating OCE_u relative to \mathbb{E} , which has $V_{\mathbb{E}}^{\max} = 1$. For instance, $V_{\text{CVaR}_{\tau}}^{\max} = \tau^{-1}$. Another intuition for V_u^{\max} is the Lipschitz constant of the risk measure (Liang and Luo, 2024; Chen et al., 2024). Please see App. B for more examples. We now instantiate this result with three optimistic oracles from the RL literature.

Algorithm 1 Meta-algorithm for optimistic oracles

- 1: **Input:** number of rounds K , optimistic oracle OPTALG satisfying Def. 3.1.
- 2: **for** round $k = 1, 2, \dots, K$ **do**
- 3: Query OPTALG in AugMDP for value func. $\widehat{V}_{1,k}(\cdot)$.
- 4: Compute initial budget \widehat{b}_k via Eq. (4).
- 5: Give initial state (s_1, \widehat{b}_k) to OPTALG and receive π^k .
- 6: Collect a trajectory with π^k starting from (s_1, \widehat{b}_k) .
- 7: **end for**

First, UCB-VI is a model-based algorithm for *tabular MDPs* with near-minimax-optimal dependence on $|\mathcal{S}|, |\mathcal{A}|, K$ (Azar et al., 2017). While the AugMDP is not tabular due to the real-valued augmented state b_h , the augmented state has a *known and deterministic* transition ($b_{h+1} = b_h - r_h$); hence, UCB-VI *only needs to learn the original MDP’s transition model* when operating in the AugMDP. Indeed, by adapting the proof of Wang et al. (2023a) in Thm. D.1, we prove that UCB-VI satisfies Def. 3.1’s conditions in the AugMDP with $\varepsilon_k^{\text{opt}} = 0$ and $\text{Reg}_{\text{Opt}}(K) \leq \tilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|HK})$. Combined with Thm. 3.2, this immediately implies an OCE regret bound of $\tilde{O}(V_u^{\max} \sqrt{SAHK})$, which is also near-minimax-optimal in $|\mathcal{S}|, |\mathcal{A}|, K$ for tabular MDPs. Moreover, in App. D.4, we obtain tight rates for CVaR using second-order oracles, showing our upper bound is tight.

The second oracle we discuss is Rep-UCB (Uehara et al., 2022), a model-based optimistic algorithm for *low-rank MDPs* (defined in Def. D.2), which are MDPs with a rank d transition kernel and potentially infinite state space (Agarwal et al., 2020); low-rank MDPs capture tabular MDPs and linear MDPs. Again, the original MDP being low-rank does not necessarily imply that the augmented MDP is low-rank. Fortunately, as before, the augmented state b_h has a known transition function and thus Rep-UCB only needs to learn the low-rank transitions of the original MDP, even when running in the AugMDP. Indeed Zhao et al. (2024) proved that Rep-UCB enjoys the same PAC bound in the AugMDP as in original low-rank MDP. We adapt their proof in Thm. D.4 and show that Rep-UCB satisfies Def. 3.1’s conditions in the AugMDP with $\text{Reg}_{\text{Opt}}(K) \leq \tilde{O}(H^3 |\mathcal{A}| d^2 \sqrt{K})$ and $\sum_k \varepsilon_k^{\text{opt}} \leq \text{Reg}_{\text{Opt}}(K)$. Together with Thm. 3.2, we thus have an OCE PAC bound of $\tilde{O}(V_u^{\max} H^3 |\mathcal{A}| d^2 \sqrt{K})$, generalizing Zhao et al. (2024) to any OCE risk measure.

We have presented two examples with model-based oracles, whose bounds naturally lifts to the AugMDP since the augmented state b_h has a known and deterministic transition. In general, model-based oracles enjoy the same bounds in the AugMDP as in the original MDP, making them convenient oracles for our meta-algorithm. Next, we discuss a model-free oracle to solve the challenging exogenous block MDP, where it is not tractable to learn the transition model.

3.1. Bounds for exogenous block MDPs

We now present a third optimistic oracle called GOLF (Jin et al., 2021) and prove the first risk-sensitive bound in the challenging exogenous block MDP (Ex-BMDP) problem, due to Efroni et al. (2022).

Definition 3.3. An Ex-BMDP has latent states $(z_h^{\text{en}}, z_h^{\text{ex}}) \in \mathcal{Z}_h^{\text{en}} \times \mathcal{Z}_h^{\text{ex}}$, where $\mathcal{Z}_h^{\text{en}}$ is the endogenous part which is tabular, and $\mathcal{Z}_h^{\text{ex}}$ is the exogenous part which is arbitrarily large. The endogenous latent transition conditions on the action $z_{h+1}^{\text{en}} \sim P_h^{\text{en}}(z_h^{\text{en}}, a_h)$, while the exogenous does not $z_{h+1}^{\text{ex}} \sim P_h^{\text{ex}}(z_h^{\text{ex}})$. The observation state is emitted from the latents $s_h \sim o_h(z_h^{\text{en}}, z_h^{\text{ex}})$ and, importantly, there exists a unique function ϕ_h^* such that $\phi_h^*(s_h) = (z_h^{\text{en}}, z_h^{\text{ex}})$ recovers the latents from the observation.

GOLF is a value-based (model-free) RL algorithm that achieves optimism by optimizing over a version space of Q -functions (Jin et al., 2021). The version space consists of candidate Q -functions from a function class $\mathcal{F} = (\mathcal{F}_h)_{h \in [H]}$, $\mathcal{F}_h \subset \mathcal{S}_{\text{aug}} \times \mathcal{A} \rightarrow \mathbb{R}$ whose Bellman backup is consistent with the data up to statistical noise. Value-based methods have shown promise in important applications such as LLM post-training (Mudgal et al., 2024; Zhou et al., 2025; Wang et al., 2025b). Due to space constraints, we defer the formal definition of GOLF in the AugMDP to App. D.3.

The standard condition required for Bellman-consistent algorithms like GOLF is completeness, which is required for Bellman backups to be well-behaved (Tsitsiklis and Van Roy, 1996; Munos and Szepesvári, 2008). Since we apply GOLF in the AugMDP, we posit completeness w.r.t. the AugMDP Bellman operator denoted as $\mathcal{T}_{\text{aug}}^{*,h}$.

Assumption 3.4. $\mathcal{T}_{\text{aug}}^{*,h} f_{h+1} \in \mathcal{F}_h$ for all $f_{h+1} \in \mathcal{F}_{h+1}$.

Another condition we require is discrete rewards (Assump. 3.5), which we use for bounding coverability in the AugMDP (Xie et al., 2023). Many safety-critical RL problems have discrete rewards, e.g., in sepsis treatment, reward is the patient’s discharge status (Johnson et al., 2016).

Assumption 3.5. We have a finite set $\mathcal{B} \subset [0, 1]$ such that $Z(\pi) \in \mathcal{B}$ w.p. 1 for all π , where $Z(\pi)$ is the cumulative reward of roll-outs from π .

This assumption can be satisfied by discretizing rewards as is common in distributional RL (Bellemare et al., 2017; Imani et al., 2024; Ayoub et al., 2024; Wang et al., 2025a) and in theory (Bastani et al., 2022; Wang et al., 2023a).

We now state our main result for Ex-BMDPs.

Theorem 3.6. For $\delta \in (0, 1)$, under Assumps. 3.4 and 3.5, running Alg. 1 with OPTALG as GOLF (Alg. 5 in App. D.3) ensures that w.p. $1 - \delta$,

$$\begin{aligned} \sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k, \widehat{b}_k) \\ \leq \tilde{O}(V_u^{\max} H \sqrt{|\mathcal{B}||\mathcal{Z}^{\text{en}}||\mathcal{A}|K \log(|\mathcal{F}|/\delta)}). \end{aligned}$$

Notably, the bound scales with the endogenous state size $|\mathcal{Z}^{\text{en}}|$ and the function class's complexity $\log(|\mathcal{F}|)$; indeed, all salient problem parameters matches the original risk-neutral bound from Xie et al. (2023). The completeness assumption may be weakened to realizability if we have local simulator access (Mhammedi et al., 2024); but otherwise is unavoidable even in risk-neutral RL (Jin et al., 2021; Xie et al., 2023). Thus, Thm. 3.6 demonstrates the versatility of our meta-algorithm to solve challenging MDPs that go beyond tabular and low-rank.

To prove Thm. 3.6, we need to show that GOLF satisfies the OPTALG conditions in Def. 3.1 with $\text{Reg}_{\text{Opt}}(K) \leq \tilde{O}(H\sqrt{|\mathcal{B}||\mathcal{Z}^{\text{en}}||\mathcal{A}|K})$ and $\varepsilon_k^{\text{opt}} = 0$. Due to completeness, we can invoke GOLF's coverability regret bound from Xie et al. (2023). Then, we prove that the AugMDP coverability is bounded by $|\mathcal{B}||\mathcal{Z}^{\text{en}}||\mathcal{A}|$ in Ex-BMDPs via a change-of-measure argument. Then, Thm. 3.6 immediately follows from Thm. 3.2. Please see App. D.3 for the full proof.

Remark on Discrete Rewards. It is possible to relax Assump. 3.5 by rounding rewards to fixed bins spaced ϵ apart. This discretization of width ϵ introduces an additional regret of at most $O(V_u^{\max}K\epsilon)$ and creates $|\mathcal{B}| = O(1/\epsilon)$ (Wang et al., 2023a). Thus, if an algorithm has a regret bound of $O(\sqrt{|\mathcal{B}|K})$, then applying this discretization would yield a regret bound of $O(V_u^{\max}\sqrt{K/\epsilon} + V_u^{\max}K\epsilon)$. Then, we can choose $\epsilon = \Theta(K^{-1/3})$ to yield a regret bound of $O(V_u^{\max}K^{2/3})$ which is sublinear. Thus, even when rewards are continuous, we can avoid Assump. 3.5 by discretizing rewards and paying a regret with rate $O(K^{2/3})$ instead of the minimax-optimal rate of $O(K^{1/2})$.

3.2. Proof for Main Reduction (Thm. 3.2)

Proof of Thm. 3.2. First, we upper bound OCE_u^* :

$$\begin{aligned} \text{OCE}_u^* &\stackrel{(i)}{=} b_1^* + V_{\text{aug}}^{\star,1}(s_1, b_1^*) \\ &\stackrel{(ii)}{\leq} b_1^* + \hat{V}_{1,k}(s_1, b_1^*) + \varepsilon_k^{\text{opt}} \\ &\stackrel{(iii)}{\leq} \hat{b}_k + \hat{V}_{1,k}(s_1, \hat{b}_k) + \varepsilon_k^{\text{opt}}, \end{aligned}$$

where (i) is by Sec. 2.1, (ii) is by optimism of the oracle (first condition of Def. 3.1), and (iii) is by the definition of \hat{b}_k via Eq. (4). Second, we lower bound $\text{OCE}_u(\pi^k, \hat{b}_k)$:

$$\begin{aligned} \text{OCE}_u(\pi^k, \hat{b}_k) &= \max_{b \in [0,1]} \{b + \mathbb{E}[u(Z(\pi^k, \hat{b}_k) - b)]\} \\ &\geq \hat{b}_k + \mathbb{E}[u(Z(\pi^k, \hat{b}_k) - \hat{b}_k)] \\ &= \hat{b}_k + V_{\text{aug}}^{\pi^k,1}(s_1, \hat{b}_k). \end{aligned}$$

Thus, $\sum_k \text{OCE}_u^* - \text{OCE}_u(\pi^k, \hat{b}_k)$ is bounded by:

$$\begin{aligned} &\leq \sum_k \hat{V}_{1,k}(s_1, \hat{b}_k) - V_{\text{aug}}^{\pi^k,1}(s_1, \hat{b}_k) + \varepsilon_k^{\text{opt}} \\ &\leq V_u^{\max} \text{Reg}_{\text{Opt}}(K) + \sum_k \varepsilon_k^{\text{opt}}, \end{aligned}$$

Algorithm 2 Meta-algorithm for PO oracles

- 1: **Input:** number of rounds K , initial budget set \mathcal{B} (Assump. 3.5), oracle POALG (Def. 4.1).
- 2: Let $d_1^{\text{aug}} = (\delta(s_1), \text{Unif}(\mathcal{B}))$ be the init. augmented state distribution. Initialize POALG with d_1^{aug} .
- 3: **for** round $k = 1, 2, \dots, K$ **do**
- 4: Obtain policy π^k and value estimate $\hat{V}_1^{\pi^k}$ from running POALG in the AugMDP.
- 5: Compute the initial budget \hat{b}_k with Eq. (5).
- 6: Give init. augmented state (s_1, \hat{b}_k) to POALG.
- 7: **end for**

where the last inequality is due to the oracle's regret bound (second condition of Def. 3.1). \square

Computational Complexity of Alg. 1. There are two sources of computational cost. The first cost is OPTALG: from the examples we discussed, UCB-VI and Rep-UCB are computationally and oracle efficient while GOLF is inefficient due to its version space optimism (Dann et al., 2018). So long as OPTALG is computationally efficient, then calls to OPTALG should also be efficient. The second cost is \hat{b}_k : computing Eq. (4) exactly is difficult since the objective is non-convex. However, it can be efficiently approximated by projecting rewards to a discrete grid with spacing $\iota = \mathcal{O}(1/K)$, so each round's Eq. (4) can be done in $\mathcal{O}(K)$ time. Notably, the approximation error per round is at most $H\iota$ and across K rounds is at most $\mathcal{O}(H\iota \cdot K) = \mathcal{O}(H)$, which is a lower order term (Wang et al., 2023a, App. H). Thus, provided that the oracle is efficient, Alg. 1 is both computationally and statistically efficient for OCE RL.

