Policy Gradient Optimization for Markov Decision Processes with Epistemic Uncertainty and General Loss Functions

Xiaoshuang Wang
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
xwang3094@gatech.edu

Yifan Lin
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
ylin429@gatech.edu

Enlu Zhou
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
enlu.zhou@isye.gatech.edu

Abstract

Motivated by many application problems, we consider Markov decision processes (MDPs) with a general loss function and unknown parameters. To mitigate the epistemic uncertainty associated with unknown parameters, we take a Bayesian approach to estimate the parameters from data and impose a coherent risk functional (with respect to the Bayesian posterior distribution) on the loss. Since this formulation usually does not admit Bellman equations and cannot be solved by approaches based on dynamic programming, we propose a policy gradient optimization method, leveraging the dual representation of coherent risk measures and extending the envelope theorem to continuous cases. We then show the stationary analysis of the algorithm with a convergence rate of $\mathcal{O}(T^{-1/2} + r^{-1/2})$, where T is the number of policy gradient iterations and r is the sample size of the gradient estimator. We further extend our algorithm to an episodic setting, and establish the global convergence of the extended algorithm and provide bounds on the number of iterations needed to achieve an error bound $\mathcal{O}(\epsilon)$ in each episode.

1 Introduction

Many applications require decision making under model uncertainty and evaluation criteria beyond expected return. We study Markov decision processes (MDPs) with unknown parameters and a general convex loss defined on the discounted occupancy measure. To represent epistemic (model) uncertainty, we adopt a Bayesian approach that forms a posterior distribution over MDP parameters and then evaluate policies through a coherent risk functional applied to that posterior. This composite objective generally breaks Bellman recursion, so dynamic programming methods do not apply. We develop a policy-gradient optimization framework for our setting. Our derivation exploits the dual representation of coherent risk measures and extends the envelope theorem to continuous environment parameter spaces, enabling a tractable gradient for general convex losses composed with Bayesian risk. We provide a non-asymptotic stationary analysis: after T gradient iterations using r posterior samples to estimate the gradient, the average gradient norm decreases at rate $\mathcal{O}(T^{-1/2} + r^{-1/2})$. We further give an episodic extension in which the posterior is updated with newly collected data;

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

under mild conditions, the algorithm converges globally under the true environment as data grow, with per-episode iteration bounds to reach an $\mathcal{O}(\varepsilon)$-accurate solution under the Bayesian formulation.

Our contributions are summarized as follows: (1) We propose a Bayesian risk formulation for MDPs with a general convex loss function and develop a policy gradient algorithm to solve for the optimal policy. The proposed formulation jointly mitigates both epistemic and intrinsic uncertainty; (2) We extend the envelope theorem to the dual representation of the coherent risk measure, and then apply the envelope theorem to derive the policy gradient. Our extension from the discrete case to the continuous case for the envelope theorem may be of independent interest; (3) We prove the convergence of the proposed algorithm and establish its convergence rate as $\mathcal{O}(T^{-1/2} + r^{-1/2})$, where T is the number of policy gradient iterations and r is the sample number of the gradient estimator; (4) We extend our policy gradient algorithm to the episodic setting, and prove the asymptotic convergence of the episodic minimizer of our Bayesian formulation to a global minimizer of the MDP problem under the true environment. Moreover, we show the number of iterations required in any episode to maintain an optimality gap $\mathcal