4. Meta-Algorithm with Policy Optimization

In this section, we propose a meta-algorithm that uses standard policy optimization (PO) oracles, which enjoy a new type of guarantee called *local improvement* that ensures the policy's performance monotonically improves at each round. First, we formalize the conditions needed for the risk-neutral PO oracle, which we remark is satisfied by gradient-based algorithms such as REINFORCE, NPG and PPO (Kakade, 2001; Agarwal et al., 2021; Grudzien et al., 2022).

Definition 4.1 (PO Oracle). Let d_1 denote an exploratory initial state distribution. At each round $k = 1, 2, \dots, K$, the oracle POALG outputs a policy π^k and its value estimate $\hat{V}_1^{\pi^k}(\cdot)$ such that $\mathbb{E}_{s_1 \sim d_1} |V_1^{\pi^k}(s_1) - \hat{V}_1^{\pi^k}(s_1)| \leq \varepsilon_k^{\text{po}}$. Then, there are two conditions on $\{\pi^k\}_{k \in [K]}$'s performance:

1. (Approx. Improvement) $V_1^{\pi^{k+1}}(d_1) \geq V_1^{\pi^k}(d_1) - \varepsilon_k^{\text{po}}$, where $V(d_1) := \mathbb{E}_{s_1 \sim d_1} V(s_1)$ and $\varepsilon_k^{\text{po}}$ are slack vars;
2. (Global Convergence) $\sum_{k=1}^K (V_1^{\pi^k}(d_1) - V_1^{\pi^*}(d_1)) \leq$

$V^{\max} \text{Reg}_{\text{PO}}(K)$ where $\text{Reg}_{\text{PO}}(K)$ bounds the total sub-optimality gap of POALG across K rounds.

Since this oracle will be applied in the AugMDP, the V^{π^k} term will be set as $V_{\text{aug}}^{\pi^k}$. Moreover, we remark that the small estimation error condition $\mathbb{E}_{s_1 \sim d_1} |V_1^{\pi^k}(s_1) - \hat{V}_1^{\pi^k}(s_1)| \leq \varepsilon_k^{\text{po}}$ can be satisfied by via off-policy evaluation (OPE) by using the data collected in prior rounds (Munos and Szepesvári, 2008; Chang et al., 2022; Kallus and Uehara, 2020). Given access to simulator, we can of course also estimate the value by collecting online Monte Carlo roll-outs.

Given the oracle POALG, our meta-algorithm (Alg. 2) proceeds as follows: at each round $k = 1, 2, \dots, K$, we query POALG with the initial budget distribution $\text{Unif}(\mathcal{B})$, where \mathcal{B} is the set of possible initial budgets that we assume to be discrete (Assump. 3.5). We require Assump. 3.5 because PO oracles, unlike optimistic ones, do not strategically explore and thus rely on an exploratory initial state distribution even in risk-neutral RL (Agarwal et al., 2021). Then, POALG returns a policy π^k and an estimate of its value $\hat{V}_1^{\pi^k}(\cdot)$, with which we compute the initial budget \hat{b}_k via

$$\hat{b}_k \leftarrow \arg \max_{b_1 \in \mathcal{B}} \{b_1 + \hat{V}_{1,k}(s_1, b_1)\}. \quad (5)$$

This is different from the optimistic version (Eq. (4)) since the maximization is restricted to \mathcal{B} . We now state our main results for Alg. 2, starting with global convergence.

Theorem 4.2 (Global Convergence). *Under Assump. 3.5 and assuming POALG satisfies the global convergence criterion of Def. 4.1, running Alg. 2 ensures that:*

$$\begin{aligned} \sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k, \hat{b}_k) \\ \leq |\mathcal{B}|(V_u^{\max} \text{Reg}_{\text{PO}}(K) + 2 \sum_k \varepsilon_k^{\text{po}}). \end{aligned}$$

This theorem bounds Alg. 2’s sub-optimality by the PO oracle’s sub-optimality, up to a factor of $|\mathcal{B}|V_u^{\max}$, plus the value estimation errors. Thus, if POALG has small value estimation errors and converges to the optimal policy, then Alg. 2 converges to the optimal OCE policy. We highlight that Thm. 4.2 is a finite-sample PAC bound whereas prior guarantees for risk-sensitive policy gradients have been asymptotic (Tamar et al., 2015a;b).

Next, toward stating the local improvement result, we introduce the risk lower bound (RLB) $\text{RLB}^{(k)}$ defined as:

$$\text{RLB}(\pi) := \max_{b_1 \in \mathcal{B}} \{b_1 + V_{\text{aug}}^{\pi,1}(s_1, b_1)\}. \quad (6)$$

Notice that the lower bound is tight at π^* : $\text{RLB}(\pi^*) = \text{OCE}_u^*$, due to Eq. (3). In general, RLB approximately lower bounds the true OCE of (π^k, \hat{b}_k) , as the following lemma proves.

Lemma 4.3 (RLB). *The $\text{RLB}^{(k)}$ approximately lower bounds the true OCE of π^k with initial budget \hat{b}_k :*

$$\text{RLB}^{(k)} - \text{OCE}_u(\pi^k, \hat{b}_k) \leq 2|\mathcal{B}|\varepsilon_k^{\text{po}}. \quad (7)$$

Moreover, the lower bound is tight on average:

$$\sum_k \text{OCE}_u(\pi^k, \hat{b}_k) - \text{RLB}^{(k)} \leq |\mathcal{B}|V_u^{\max} \text{Reg}_{\text{PO}}(K).$$

We now state the approximate improvement guarantee.

Theorem 4.4 (Local Improvement). *Under Assump. 3.5 and assuming POALG satisfies the approximate improvement criterion of Def. 4.1, running Alg. 2 ensures that:*

$$\forall k \in [K] : \text{RLB}^{(k+1)} \geq \text{RLB}^{(k)} - |\mathcal{B}|\varepsilon_k^{\text{po}}.$$

This theorem shows that Alg. 2’s policies are approximately improving at each step, up to an error $|\mathcal{B}|\varepsilon_k^{\text{po}}$. Hence, if the value estimation error from POALG is small, then the new policy’s RLB cannot be much worse than the current. While local improvement is known for risk-neutral RL (Agarwal et al., 2021), to the best of our knowledge, Thm. 4.4 is the first analog in risk-sensitive RL. In sum, Thm. 4.2 and Thm. 4.4 provide complementary global and local guarantees for OCE RL, stated in a finite-sample manner.

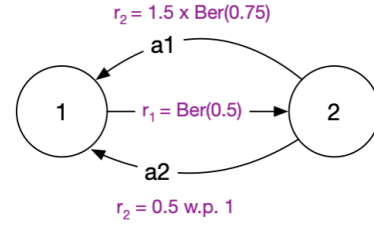


Figure 1. A simple MDP where the optimal CVaR policy is history-dependent. Each policy’s cumulative reward distribution is shown in the table below.

Action at s_2	Cumulative-Reward Dist.	CVaR _{0.25}
a_1 (Markov)	$\{(0, 1/8), (1, 1/8), (1.5, 3/8), (2.5, 3/8)\}$	0.5
a_2 (Markov)	$\{(0.5, 1/2), (1.5, 1/2)\}$	0.5
a_1 if $r_1=0$ a_2 if $r_1=1$	$\{(0, 1/8), (1.5, 7/8)\}$	0.75

Table 1. The cumulative reward distribution of Markovian policies (rows 1-2) and the optimal policy (row 3). For the distribution, $\{(v_i, p_i)\}_i$ denotes a random variable that takes value v_i w.p. p_i , s.t. $\sum_i p_i = 1$. The optimal CVaR policy has $\text{CVaR}_{0.25}^* = 0.75$ while both Markovian policies have $\text{CVaR}_{0.25} = 0.5$.

4.1. Case Study: Natural Policy Gradient

As an example POALG, we consider natural policy gradients (NPG; Kakade, 2001) whose core ideas underpin TRPO and PPO (Schulman et al., 2015; 2017). Let π_h^θ denote

OCE	Optimistic Alg (Alg. 1)	PO Alg (Alg. 2)	Best Markovian	Markovian Optimal?
$\mathbb{E}[X] - \text{Var}(X)$	1.07 ± 0.01	1.06 ± 0.01	0.95	✗
$\mathbb{E}[X] - 2\text{Var}(X)$	0.81 ± 0.01	0.75 ± 0.08	0.5	✗
$\text{Entr}_{-1.0}(X)$	1.25 ± 0.01	1.23 ± 0.03	1.25	✓
$\text{Entr}_{-2.0}(X)$	0.90 ± 0.02	0.90 ± 0.01	0.91	✓
$\text{CVaR}_{0.25}(X)$	0.75 ± 0.02	0.71 ± 0.08	0.5	✗
$\text{CVaR}_{0.5}(X)$	1.12 ± 0.03	1.12 ± 0.03	1.0	✗

Table 2. We benchmark our optimistic Alg. 1 with UCB-VI as OPTALG and our PO Alg. 2 with NPG as POALG against the best Markovian policy for various OCEs. We repeat the experiment 10 times and report 95% confidence intervals for the average performance.

an augmented policy with parameters θ_h . Then, the NPG update is given by

$$\theta_h^{k+1} = \theta_h^k + \eta F_{h,k}^\dagger \nabla_{\theta_h} V_{\text{aug}}^{\pi^k}, \quad (8)$$

where $F_{h,k} = \mathbb{E}_{\pi^k}[\nabla_{\theta} \log \pi_h^k(a_h | s_h, b_h) \nabla_{\theta} \log \pi_h^k(a_h | s_h, b_h)^\top]$ is the Fisher info matrix and $F_{h,k}^\dagger$ is its pseudo-inverse. A special class of policies are softmax policies: $\pi_h^\theta(a | s, b) \propto \exp(\theta_{s,b,a}^h)$ with $\theta_{s,b,a}^h \in \mathbb{R}$. Under softmax parameterization, the NPG update in Eq. (8) is equivalent to soft policy iteration (Kakade, 2001):

$$\pi^{k+1}(a | s, b) \propto \pi^k(a | s, b) \exp(\eta \cdot Q_{\text{aug}}^{\pi^k}(s, b, a)).$$

We now show that NPG in the AugMDP satisfies both conditions in Def. 4.1. First, for local improvement, Agarwal et al. (2021, Lemma 5.2) implies that $\mathbb{E}_{b \sim \text{Unif}(\mathcal{B})}[V_{\text{aug}}^{\pi^{k+1}}(s_1, b) - V_{\text{aug}}^{\pi^k}(s_1, b)] \geq 0$, which uses the fact that the initial budget distribution is $\text{Unif}(\mathcal{B})$. Then, for global convergence, Agarwal et al. (2021, Theorem 5.3) proves that $\text{Reg}_{\text{PO}}(K) \leq \mathcal{O}(H)$, provided we set an appropriate learning rate $\eta = H \log |\mathcal{A}|$. Thus, NPG satisfies Def. 4.1 so we can apply both Thm. 4.2 and Thm. 4.4. To the best of our knowledge, these are the first non-asymptotic guarantees for gradient-based risk-sensitive RL. In App. E.1, we extend this analysis further to smooth policy parameterization and unknown Q_{aug} functions, using compatible function approximation à la Agarwal et al. (2021); Xiao (2022).

5. Simulation Experiments

We describe a numerical simulation to demonstrate the importance of learning history-dependent policies for OCE RL and to empirically evaluate our algorithms. Our code can be found at <https://github.com/kaiwenw/oce-rl>.

Setting up synthetic MDP. The proof-of-concept MDP is shown in Figure 1 and has two states. At s_1 , all actions lead to a random reward $r_1 \sim \text{Ber}(0.5)$ and transits to s_2 . At s_2 , the first action a_1 gives a random reward $r_2 | s_2, a_1 \sim 1.5 \cdot \text{Ber}(0.75)$, while another action a_2 gives a deterministic reward $r_2 | s_2, a_2 = 0.5$. The trajectory ends after s_2 .

The MDP is designed so that the optimal CVaR policy is non-Markovian. Due to the MDP’s simplicity, we can compute the optimal CVaR and the Markovian policies’ CVaR in closed form, which we list in Table 1. Specifically, the optimal action at s_2 depends on the random outcome of the first reward r_1 : if $r_1 = 0$ then it should be risky and pick a_1 ; if $r_1 = 1$ then it should be conservative and pick a_2 . This optimal policy achieves $\text{CVaR}_{0.25}^* = 0.75$, while Markovian policies, which do not react to r_1 , can only achieve $\text{CVaR}_{0.25} = 0.5$. This reinforces our thesis that history-dependent policies are required for optimality.

Experiment with tabular policies. We now apply our meta-algorithms to the synthetic MDP. For our optimistic meta-algorithm (Alg. 1), we use UCB-VI (Azar et al., 2017) as the oracle OPTALG. For our gradient-based meta-algorithm (Alg. 2), we use NPG as the oracle POALG. We evaluate our algorithms and compute the best Markovian policies for six OCEs: two CVaRs, two entropic risks and two mean-variances, where recall CVaR and entropic risks are defined as:

$$\begin{aligned} \text{CVaR}_\tau(X) &:= \max_{b \in \mathbb{R}} \{b - \tau^{-1} \mathbb{E}[(b - X)_+]\}, \\ \text{Entr}_\beta(X) &:= \frac{1}{\beta} \ln \mathbb{E}[\exp(\beta X)]. \end{aligned}$$

The results are in Table 2 and we see that our algorithms are consistently better than the best Markovian policy. We remark that entropic risk is a special case where the optimal policy is Markovian (Fei et al., 2020; 2021); nonetheless, our algorithms still consistently learn the optimal policy despite AugMDP not being necessary for entropic risk.

Experiment with neural network policies. To test the versatility of our PO algorithm (Alg. 2), we also evaluate three deep RL oracles: PPO with forward KL (Schulman et al., 2017), PPO with backward KL (Hsu et al., 2020), and REINFORCE (Sutton et al., 1999). Using neural networks to approximate the policy and Q functions, we train all three oracles to maximize $\text{CVaR}_{0.25}$ in the synthetic MDP.

We plot the learning curves in Figure 2, where we see that Alg. 2 consistently converges to the optimal $\text{CVaR}_{0.25}^*$ value of 0.75 for all three oracles. An interesting trend is that

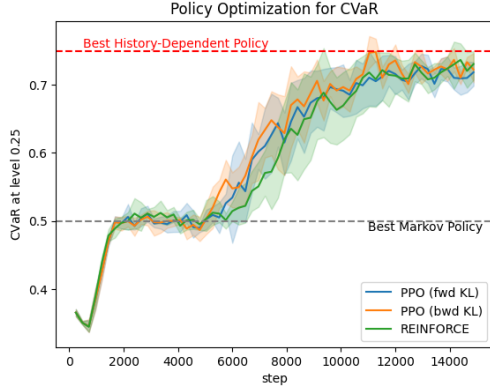


Figure 2. Learning curves for Alg. 2 with three oracles: REINFORCE and PPO with fwd & bwd KL. We repeat runs five times and report 95% confidence intervals for the mean performance.

Alg. 2 tends to first plateau at a Markovian policy at $\sim 2k$ steps, and then eventually converge to the optimal non-Markovian policy at $\sim 8k$ steps, suggesting that Markovian policies are perhaps easier-to-learn local optima in the AugMDP. Nonetheless, our algorithm consistently converges to the optimal policy given enough gradient steps. In sum, our experiments demonstrate the importance of history-dependent policies in risk-sensitive RL and the effectiveness of our AugMDP algorithms. We report all hyperparameters and training details in App. C.

6. Conclusion

In this paper, we proposed two meta-algorithms for RSRL with the static OCE risk. These meta-algorithms provide a general reduction from OCE RL to risk-neutral RL in the augmented MDP framework. First, we proposed an optimistic meta-algorithm (Alg. 1) that generalizes all prior bounds in CVaR RL to the OCE setting when UCB-VI and Rep-UCB are used as the optimistic oracle. Moreover, using GOLF as the optimistic oracle, we proved the first risk-sensitive regret bounds for exogenous block MDPs. Second, we proposed a gradient-based meta-algorithm (Alg. 2) that enjoys both global convergence and local improvement. With this framework, we deduced the first finite-sample bounds for policy gradients in risk-sensitive RL. Finally, we evaluate our algorithms in a proof-of-concept MDP where they consistently learn the optimal policy and outperform the best Markovian policy. A promising direction for future work is to scale our ideas to more complex tasks such as robotics or finetuning LLMs (Ouyang et al., 2022). Given the plurality of values and risks, it is also promising to apply our ideas to design algorithms for risk-sensitive pluralistic alignment (Sorensen et al., 2024; Wang et al., 2024a).

Acknowledgements. This work was supported by a Google PhD fellowship and grants NSF IIS-2154711, NSF CAREER 2339395, IIS-1846210, and DARPA LANCER: LeArning Network CyBERagents.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.
- Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.
- Maurice Allais. Allais paradox. In *Utility and probability*, pages 3–9. Springer, 1990.
- Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.
- Alex Ayoub, Kaiwen Wang, Vincent Liu, Samuel Robertson, James McInerney, Dawen Liang, Nathan Kallus, and Csaba Szepesvári. Switching the loss reduces the cost in batch reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024.
- Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.
- Osbert Bastani, Yecheng Jason Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=yJEUDfzsTX7>.
- Nicole Bäuerle and Alexander Glauner. Minimizing spectral risk measures applied to markov decision processes. *Mathematical Methods of Operations Research*, 94(1): 35–69, 2021.
- Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.
- Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

- 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.
- Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust markov decision processes. *Advances in Neural Information Processing Systems*, 2024.
- Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.
- Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for off-line policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.
- Yu Chen, XiangCheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=0xmExPqFf>.
- Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27, 2014.
- Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.
- Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial intelligence in medicine*, 109:101964, 2020.
- Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.
- Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.
- 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. URL <https://openreview.net/forum?id=Yn0xg-kHNW->.
- Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=RQLLzMCefQu>.
- 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, 2021.
- Hans Föllmer and Alexander Schied. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.
- Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:32639–32652, 2022.
- Jakub Grudzien, Christian A Schroeder De Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *International Conference on Machine Learning*, pages 7825–7844. PMLR, 2022.
- Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7): 356–369, 1972.
- Chloe Ching-Yun Hsu, Celestine Mendler-Dünner, and Moritz Hardt. Revisiting design choices in proximal policy optimization. *arXiv preprint arXiv:2009.10897*, 2020.
- Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank mdps with density features. In *International Conference on Machine Learning*, pages 13710–13752. PMLR, 2023.
- Ehsan Imani, Kai Luedemann, Sam Scholnick-Hughes, Esraa Elelimy, and Martha White. Investigating the histogram loss in regression. *arXiv preprint arXiv:2402.13425*, 2024.
- Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.
- Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.
- Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.
- Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.
- Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.
- Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Yehuda Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=NLpXRrjpa6>.
- 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*, 2023. URL <https://openreview.net/forum?id=-RwZOVybbj>.
- 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*, pages 1774–1782. PMLR, 2024.
- Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, 2022.
- Xiaoteng Ma, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.
- Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.
- HM Markowitz. Portfolio selection, the journal of finance. 7 (1). N, 1:71–91, 1952.
- Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=7sACcaOmGi>.
- Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. In *International Conference on Machine Learning*, pages 36486–36503. PMLR, 2024.
- Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Fabio Pardo, Arash Tavakoli, Vitaly Levnik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning*, pages 4045–4054. PMLR, 2018.
- Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- 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, 2023.
- R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.
- Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125:235–261, 2010.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *International Conference on Machine Learning*, 2024.
- Richard S. Sutton. The reward hypothesis. <http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html>, 2004.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015a.
- Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015b.
- John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.
- Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *ICLR*, 2022. URL <https://openreview.net/forum?id=J4iSIR9fhY0>.
- Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. In *International Conference on Learning Representations*, 2021.
- Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *International Conference on Machine Learning*, 2023a.
- Kaiwen Wang, Junxiong Wang, Yueying Li, Nathan Kallus, Immanuel Trummer, and Wen Sun. Joingym: An efficient query optimization environment for reinforcement learning. *arXiv preprint arXiv:2307.11704*, 2023b.
- Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghu Gupta, Avinava Dubey, et al. Conditional language policy: A general framework for steerable multi-objective finetuning. *Findings of Empirical Methods in Natural Language Processing*, 2024a.
- Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024b. URL <https://openreview.net/forum?id=kZBCFQe1Ej>.
- Kaiwen Wang, Nathan Kallus, and Wen Sun. The central role of the loss function in reinforcement learning. *Statistical Science*, 2025a.
- Kaiwen Wang, Jin Peng Zhou, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kianté Brantley, and Wen Sun. Value-guided search for efficient chain-of-thought reasoning. *arXiv preprint arXiv:2505.17373*, 2025b.
- Zhiyong Wang, Dongruo Zhou, John Lui, and Wen Sun. Model-based rl as a minimalist approach to horizon-free and second-order bounds. *arXiv preprint arXiv:2408.08994*, 2024c.
- Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.
- Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23 (282):1–36, 2022.
- Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=LQIjzPdDt3q>.
- Wenhao Xu, Xuefeng Gao, and Xuedong He. Regret bounds for markov decision processes with recursive optimized certainty equivalents. *International Conference on Machine Learning*, 2023.
- Menahem E Yaari. The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, pages 95–115, 1987.
- Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.
- Yulai Zhao, Wenhao Zhan, Xiaoyan Hu, Ho fung Leung, Farzan Farnia, Wen Sun, and Jason D. Lee. Provably efficient CVAR RL in low-rank MDPs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=9x6yrFAPnx>.

Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q Weinberger, Kianté Brantley, and Wen Sun. *q#*: Provably optimal distributional rl for llm post-training. *arXiv preprint arXiv:2502.20548*, 2025.

Ruiwen Zhou, Minghuan Liu, Kan Ren, Xufang Luo, Weinan Zhang, and Dongsheng Li. Is risk-sensitive reinforcement learning properly resolved? *arXiv preprint arXiv:2307.00547*, 2023.

Appendices

A. List of Notations

Table 3. List of notations used in the paper.

\mathcal{S}, \mathcal{A}	State and action spaces.
H	Time horizon.
$\Delta(\mathcal{S})$	The set of distributions supported by set \mathcal{S} .
Π_{Markov}	Class of Markovian policies that act only based on current state.
Π_{HD}	Class of history-dependent policies.
$\Pi_{\text{aug}}, \Pi_{\text{aug}}^{\text{HD}}$	Class of Markovian and history-dependent policies in the AugMDP.
$Q_{\text{aug}}^{\pi, h}, V_{\text{aug}}^{\pi, h}$	Q and value functions of π in the AugMDP.
$Q_{\text{aug}}^{*, h}, V_{\text{aug}}^{*, h}$	Optimal Q and value functions in the AugMDP.
(π, b)	Run the policy π from an initial state (s_1, b) in the AugMDP.
$Z(\pi)$	The cumulative reward distribution of policy π .
$\text{OCE}_u(\pi)$	Optimized certainty equivalent with utility u for $Z(\pi)$ (see App. B for primer on OCE).
OCE_u^*	Optimal OCE of cumulative rewards by history-dependent policy, <i>i.e.</i> , $\max_{\pi \in \Pi_{\text{HD}}} \text{OCE}_u(\pi)$.
$\text{RLB}^{(k)}$	The risk lower bound at round k for policy optimization algorithms (defined in Eq. (6)).
V_u^{\max}	Scale of utility u defined as $V_u^{\max} := \max_{c \in [-1, 1]} u(c) $.

B. Primer on Optimized Certainty Equivalents (OCE)

In section is a short primer on optimized certainty equivalents (OCE), which capture many important risk measures including conditional value-at-risk (CVaR) (Rockafellar and Uryasev, 2000), entropic risk (Föllmer and Schied, 2011) and Markowitz’s mean-variance (Markowitz, 1952). For a utility function $u : \mathbb{R} \rightarrow [-\infty, \infty)$, the OCE of a random variable X is defined as:

$$\text{OCE}_u(X) := \max_{b \in \mathbb{R}} \{b + \mathbb{E}[u(X - b)]\}.$$

In the paper, we focused on the case where X is the cumulative reward distribution of some policy π , and the goal is to learn the policy that maximizes the OCE. While the OCE is well-defined for any utility function, there are common regularity conditions proposed by Ben-Tal and Teboulle (2007) that ensure the OCE is well-behaved. These conditions are:

- [R1] u is proper meaning that its domain is non-empty, *i.e.*, $\text{dom}(u) := \{t \in \mathbb{R} : u(t) > -\infty\} \neq \emptyset$;
- [R2] u is closed meaning that its hypograph $\text{hyp}(u) := \{(t, r) \in \mathbb{R} \times \mathbb{R} : r \leq u(t)\}$ is a closed set;
- [R3] u is concave meaning that its hypograph is convex;
- [R4] u is non-decreasing;
- [R5] $u(0) = 0$ and $1 \in \partial u(0)$, where $\partial u(x) = \{g : \forall z \in \mathbb{R}, g \leq (u(z) - u(x))/(z - x)\}$ is the subgradient.

Ben-Tal and Teboulle (2007, Theorem 2.1) showed that, if u satisfies R1-R5, then OCE_u enjoys the following properties:

- [P1] Translation invariance: $\text{OCE}_u(X + c) = \text{OCE}_u(X) + c$ for constant $c \in \mathbb{R}$;
- [P2] Monotonicity: if $X(\omega) \leq Y(\omega), \forall \omega \in \Omega$, then $\text{OCE}_u(X) \leq \text{OCE}_u(Y)$;
- [P3] Concavity: for any $X, Y \in L_\infty, \lambda \in [0, 1]$, we have $\text{OCE}_u(\lambda X + (1 - \lambda)Y) \geq \lambda \text{OCE}_u(X) + (1 - \lambda) \text{OCE}_u(Y)$;
- [P4] Consistency: $\text{OCE}_u(0) = 0$.

Translation invariance (P1) and consistency (P4) are intuitive since a deterministic payoff of c should have a value of c . Monotonicity (P2) states that if X is dominated by Y , then the risk of X should be no greater than the risk of Y . Concavity (P3) implies that $-\text{OCE}_u$ is a convex risk measure. However, we remark that these properties do not imply that OCE_u is a coherent risk measure (Artzner et al., 1999), which would require sub-additivity and positive homogeneity. Indeed, convexity is a relaxation of sub-additivity and positive homogeneity, which are not always satisfied by OCE_u even if u satisfies R1-R5. For example, the entropic risk is an OCE with an exponential utility function satisfying R1-R5, but it is not a coherent risk measure.

In Table 4 below, we provide important examples of OCE, which includes (1) expectation, (2) entropic risk, (3) CVaR and (4) Markowitz's mean-variance. Recall that CVaR and Entropic risk are defined by:

$$\text{CVaR}_\tau(X) := \max_{b \in \mathbb{R}} \{b - \tau^{-1} \mathbb{E}[(b - X)_+]\}, \quad \text{Entr}_\beta(X) := \frac{1}{\beta} \log \mathbb{E} \exp(\beta X).$$

Risk Name	Parameter	Utility function u	V_u^{\max}
Mean ($\mathbb{E}X$)	None	$u(t) = t$	1
Cond. Value-at-Risk ($\text{CVaR}_\tau(X)$)	$\tau \in (0, 1]$	$u(t) = -\tau^{-1}(-t)_+$	τ^{-1}
Entropic Risk ($\text{Entr}_\beta(X)$)	$\beta \in (-\infty, 0)$	$u(t) = \frac{1}{\beta} (\exp(\beta t) - 1)$	$\frac{1}{ \beta } (\exp(\beta) - 1)$
Mean-Variance ($\mathbb{E}X - c \cdot \text{Var}(X)$)	$c > 0$	$u(t) = t - ct^2$ if $t \leq 1/(2c)$ else $u(t) = 1/(4c)$	$1 + c$
Mean-CVaR $_\tau$ ($\kappa_1 \mathbb{E}X + (1 - \kappa_1) \cdot \text{CVaR}_\tau(X)$)	$\kappa_1 \in [0, 1]$	$u(t) = \kappa_1(t)_+ - \kappa_2(-t)_+$, where $\kappa_2 = \tau^{-1}(1 - \kappa_1) + \kappa_1$.	κ_2

Table 4. Examples of OCE risk measures with their corresponding u and V_u^{\max} .

Table 4 also lists each OCE's V_u^{\max} which recall is defined as $V_u^{\max} := \max_{c \in [-1, 1]} |u(c)|$ and roughly measures the statistical hardness of learning OCE_u relative to \mathbb{E} (which has $V_u^{\max} = 1$). We remark that OCE is a Lipschitz risk measures (Liang and Luo, 2024; Chen et al., 2024) with Lipschitz constant V_u^{\max} , and so V_u^{\max} can also be interpreted in this way.

The first four rows of Table 4 are well-known risk measures and are known to be OCEs. The final row Mean-CVaR $_\tau$ is a lesser known example of OCE, but is perhaps more relevant in practice since it captures both the average and tail risks together. We now provide a proof that Mean-CVaR $_\tau$ is an OCE with the piecewise linear utility.

Theorem B.1. *Let $0 \leq \kappa_1 < 1 < \kappa_2$ and consider the piece-wise linear utility:*

$$u_{\kappa_1, \kappa_2}(t) = \kappa_1(t)_+ - \kappa_2(-t)_+.$$

where we denote $(y)_+ = \max(y, 0)$. Then, with $\tau = \frac{1-\kappa_1}{\kappa_2-\kappa_1}$, we have that

$$\text{OCE}_{u_{\kappa_1, \kappa_2}}(X) = \kappa_1 \mathbb{E}X + (1 - \kappa_1) \text{CVaR}_\tau(X).$$

Proof. By Ben-Tal and Teboulle (2007, Example 2.3),

$$\text{OCE}_{u_{\kappa_1, \kappa_2}}(X) = \max_{b \in [0, 1]} \{b + \kappa_1 \mathbb{E}[(X - b)_+] - \kappa_2 \mathbb{E}[(b - X)_+]\}. \quad (9)$$

Moreover, the optimal dual variable is $b^* = F^\dagger((1 - \kappa_1)/(\kappa_2 - \kappa_1))$, where $F^\dagger(t) = \inf\{x \mid F(x) \geq t\}$ is the quantile function and F is the CDF of X . Expanding Eq. (9) and using the fact that $(t)_+ - (-t)_+ = t$, we get

$$\begin{aligned} \text{OCE}_{u_{\kappa_1, \kappa_2}}(X) &= \kappa_1 \mathbb{E}X + \max_{b \in [0, 1]} \{(1 - \kappa_1)b - (\kappa_2 - \kappa_1) \mathbb{E}[(b - X)_+]\} \\ &= \kappa_1 \mathbb{E}X + (1 - \kappa_1) \max_{b \in [0, 1]} \{b - \frac{\kappa_2 - \kappa_1}{1 - \kappa_1} \mathbb{E}[(b - X)_+]\} \\ &= \kappa_1 \mathbb{E}X + (1 - \kappa_1) \text{CVaR}_\tau(X), \end{aligned}$$

where $\tau = \frac{1-\kappa_1}{\kappa_2-\kappa_1}$. This completes the proof. \square

C. More Details on Experimental Setup

C.1. Network Architecture and Input Features

We parameterize both the augmented Q -values and the softmax policy using multilayer perceptrons (MLPs) with two hidden layers of dimension 64. For the policy network, we one-hot encode the budget b_h before passing it into the network. This transformation does not introduce approximation error since rewards are discrete in the Markov Decision Process (MDP). We observed that when b_h was inputted directly as a real number, the training of the policy was less stable and often diverged. In contrast, the Q -network demonstrated greater robustness to how b_h is featurized, likely due to the augmented value function being Lipschitz and monotonic in b_h . Unlike the policy network, the Q -network receives b_h as a direct input without encoding. Finally, since the task is finite-horizon, we incorporate the time step h as part of the state representation, following [Pardo et al. \(2018\)](#).

C.2. Regularization

Log-barrier regularization played a crucial role in stabilizing and preventing the optimization from getting stuck due to vanishing gradients. Indeed in softmax policies, the gradients tend to vanish as the policy becomes more deterministic. To mitigate this issue, we applied a regularization weight of 0.1 which helped to prevent the policy from becoming too deterministic. We remark that policy gradients with log-barrier regularization also have theoretical guarantees [Agarwal et al. \(2021, Theorem 12\)](#), making this a valid POALG for our PO algorithm.

C.3. Hyperparameter Settings

Component	Value/Description
Policy Network	Softmax policy with MLP with two hidden layers of dimension 64
Value Network	MLP with two hidden layers of dimension 64
Budget Encoding (Policy)	One-hot
Budget Encoding (Q-network)	Direct input
Optimizer	Adam with $\beta_1 = 0.9, \beta_2 = 0.999$
Batch Size	256
Learning Rate	5×10^{-3}
GAE λ	0.95
PPO KL weight	0.1
Regularization Weight	0.1

Table 5. Hyperparameter settings used in our experiments.

D. Proofs for Optimistic Meta-Algorithm

In this section, we study three illustrative examples of optimistic oracles that satisfy Def. 3.1 in the AugMDP. In particular, all these oracles can be used with our meta-algorithm Alg. 1 to form an optimistic OCE algorithm with the regret / PAC bounds in Thm. 3.2. The three oracles we consider are UCB-VI (Azar et al., 2017), Rep-UCB (Uehara et al., 2022) and GOLF (Jin et al., 2021).

First, UCB-VI is a model-based algorithm based on optimistic value iteration with exploration bonuses and achieves minimax-optimal regret in tabular MDPs (Azar et al., 2017). Second, Rep-UCB is a model-based algorithm based on elliptical exploration bonuses and achieves PAC bounds in low-rank MDPs (Uehara et al., 2022). While the AugMDP may initially seem more complex, since it is no longer tabular or low-rank, the regret bounds of these algorithms actually easily transfers to the AugMDP when the underlying MDP (*i.e.*, the MDP being augmented) is tabular or low-rank. The intuition is because the AugMDP does not introduce any new unknowns to the model, as the augmented transition and reward functions are known and deterministic, as highlighted in Sec. 2.1. Hence, the model that needs to be learned is the same in the original MDP and the AugMDP, and the prior regret bounds naturally translate to the AugMDP, which we formalize in the following subsections. In particular, we will see that prior works such as (Wang et al., 2023a; Zhao et al., 2024) have already implicitly proven that UCB-VI and Rep-UCB satisfies Def. 3.1 in the CVaR AugMDP.

The third optimistic oracle we consider is GOLF, a model-free algorithm based on version space optimism and can achieve regret bounds in exogenous block MDPs (Xie et al., 2023). Unlike model-based algorithm, extending GOLF to the AugMDP requires more care since the value functions being learned also takes in the augmented state b as input; that is, the value function class is more complex in the AugMDP. To handle this added complexity, we posit a discreteness assumption in Assump. 3.5. Under this premise, we prove the previous GOLF analysis based on coverability can be extended to the AugMDP. Consequently, we derive the first OCE regret bounds in exogenous block MDPs (Efroni et al., 2022).

Finally, we conclude this section by showing how second-order bounds for the oracle can lead to tight and optimal regret for CVaR_τ in tabular MDPs, recovering the main result of (Wang et al., 2023a). In particular, we observe that the key idea of the complex argument of (Wang et al., 2023a) is simply access to an oracle with second-order regret. We believe this observation can lead to tighter RSRL bounds, especially since second-order bounds have recently been possible in much more general MDPs via distributional RL (Wang et al., 2024b; 2025a; 2024c).

Algorithm 3 Optimistic Oracle: UCB-VI (Azar et al., 2017)

- 1: **Input:** Number of rounds K , failure probability δ .
- 2: **for** round $k = 1, 2, \dots, K$ **do**
- 3: Compute counts and empirical transition estimate,

$$N_k(s, a, s') = \sum_{h=1}^H \sum_{i=1}^{k-1} \mathbb{I}[(s_{h,i}, a_{h,i}, s_{h+1,i}) = (s, a, s')],$$

$$N_k(s, a) = 1 \vee \sum_{s' \in \mathcal{S}} N_k(s, a, s'), \quad \hat{P}_k(s' | s, a) = \frac{N_k(s, a, s')}{N_k(s, a)},$$

- 4: For all $s \in \mathcal{S}, b \in [0, 1]$, set $\hat{V}_{H+1,k}(s, b) = u(-b)$.
- 5: **for** $h = H, H-1, \dots, 1$ **do**
- 6: For all s, b, a ,

$$\hat{Q}_{h,k}(s, b, a) = \hat{P}_k(s, a)^\top \mathbb{E}_{r_h \sim R(s,a)} [\hat{V}_{h+1,k}(\cdot, b - r_h)] + \sqrt{\frac{\log(HSAK/\delta)}{N_k(s, a)}},$$

$$\pi_h^k(s, b) = \arg \max_a \hat{Q}_{h,k}(s, b, a), \quad \hat{V}_{h,k}(s, b) = \min \left\{ \hat{Q}_{h,k}(s, b, \pi_h^k(s, b)), V^{\max} \right\}.$$

- 7: **end for**
 - 8: Output optimistic value function $\hat{V}_{1,k}(s_1, \cdot)$.
 - 9: Receive adversarial \hat{b}_k .
 - 10: Collect $\{(s_{h,k}, a_{h,k}, r_{h,k})\}_{h \in [H]}$ by executing π^k starting from (s_1, \hat{b}_k) in AugMDP.
 - 11: **end for**
-

D.1. Example 1: UCB-VI

UCB-VI (Azar et al., 2017) is a model-based algorithm for tabular MDPs based on value iteration with exploration bonuses. Alg. 3 formalizes the UCB-VI algorithm in the AugMDP. The following theorem recovers the results from Wang et al. (2023a, Theorem 5.2), who focused on the more restricted CVaR RL setting.

Theorem D.1. For any $\delta \in (0, 1)$, running Alg. 3 enjoys the following w.p. $1 - \delta$:

1. (Optimism) $\hat{V}_{1,k}(s_1, b_1) \geq V_{\text{aug}}^{*,1}(s_1, b_1)$ for all $k \in [K], b_1 \in [0, 1]$;
2. (Regret) The regret in the AugMDP is at most,

$$\sum_{k=1}^K \hat{V}_{1,k}(s_1, \hat{b}_k) - V_{\text{aug}}^{\pi^k,1}(s_1, \hat{b}_k) \leq \tilde{\mathcal{O}}(V_u^{\max}(\sqrt{SAHK \log(1/\delta)} + S^2AH)).$$

where $\tilde{\mathcal{O}}(\cdot)$ ignores terms logarithmic in S, A, H, K .

Proof Sketch. Thm. D.1 can be proved by following the same argument in (Wang et al., 2023a, Appendix G.3), except that the CVaR utility $\tau^{-1} \min(t, 0)$ is replaced by any generic OCE utility u . In particular, optimism is ensured by Lemma G.3 and the Equation Bon★. The regret bound is ensured by the proof of Theorem 5.2 on Page 35. Thus, by applying the argument in (Wang et al., 2023a), we can show that UCB-VI satisfies the oracle conditions in Def. 3.1. \square

Therefore, Thm. D.1 shows that UCB-VI satisfies the oracle conditions Def. 3.1 in the AugMDP with $\varepsilon_k^{\text{opt}} = 0$ and $\text{Reg}_{\text{opt}}(K) \leq \tilde{\mathcal{O}}(\sqrt{SAHK \log(1/\delta)} + S^2AH)$.

Algorithm 4 Optimistic Oracle: Rep-UCB (Uehara et al., 2022)

- 1: **Input:** No. episodes K , Model Classes (Φ, Υ) .
 - 2: Initialize $\mathcal{D}_h^{(0)} = \tilde{\mathcal{D}}_h^{(0)} = \emptyset$ for all $h \in [H]$.
 - 3: **for** round $k = 1, 2, \dots, K$ **do**
 - 4: Learn the model via maximum likelihood estimation (MLE): for all $h \in [H]$:

$$\phi_h^{(k)}, \mu_h^{(k)} := \arg \max_{\phi \in \Phi, \mu \in \Upsilon} \sum_{s, a, s' \in \mathcal{D}_h^{(k-1)} \cup \tilde{\mathcal{D}}_h^{(k-1)}} \log \phi(s_h, i, a_{h,i})^\top \mu(s_{h+1}, i)$$
 - 5: Define the empirical covariance $\Sigma_{h,k} = \sum_{s, a \in \mathcal{D}_h^{(k-1)}} \phi_h^{(k)}(s, a) (\phi_h^{(k)}(s, a))^\top + \lambda_k I$.
 - 6: Define the bonus, $b_{h,k}(s, a) = \min(\alpha_k \|\phi_h^{(k)}(s, a)\|_{\Sigma_{h,k}^{-1}}, 2)$ if $h < H$. Otherwise $b_{H,k} = 0$.
 - 7: Let $\hat{Q}_{h,k}(s, b, a)$ denote quality function of the AugMDP with reward $r_{\text{aug}}^{h,k}(s, b, a) + b_{h,k}(s, a)$ and transitions $\mathcal{P}_h^{(k)}(s' | s, a) = \phi_h^{(k)}(s, a)^\top \mu_h^{(k)}(s')$. Let $\hat{V}_{h,k}(s, b) := \max_a \hat{Q}_{h,k}(s, b, a)$ and $\pi^k := \arg \max_{a \in \mathcal{A}} \hat{Q}_{h,k}(s, b, a)$. This can be achieved via a planning oracle (Zhao et al., 2024).
 - 8: Output optimistic value function $\hat{V}_{1,k}(s_1, \cdot)$.
 - 9: Receive adversarial \hat{b}_k .
 - 10: For each $h \in [H]$, execute π^k starting from (s_1, \hat{b}_k) in AugMDP until s_h, b_h and take two uniform actions. That is, $a_h \sim \text{Unif}(\mathcal{A})$, observe $s_{h+1} \sim P_h^*(s_h, a_h)$, and take $a_{h+1} \sim \text{Unif}(\mathcal{A})$, observe $s_{h+2} \sim P_{h+1}^*(s_{h+1}, a_{h+1})$. Then, add to datasets: $\mathcal{D}_h^{(k)} \leftarrow \mathcal{D}_h^{(k-1)} \cup \{(s_h, a_h, s_{h+1})\}$ and $\tilde{\mathcal{D}}_{h+1}^{(k)} \leftarrow \tilde{\mathcal{D}}_{h+1}^{(k-1)} \cup \{(s_{h+1}, a_{h+1}, s_{h+2})\}$.
 - 11: **end for**
-

D.2. Example 2: Rep-UCB

Rep-UCB (Uehara et al., 2022) is a model-based algorithm for low-rank MDPs based on elliptical bonuses. We first recall the definition of low-rank MDP (Agarwal et al., 2020).

Definition D.2 (Low-Rank MDP). An MDP is has rank d if its transitions have a low-rank decomposition $P_h(s' | s, a) = \phi_h^*(s, a)^\top \mu_h^*(s')$ where $\phi_h^*(s, a), \mu_h^*(s') \in \mathbb{R}^d$ are unknown features that satisfy $\sup_{s,a} \|\phi_h^*(s, a)\|_2 \leq 1$ and $\|\int g(s') d\mu_h^*(s')\| \leq \|g\|_\infty \sqrt{d}$ for all $g : \mathcal{S} \rightarrow \mathbb{R}$.

Rep-UCB utilizes a model class (Φ, Υ) to learn the low-rank transition $\phi_h^*(s, a)^\top \mu_h^*(s')$.

Assumption D.3 (Realizability). $\phi_h^* \in \Phi$ and $\mu_h^* \in \Upsilon$ for all $h \in [H]$.

Rep-UCB as an oracle in the AugMDP is presented in Alg. 4. Our theory recovers the results of Zhao et al. (2024), who focused on the more restricted CVaR setting.

Theorem D.4. Under Def. D.2 and Assump. D.3, for any $\delta \in (0, 1)$, Alg. 4 enjoys the following w.p. $1 - \delta$:

1. (Optimism) For all $k \in [K]$ and b_1 , we have $V_{\text{aug}}^{*,1}(s_1, b_1) - \hat{V}_{1,k}(s_1, b_1) \leq \mathcal{O}(H\sqrt{AL/k})$ where $L = \log(|\Phi||\Upsilon|HK/\delta)$;
2. (Total sub-optimality gap)

$$\sum_{k=1}^K \hat{V}_{1,k}(s_1, \hat{b}_k) - V_{\text{aug}}^{\pi^k,1}(s_1, \hat{b}_k) \leq \mathcal{O}(V_u^{\max}(H^3 A d^2 \sqrt{KL})).$$

Proof Sketch. Thm. D.4 can be proved by following the same argument in Zhao et al. (2024), except that the CVaR utility $\tau^{-1} \min(t, 0)$ is replaced by any generic OCE utility u . In particular, optimism is ensured by Lemma C.2. The total sub-optimality gap bound was proven in Lemma C.1, where any $\tau^{-1}(t - R)^+$ can be replaced by $u(R - t)$. Thus, by applying the argument in (Zhao et al., 2024), we can show that Rep-UCB satisfies the oracle conditions in Def. 3.1. \square

Therefore, Thm. D.4 shows that Rep-UCB satisfies the oracle conditions Def. 3.1 in the AugMDP with $\text{Reg}_{\text{Opt}}(K) \leq \mathcal{O}(H^3 A d^2 \sqrt{KL})$ and $\sum_k \varepsilon_k^{\text{opt}} \leq \mathcal{O}(H\sqrt{AKL})$.

Algorithm 5 Optimistic Oracle: GOLF (Jin et al., 2021)

- 1: **Input:** No. rounds K , Function class $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_H$, threshold β .
- 2: Initialize $\mathcal{D}_h^{(0)} = \emptyset$ for all h , and $\mathcal{F}^{(0)} = \mathcal{F}$.
- 3: **for** round $k = 1, 2, \dots, K$ **do**
- 4: For each $b_1 \in [0, 1]$, define $\hat{Q}_k(\cdot; b_1) := \arg \max_{f \in \mathcal{F}^{(k-1)}} \max_{a \in \mathcal{A}} f_1(s_1, b_1, a)$.
- 5: Output optimistic value functions $\hat{V}_k(s, b; b_1) = \max_{a \in \mathcal{A}} \hat{Q}_k(s, b, a; b_1)$.
- 6: Receive adversarial \hat{b}_k .
- 7: Define policy $\pi_h^k(a | s, b) := \arg \max_{a \in \mathcal{A}} \hat{Q}_{h,k}(s, b, a; \hat{b}_k)$.
- 8: Execute π^k from \hat{b}_k in the AugMDP and collect a trajectory $(s_h^{(k)}, b_h^{(k)}, a_h^{(k)}, r_h^{(k)})_{h \in [H], k \in [K]}$.
- 9: Update dataset $\mathcal{D}_h^{(k)} = \mathcal{D}_h^{(k-1)} \cup \{(s_h^{(k)}, b_h^{(k)}, a_h^{(k)}, r_h^{(k)}, s_{h+1}^{(k)}, b_{h+1}^{(k)})\}$.
- 10: Compute confidence set:

$$\mathcal{F}^{(k)} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(k)}(f_h, f_{h+1}) \leq \min_{g \in \mathcal{F}_h} \mathcal{L}_h^{(k)}(g, f_{h+1}) + \beta, \forall h \in [H] \right\},$$

$$\text{where } \mathcal{L}_h^{(k)}(g, f') = \sum_{i \leq k} (g(s_h^{(i)}, b_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \max_{a \in \mathcal{A}} f'(s_{h+1}^{(i)}, b_{h+1}^{(i)}, a))^2$$

11: **end for**

D.3. Example 3: GOLF

GOLF (Jin et al., 2021) is a model-free algorithm that establishes optimism by optimizing over a version space of candidate Q -functions. In this section, we show that GOLF's regret in the AugMDP can be bounded under a discreteness premise on the cumulative rewards (cf. Assump. 3.5).

We consider a function class $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_H$ with elements $f = (f_1, \dots, f_H) \in \mathcal{F}$ such that $f_h : \mathcal{S} \times [0, 1] \times \mathcal{A} \rightarrow [-V_u^{\max}, V_u^{\max}]$. As convention, we set $f_{H+1}(s, b, a) = u(-b)$.

GOLF constructs the version space by selecting a subset of \mathcal{F} such that the TD-error is small for all h . This is formalized in Alg. 5 with the squared TD-error. To ensure the regression at each step can succeed, GOLF requires Bellman completeness (Jin et al., 2021; Xie et al., 2023). To state Bellman completeness, we first define the Bellman optimality operator in the AugMDP: for a function $f : \mathcal{S} \times [0, 1] \times \mathcal{A} \rightarrow \mathbb{R}$ and any $h < H$:

$$\mathcal{T}_{\text{aug}}^{*,h} f(s_h, b_h, a_h) = \mathbb{E}_{s_{h+1} \sim P_h(s_h, a_h), r_h \sim R_h(s_h, a_h)} [\max_{a' \in \mathcal{A}} f(s_{h+1}, b_h - r_h, a')],$$

$$\mathcal{T}_{\text{aug}}^{*,H} f_{H+1}(s_H, b_H, a_H) = \mathbb{E}_{r_H \sim R_H(s_H, a_H)} [u(r_H - b_H)].$$

Assumption D.5 (AugMDP Bellman Completeness). $\mathcal{T}_{\text{aug}}^{*,h} f_{h+1} \in \mathcal{F}_h$ for all $f_{h+1} \in \mathcal{F}_{h+1}, h \in [H]$.

To prove regret bounds for exogenous block MDPs, we recall the definition of coverability from Xie et al. (2023). Coverability is a complexity measure that captures the minimum possible concentrability coefficient in the MDP:

$$\text{Cov} := \min_{\mu_h \in \Delta(\mathcal{S} \times \mathcal{A})} \max_{\pi \in \Pi_{\text{Markov}}, h \in [H]} \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty,$$

where Π_{Markov} is the set of policies in the original MDP. We extend this notion to the AugMDP:

$$\text{Cov}_{\text{aug}} := \min_{\mu_h \in \Delta(\mathcal{S} \times [0, 1] \times \mathcal{A})} \max_{\pi \in \Pi_{\text{aug}}, h \in [H], b_1 \in [0, 1]} \left\| \frac{d_h^{\pi, b_1}}{\mu_h} \right\|_\infty.$$

We now state the AugMDP regret bound for GOLF.

Theorem D.6. Under Assump. D.5, for any $\delta \in (0, 1)$, running Alg. 5 with $\beta = \Theta(\log(KH|\mathcal{F}|/\delta))$ enjoys the following w.p. $1 - \delta$:

1. (Optimism) $\hat{V}_{1,k}(s_1, b_1; b_1) \geq V_{\text{aug}}^{*,1}(s_1, b_1)$ for all $k \in [K], b_1 \in [0, 1]$;

2. (Regret bound)

$$\sum_{k=1}^K \widehat{V}_{1,k}(s_1, \widehat{b}_k; \widehat{b}_k) - V_{\text{aug}}^{\pi^k, 1}(s_1, \widehat{b}_k) \leq \mathcal{O}(V_u^{\max} H \sqrt{\text{Cov}_{\text{aug}} \beta K \log(K)}).$$

Proof Sketch. Thm. D.6 is a direct consequence of Jin et al. (2021); Xie et al. (2023), since all that we have done is rewritten everything (algorithms, assumptions and theorems) in the AugMDP notation. For example, optimism is proved by Jin et al. (2021, Section 4.2, Page 9). Moreover, the regret bound is proved by Xie et al. (2023, Theorem 1). \square

Therefore, Thm. D.6 shows that GOLF satisfies the oracle conditions Def. 3.1 in the AugMDP with $\varepsilon_k^{\text{opt}} = 0$ and $\text{Reg}_{\text{Opt}}(K) \leq (H \sqrt{\text{Cov}_{\text{aug}} \beta K \log(K)})$.

The main question now is whether we can bound Cov_{aug} by the original coverability Cov , which we know is bounded for low-rank MDPs (Def. D.2) and exogenous block MDPs (Def. 3.3). We now show that if cumulative rewards live in a discrete set \mathcal{B} (cf. Assump. 3.5), we can bound $\text{Cov}_{\text{aug}} \leq |\mathcal{B}| \text{Cov}$ by a simple importance sampling argument.

Lemma D.7. *Under Assump. 3.5, we have*

$$\text{Cov}_{\text{aug}} \leq |\mathcal{B}| \text{Cov}.$$

Proof.

$$\begin{aligned} \text{Cov}_{\text{aug}} &\stackrel{(i)}{=} \max_{h \in [H]} \sum_{s, b, a \in \mathcal{S} \times [0, 1] \times \mathcal{A}} \max_{\pi \in \Pi_{\text{aug}}, b_1 \in [0, 1]} d_h^{\pi, b_1}(s, b, a) \\ &\stackrel{(ii)}{\leq} \max_{h \in [H]} \sum_{s, b, a \in \mathcal{S} \times [0, 1] \times \mathcal{A}} \max_{\pi \in \Pi_{\text{aug}}, b_1 \in [0, 1]} d_h^{\pi, b_1}(s, a) \\ &\stackrel{(iii)}{\leq} \max_{h \in [H]} \sum_{s, b, a \in \mathcal{S} \times [0, 1] \times \mathcal{A}} \max_{\pi \in \Pi_{\text{HD}}} d_h^{\pi}(s, a) \\ &\stackrel{(iv)}{=} \max_{h \in [H]} \sum_{s, b, a \in \mathcal{S} \times [0, 1] \times \mathcal{A}} \max_{\pi \in \Pi_{\text{Markov}}} d_h^{\pi}(s, a) \\ &\stackrel{(v)}{=} |\mathcal{B}| \max_{h \in [H]} \sum_{s, a \in \mathcal{S} \times \mathcal{A}} \max_{\pi \in \Pi_{\text{Markov}}} d_h^{\pi}(s, a) \\ &\stackrel{(vi)}{=} |\mathcal{B}| \text{Cov}, \end{aligned}$$

where (i) is by coverability's equivalence to cumulative reachability (Xie et al., 2023, Lemma 3); (ii) is by $d_h^{\pi, b_1}(s, b, a) \leq d_h^{\pi, b_1}(s, a)$ since \mathcal{B} is discrete; (iii) is since π, b_1 can be viewed as a policy in the original MDP; (iv) is by the Markov optimality theorem for risk-neutral RL: $\max_{\pi \in \Pi_{\text{HD}}} d_h^{\pi}(s', a')$ is equivalent to standard RL with the reward $r_h(s, a) = \mathbb{I}[(s, a) = (s', a')]$ and zero otherwise, and we know that Markovian policies are optimal for risk-neutral RL; (v) is by discreteness of \mathcal{B} to collect common terms; and (vi) is again by Xie et al. (2023, Lemma 3) in the original MDP. \square

We can now state our OCE regret guarantees for low-rank MDPs and exogenous block MDPs. Note these are the first risk-sensitive regret bounds for these rich-observation MDPs.

Theorem D.8. *In an exogenous block MDP (Def. 3.3), under Assump. 3.5, we have*

$$\text{Cov}_{\text{aug}} \leq |\mathcal{B}| |\mathcal{Z}^{\text{en}}| |\mathcal{A}|.$$

Therefore, under Assump. D.5, running Alg. 1 with the GOLF oracle (cf. Alg. 5) enjoys the regret bound:

$$\sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k, \widehat{b}_k) \leq \mathcal{O}(V_u^{\max} H \sqrt{|\mathcal{B}| |\mathcal{Z}^{\text{en}}| |\mathcal{A}| \beta K \log(K)}).$$

Proof. To prove the first statement, recall that $\text{Cov} \leq |\mathcal{Z}^{\text{en}}| |\mathcal{A}|$, as was proved by Xie et al. (2023, Proposition 5). Thus, Lemma D.7 implies that $\text{Cov}_{\text{aug}} \leq |\mathcal{B}| \cdot |\mathcal{Z}^{\text{en}}| |\mathcal{A}|$. To prove the second statement, simply combine our meta-algorithm guarantee in Thm. 3.2 with the GOLF oracle guarantee in Thm. D.6. \square

Theorem D.9. *In a low-rank MDP with rank d (Def. D.2), under Assump. 3.5, we have*

$$\text{Cov}_{\text{aug}} \leq |\mathcal{B}| d |\mathcal{A}|.$$

Therefore, under Assump. D.5, running Alg. 1 with the GOLF oracle (cf. Alg. 5) enjoys the regret bound:

$$\sum_{k=1}^K \text{OCE}_u^* - \text{OCE}_u(\pi^k, \widehat{b}_k) \leq \mathcal{O}(V_u^{\max} H \sqrt{|\mathcal{B}| d |\mathcal{A}| \beta K \log(K)}).$$

Proof. To prove the first statement, recall that $\text{Cov} \leq d|\mathcal{A}|$, as was proved by Huang et al. (2023, Proposition 3). Thus, Lemma D.7 implies that $\text{Cov}_{\text{aug}} \leq |\mathcal{B}| \cdot d|\mathcal{A}|$. To prove the second statement, simply combine our meta-algorithm guarantee in Thm. 3.2 with the GOLF oracle guarantee in Thm. D.6. \square

Note the above is the first risk-sensitive regret bound for low-rank MDPs. The previous Rep-UCB result ((Zhao et al., 2024) or Thm. D.4) is technically not a regret bound since the data collection policy takes uniform exploratory actions, *i.e.*, Rep-UCB can only yield PAC bounds.

D.4. Tight bounds for CVaR via an oracle with second-order regret

Recall that Thm. 3.2 combined with UCB-VI (Thm. D.1) gave an OCE regret bound of $\mathcal{O}(V_u^{\max} \sqrt{SAHK})$. Specializing to CVaR, we have $\mathcal{O}(\tau^{-1} \sqrt{SAHK})$. A weakness in this bound is that it is not tight in τ , since the minimax-optimal rate for CVaR RL actually scales with $\tau^{-1/2}$ (Wang et al., 2023a, Theorem 3.1).

In this section, we address this issue by assuming that the oracle has a second-order regret. Second-order regret, *a.k.a.* variance-dependent regret, is an instance-dependent regret bound that scales with the variance of returns throughout K episodes, and is strictly stronger than the standard \sqrt{K} minimax-optimal regret. Specifically, an oracle with second-order regret takes the following form.

Assumption D.10 (Second-Order Oracle). $\text{Reg}_{\text{Opt}}(K) \leq \sqrt{C_1 \sum_k \text{Var}(Z(\pi^k, \hat{b}_k))} + C_2$ for constants C_1, C_2 .

For example, optimistic model-based algorithms such as UCB-VI with Bernstein bonus (Azar et al., 2017; Zanette and Brunskill, 2019) can achieve second-order bounds in tabular MDPs. Indeed, this property of UCB-VI was used to prove the tight $\tau^{-1/2}$ rate for CVaR RL in Wang et al. (2023a); our following result generalizes this argument. Recently, distributional RL has been used to also prove second-order bounds in much more general MDPs, such as low-rank MDPs (Wang et al., 2024b; 2025a).

We also assume a continuous returns assumption of (Wang et al., 2023a).

Assumption D.11 (Continuously Distributed Returns). For all policies $\pi \in \Pi_{\text{HD}}$, the cumulative reward $Z(\pi)$ is continuously distributed with density lower-bounded by p_{\min} .

The following result has a sharp $\tau^{-1/2}$ dependence and generalizes the minimax-optimal results of (Wang et al., 2023a).

Theorem D.12 ($\tau^{-1/2}$ -regret for CVaR). *Under Assumps. D.10 and D.11, Alg. 1 enjoys*

$$\text{Reg}_{\text{CVaR}_\tau}(K) \leq 4\tau^{-1/2} \sqrt{C_1 K} + 4\tau^{-1}(C_1 p_{\min}^{-1} + C_2).$$

We note that the density lower bound p_{\min} only scales a lower-order term, which is independent of K . The proof uses the fact that the AugMDP return variance is at most τ under the true initial budget $b_k^* = \arg \max_{b \in [0,1]} \{b + V_{\text{aug}}^{\pi^k}(s_1, b)\}$. Then, under Assump. D.11, the approximation $(\hat{b}_k - b_k^*)^2$ can be related to Reg_{Opt} , which leads to a self-bounding inequality that solves to $\text{Reg}_{\text{Opt}}(K) \leq \sqrt{C_1 K \tau} + 2C_1 p_{\min}^{-1} + 2C_2$. Compared to Wang et al. (2023a, Theorem 5.5), our result is more general (applies beyond tabular MDPs) and Thm. D.12 also sharpens lower-order terms (p_{\min} is not multiplied with $K^{1/4}$).

Proof. First, we want to show: under continuous returns (Assump. D.11), we have

$$\text{Reg}_{\text{Opt}}(K) \leq 2\sqrt{C_1 K \tau} + 2C_1 p_{\min}^{-1} + 2C_2. \quad (10)$$

Recall for CVaR, the normalized reward in the AugMDP is precisely $(\hat{b}_k - Z(\pi^k, \hat{b}_k))_+$. Let b_k^* denote the true

τ -quantile of $Z(\pi^k, \hat{b}_k)$.

$$\begin{aligned}
 & \text{Reg}_{\text{Opt}}(K) \\
 & \stackrel{(i)}{\leq} \sqrt{C_1 \sum_k \text{Var}(b_k^* - Z(\pi^k, \hat{b}_k)_+)} + \sqrt{C_1 \sum_k \text{Var}((\hat{b}_k - Z(\pi^k, \hat{b}_k))_+ - (b_k^* - Z(\pi^k, \hat{b}_k))_+)} + C_2 \\
 & \stackrel{(ii)}{\leq} \sqrt{C_1 K \tau} + \sqrt{C_1 \sum_k \text{Var}((\hat{b}_k - Z(\pi^k, \hat{b}_k))_+ - (b_k^* - Z(\pi^k, \hat{b}_k))_+)} + C_2 \\
 & \stackrel{(iii)}{\leq} \sqrt{C_1 K \tau} + \sqrt{C_1 \sum_k (\hat{b}_k - b_k^*)^2} + C_2 \\
 & \stackrel{(iv)}{\leq} \sqrt{C_1 K \tau} + \sqrt{2C_1 p_{\min}^{-1} \text{Reg}_{\text{Opt}}(K)} + C_2 \\
 & \leq \sqrt{C_1 K \tau} + C_1 p_{\min}^{-1} + \frac{1}{2} \text{Reg}_{\text{Opt}}(K) + C_2. \tag{AM-GM}
 \end{aligned}$$

(i) holds since $\sqrt{\text{Var}(X + Y)} \leq \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)}$ for any random variables X, Y . (ii) holds since $\text{Var}(X_+) \leq \mathbb{E}[X^2 \mathbb{I}[X \geq 0]] \leq \Pr(X \geq 0)$ for any bounded $X \in [-1, 1]$ and b_k^* is the true τ -quantile. (iii) holds since $(\hat{b}_k - Z_k)_+ - (b_k^* - Z_k)_+ \leq |\hat{b}_k - b_k^*|$ almost surely. (iv) holds by (Wang et al., 2023a, Lemma G.10): it states that under Assump. D.11, the choice of \hat{b}_k ensures:

$$(\hat{b}_k - b_k^*)^2 \leq 2p_{\min}^{-1} (\hat{V}_{1,k}(s_1, \hat{b}_k) - V_{\text{aug}}^{\pi^k, 1}(s_1, \hat{b}_k)).$$

Rearranging $\text{Reg}_{\text{Opt}}(K)$ implies the desired Eq. (10). Therefore, we can conclude the proof applying Thm. 3.2 and using $V_u^{\max} = \tau^{-1}$ and Eq. (10):

$$\text{Reg}_{\text{OCE}}(K) = 4\tau^{-1/2} \sqrt{C_1 K} + 4\tau^{-1} (C_1 p_{\min}^{-1} + C_2).$$

□

E. Proofs for Policy Optimization Meta-Algorithm

We begin by proving a stronger global convergence guarantee than the one stated in the main paper.

Theorem E.1 (Strong Global Convergence). *Under Assump. 3.5 and Def. 4.1, we have*

$$\sum_{k=1}^K \text{OCE}_u^* - \text{RLB}^{(k)} \leq |\mathcal{B}| V_u^{\max} \text{Reg}_{\text{PO}}(K).$$

Proof.

$$\begin{aligned}
 \sum_{k=1}^K \text{OCE}_u^* - \text{RLB}^{(k)} &= \sum_{k=1}^K \{b_1^* + V_{\text{aug}}^*(s_1, b_1^*)\} - \max_{b \in [0, 1]} \{b + V_{\text{aug}}^{\pi^k, 1}(s_1, b)\} \\
 &\leq \sum_{k=1}^K V_{\text{aug}}^*(s_1, b_1^*) - V_{\text{aug}}^{\pi^k, 1}(s_1, b_1^*) \\
 &\leq \sum_{k=1}^K |\mathcal{B}| \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} (V_{\text{aug}}^*(s_1, b) - V_{\text{aug}}^{\pi^k, 1}(s_1, b)) \\
 &\leq |\mathcal{B}| \text{Reg}_{\text{PO}}(K),
 \end{aligned}$$

where the last step uses the global convergence premise of POALG in the AugMDP. □

This is stronger than the global convergence in the main paper (cf. Thm. 4.2), since the RLB approximately lower bounds $\text{OCE}_u(Z(\pi^k, \hat{b}_k))$, as we prove next. Therefore, Thm. 4.2 is a direct consequence of Thm. E.1 and the following RLB lemma (Eq. (7)).

Lemma 4.3 (RLB). *The $\text{RLB}^{(k)}$ approximately lower bounds the true OCE of π^k with initial budget \hat{b}_k :*

$$\text{RLB}^{(k)} - \text{OCE}_u(\pi^k, \hat{b}_k) \leq 2|\mathcal{B}| \varepsilon_k^{\text{po}}. \tag{7}$$

Moreover, the lower bound is tight on average:

$$\sum_k \text{OCE}_u(\pi^k, \hat{b}_k) - \text{RLB}^{(k)} \leq |\mathcal{B}| V_u^{\max} \text{Reg}_{\text{PO}}(K).$$

Proof. First, we prove Eq. (7). By importance sampling and the value estimate guarantee of Def. 4.1: for all $b_1 \in \mathcal{B}$, we have $|V_1^{\pi^k}(s_1, b_1) - \widehat{V}_1^{\pi^k}(s_1, b_1)| \leq |\mathcal{B}| \varepsilon_k^{\text{po}} =: \varepsilon'_k$. Thus,

$$\begin{aligned}
 \text{OCE}_u(\pi^k, \widehat{b}_k) &= \max_{b \in \mathcal{B}} \{b + \mathbb{E}[u(Z(\pi^k, \widehat{b}_k) - b)]\} \\
 &\geq \{\widehat{b}_k + \mathbb{E}[u(Z(\pi^k, \widehat{b}_k) - \widehat{b}_k)]\} && (\widehat{b}_k \in \mathcal{B}, \text{ by Eq. (5)}) \\
 &= \{\widehat{b}_k + V_{\text{aug}}^{\pi^k, 1}(s_1, \widehat{b}_k)\} && (\text{def. of } V^{\pi^k}) \\
 &\stackrel{(i)}{\geq} \{\widehat{b}_k + \widehat{V}_1^{\pi^k}(s_1, \widehat{b}_k)\} - \varepsilon'_k \\
 &= \max_{b_1 \in \mathcal{B}} \{b_1 + \widehat{V}_1^{\pi^k}(s_1, b_1)\} - \varepsilon'_k && (\text{def. of } \widehat{b}_k, \text{ by Eq. (5)}) \\
 &\stackrel{(ii)}{\geq} \max_{b_1 \in \mathcal{B}} \{b_1 + V_{\text{aug}}^{\pi^k, 1}(s_1, b_1)\} - 2\varepsilon'_k \\
 &= \text{RLB}^{(k)} - 2\varepsilon'_k, && (\text{def. of RLB, by Eq. (6)})
 \end{aligned}$$

where the inequalities (i,ii) are due to the value estimate guarantee. This finishes the proof of Eq. (7).

To prove the second statement, simply apply the stronger global convergence result Thm. E.1 with the fact that $\text{OCE}_u(Z(\pi^k, \widehat{b}_k)) \leq \text{OCE}_u^*$. \square

Theorem 4.4 (Local Improvement). *Under Assump. 3.5 and assuming POALG satisfies the approximate improvement criterion of Def. 4.1, running Alg. 2 ensures that:*

$$\forall k \in [K] : \text{RLB}^{(k+1)} \geq \text{RLB}^{(k)} - |\mathcal{B}| \varepsilon_k^{\text{po}}.$$

Proof. Let $b_k^* = \arg \max_{b \in \mathcal{B}} \{b + V_{\text{aug}}^{\pi^k, 1}(s_1, b)\}$. Then,

$$\begin{aligned}
 \text{RLB}^{(k)} - \text{RLB}^{(k+1)} &= \{b_k^* + V_{\text{aug}}^{\pi^k, 1}(s_1, b_k^*)\} - \max_{b \in [0, 1]} \{b + V_{\text{aug}}^{\pi^{k+1}, 1}(s_1, b)\} \\
 &\leq V_{\text{aug}}^{\pi^k, 1}(s_1, b_k^*) - V_{\text{aug}}^{\pi^{k+1}, 1}(s_1, b_k^*) \\
 &\leq |\mathcal{B}| \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} (V_{\text{aug}}^{\pi^k, 1}(s_1, b) - V_{\text{aug}}^{\pi^{k+1}, 1}(s_1, b)) \\
 &\leq |\mathcal{B}| \varepsilon_k^{\text{po}},
 \end{aligned}$$

where the last step uses local improvement premise of POALG in the AugMDP. \square

E.1. Primer on Natural Policy Gradient in Finite-Horizon MDPs

In this section, we provide a primer on natural policy gradient (NPG) for finite-horizon MDPs and its guarantees, supplementing the case study (Sec. 4.1) of the main paper. Our analysis is based on the infinite-horizon analysis of Agarwal et al. (2021). Let d_1 denote an fixed initial state distribution. For a policy π , let $Q_h^\pi(s_h, a_h) = \mathbb{E}_\pi[r_h + r_{h+1} + \dots + r_H \mid s_h, a_h]$ denote the Q -function, $V_h^\pi(s_h) = \mathbb{E}_{a_h \sim \pi(s_h)}[Q_h^\pi(s_h, a_h)]$ denote the value function, and $A_h^\pi(s, a) = Q_h^\pi(s, a) - V_h^\pi(s)$ denote the advantage function.

We consider policies $\pi^\theta = (\pi_1^\theta, \dots, \pi_H^\theta)$ parameterized by weight vectors $\theta = (\theta_1, \dots, \theta_H)$ where $\theta_h \in \mathbb{R}^m$. Initialize $\theta_h^{(0)} = \mathbf{0}$ for all $h \in [H]$. Consider the generic update rule for $k = 1, 2, \dots, K$: for all h ,

$$\theta_h^{(k+1)} = \theta_h^{(k)} + \eta w_h^{(k)},$$

where $w_h^{(k)} \in \mathbb{R}^m$ have ℓ_2 -norm at most W . Let π^k denote the policy parameterized by $\theta^{(k)}$. Then, we have the following lemma that bounds the sub-optimality of the update rule.

Lemma E.2. *Suppose the log-probability parameterization $\theta_h \mapsto \log \pi_{\theta_h}(a \mid s)$ is β -smooth for all h, s, a . Then, setting $\eta = \sqrt{\frac{2 \log(A)}{\beta K W^2}}$, we have*

$$\begin{aligned}
 \sum_{k=1}^K \mathbb{E}_{s_1 \sim d_1} [V_1^*(s_1) - V_1^{\pi^k}(s_1)] &\leq HW \sqrt{2\beta K \log(A)} + \sum_{h,k} \text{err}_{h,k}, \\
 \text{where } \text{err}_{h,k} &:= \mathbb{E}_{s_1 \sim d_1, \pi^*} [A_h^{\pi^k}(s_h, a_h) - w_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta_h^{(k)}}(a_h \mid s_h)].
 \end{aligned}$$

Proof. First, notice that the smoothness assumption implies that for all h, s, a ,

$$\begin{aligned} \log \pi_h^{k+1}(a | s) - \log \pi_h^k(a | s) &\geq (\theta_h^{(k+1)} - \theta_h^{(k)}) \cdot \nabla_{\theta_h} \log \pi_{\theta_h^{(k)}}(a | s) - \frac{\beta}{2} \|\theta_h^{(k+1)} - \theta_h^{(k)}\|_2^2 \\ &\geq \eta w_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta_h^{(k)}}(a | s) - \frac{\beta \eta^2 W^2}{2}. \end{aligned} \quad (\text{by update rule})$$

Rearranging, we have

$$w_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta_h^{(k)}}(a | s) \leq \eta^{-1} \log(\pi_h^{k+1}(a | s) / \pi_h^k(a | s)) + \frac{\beta \eta W^2}{2}. \quad (11)$$

Then, we can bound the sub-optimality at the k -th round by:

$$\begin{aligned} &\mathbb{E}_{s_1 \sim d_1} [V_1^*(s_1) - V_1^{\pi^k}(s_1)] \\ &= \sum_{h=1}^H \mathbb{E}_{s_1 \sim d_1, \pi^*} [A_h^{\pi^k}(s_h, a_h)] \quad (\text{by performance difference lemma}) \\ &= \sum_h \text{err}_{h,k} + \mathbb{E}_{s_1 \sim d_1, \pi^*} [w_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta_h^{(k)}}(a_h | s_h)] \\ &\leq \sum_h \text{err}_{h,k} + \eta^{-1} \mathbb{E}_{s_1 \sim d_1, \pi^*} [\log(\pi_h^{k+1}(a_h | s_h) / \pi_h^k(a_h | s_h))] + \frac{\beta \eta W^2}{2} \quad (\text{by Eq. (11)}) \\ &= \sum_h \text{err}_{h,k} + \eta^{-1} \mathbb{E}_{s_1 \sim d_1, \pi^*} [D_{KL}(\pi_h^*(s_h) \parallel \pi_h^k(s_h)) - D_{KL}(\pi_h^*(s_h) \parallel \pi_h^{k+1}(s_h))] + \frac{\beta \eta W^2}{2}, \end{aligned}$$

where D_{KL} is the KL-divergence.

Finally, summing over k implies the final result by telescoping:

$$\begin{aligned} &\sum_{k=1}^K \mathbb{E}_{s_1 \sim d_1} [V_1^*(s_1) - V_1^{\pi^k}(s_1)] \\ &\leq \frac{\beta \eta H W^2 K}{2} + \eta^{-1} \sum_h \mathbb{E}_{s_1 \sim d_1, \pi^*} [D_{KL}(\pi_h^*(s_h) \parallel \pi_h^1(s_h)) - D_{KL}(\pi_h^*(s_h) \parallel \pi_h^K(s_h))] + \sum_{h,k} \text{err}_{h,k} \\ &\leq \frac{\beta \eta H W^2 K}{2} + \eta^{-1} H \log(A) + \sum_{h,k} \text{err}_{h,k} \quad (\pi^{(1)} \text{ is uniform}) \\ &= 2\sqrt{\beta H^2 W^2 K \log(A)/2} + \sum_{h,k} \text{err}_{h,k}. \quad (\text{choice of } \eta) \end{aligned}$$

This finishes the proof. \square

A natural idea is to set $w_h^{(k)}$ to minimize the $\text{err}_{h,k}$ error terms, which exactly motivates the NPG update. Specifically, the (idealized) NPG update vector is defined as:

$$\begin{aligned} \tilde{w}_h^{(k)} &= \arg \min_{\|w\|_2 \leq W} L_h(w; \theta^{(k)}, d_h^{\pi^k}), \\ \text{where } L_h(w_h; \theta', \nu_h) &:= \mathbb{E}_{s_h, a_h \sim \nu_h} [(A_h^{\pi_{\theta'}}(s_h, a_h) - w_h \cdot \nabla \log \pi_{\theta'}(a_h | s_h))^2], \end{aligned}$$

for $\theta_h \in \mathbb{R}^d$ and $\nu_h \in \Delta(\mathcal{S} \times \mathcal{A})$. In other words, the idealized NPG update vector minimizes the squared error term while having Euclidean norm at most W . Since the true mean is unknown, we approximate the above using samples. Thus, the actual NPG update we consider minimizes the empirical error

$$w_h^{(k)} := \arg \min_{\|w\|_2 \leq W} \hat{L}_h(w; \theta^{(k)}, \{s_{h,i}, a_{h,i}\}_{i \in [N]}), \quad (12)$$

where $\hat{L}_h(w_h; \theta', \{s_{h,i}, a_{h,i}\}_{i \in [N]}) = \frac{1}{N} \sum_i (A_h^{\pi_{\theta'}}(s_{h,i}, a_{h,i}) - w_h \cdot \nabla \log \pi_{\theta'}(a_{h,i} | s_{h,i}))^2$ and $s_{h,i}, a_{h,i} \sim d_h^{\pi^k}$ are N i.i.d. samples from roll-outs.

We now bound the sub-optimality of the NPG update. We define two sources of error. First, there is the statistical estimation error from using the empirical loss \hat{L}_h instead of the true L_h . This error is expected to converge at a $O(1/\sqrt{N})$ rate or faster, where N is the number of samples per round. Second, there is a transfer / approximation error that measures the performance of the best population vector $\tilde{w}_h^{(k)}$ under the optimal policy's visitations $d_h^{\pi^*}$. This quantity is small if the training policies for data collection cover the traces of the optimal policy π^* . The first estimation error is also measured under the distribution induced by π^* , but can be handled by using the relative condition number, which is a weaker measure of coverage than ℓ_∞ density ratio bounds.

The following results are based on the infinite-horizon MDP analysis of [Agarwal et al. \(2021, Section 6.3\)](#).

Assumption E.3. The update vectors generated by NPG satisfy for all $k \in [K], h \in [H]$:

1. (Excess risk) The statistical estimation error is bounded by $\varepsilon_{\text{stat}}$:

$$L(w_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}) - L(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}) \leq \varepsilon_{\text{stat}}.$$

2. (Coverage) The relative condition number of the covariance under $d_h^{\pi^k}$ and $d_h^{\pi^*}$ is bounded by κ :

$$\sup_{w \in \mathbb{R}^d} \frac{w^\top \Sigma_{h,k,\pi^*} w}{w^\top \Sigma_{h,k,\pi^k} w} \leq \kappa,$$

where $\Sigma_{h,k,\pi^*} := \mathbb{E}_{\pi^*} [\nabla_{\theta_h} \log \pi_h^k(a_h | s_h) (\nabla_{\theta_h} \log \pi_h^k(a_h | s_h))^\top]$.

3. (Transfer Error) Assume the transfer error is bounded by $\varepsilon_{\text{bias}}$:

$$L(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^{\pi^*}) \leq \varepsilon_{\text{bias}}.$$

Theorem E.4 (Agnostic NPG). *Under Assump. E.3, running the NPG update in Eq. (12) enjoys:*

$$\sum_{k=1}^K \mathbb{E}_{s_1 \sim d_1} [V_1^*(s_1) - V_1^{\pi^k}(s_1)] \leq HW \sqrt{2\beta K \log(A)} + HK \sqrt{\varepsilon_{\text{bias}}} + HK \sqrt{\kappa \varepsilon_{\text{stat}}}.$$

This is a finite-horizon analog of Agarwal et al. (2021, Theorem 6.2), where their effective horizon $1/(1 - \gamma)$ is replaced with the horizon H .

Proof. The proof focuses on decomposing the error term in Lemma E.2.

$$\begin{aligned} \text{err}_{h,k} &= \mathbb{E}_{s_1 \sim d_1, \pi^*} [A_h^{\pi^k}(s_h, a_h) - w_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h)] \\ &= \underbrace{\mathbb{E}_{s_1 \sim d_1, \pi^*} [A_h^{\pi^k}(s_h, a_h) - \tilde{w}_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h)]}_{\text{approx.error}} + \underbrace{(\tilde{w}_h^{(k)} - w_h^{(k)}) \cdot \mathbb{E}_{s_1 \sim d_1, \pi^*} [\nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h)]}_{\text{est.error}}. \end{aligned}$$

Let $d_h^* = d_{1,d_1}^{\pi^*}$ be the h -th visitation distribution of rolling in π^* from d_1 . The approximation error can be handled by:

$$\begin{aligned} &\mathbb{E}_{s_1 \sim d_1, \pi^*} [A_h^{\pi^k}(s_h, a_h) - \tilde{w}_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h)] \\ &\leq \sqrt{\mathbb{E}_{s_1 \sim d_1, \pi^*} [(A_h^{\pi^k}(s_h, a_h) - \tilde{w}_h^{(k)} \cdot \nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h))^2]} = \sqrt{L_h(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^*)}. \end{aligned}$$

The estimation error can be handled by:

$$\left| (\tilde{w}_h^{(k)} - w_h^{(k)}) \cdot \mathbb{E}_{s_1 \sim d_1, \pi^*} [\nabla_{\theta_h} \log \pi_{\theta^{(k)}}(a_h | s_h)] \right| \leq \|\tilde{w}_h^{(k)} - w_h^{(k)}\|_{\Sigma_{h,k,\pi^*}}.$$

Then by definition of the relative condition number $\kappa_{h,k} = \|(\Sigma_{h,k,\pi^{(k)}})^{-1/2} \Sigma_{h,k,\pi^*} (\Sigma_{h,k,\pi^{(k)}})^{-1/2}\|_2$, we have

$$\begin{aligned} \|\tilde{w}_h^{(k)} - w_h^{(k)}\|_{\Sigma_{h,k,\pi^*}} &\leq \sqrt{\kappa_{h,k}} \|\tilde{w}_h^{(k)} - w_h^{(k)}\|_{\Sigma_{h,k,\pi^k}} \\ &= \sqrt{\kappa_{h,k} (L_h(w_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}) - L_h(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}))}. \end{aligned}$$

Therefore, we have shown that

$$\begin{aligned} \text{err}_{h,k} &\leq \sqrt{L_h(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^*)} + \sqrt{\kappa_{h,k} (L_h(w_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}) - L_h(\tilde{w}_h^{(k)}; \theta^{(k)}, d_h^{\pi^k}))} \\ &\leq \sqrt{\varepsilon_{\text{bias}}} + \sqrt{\kappa \varepsilon_{\text{stat}}}, \end{aligned}$$

which concludes the proof. \square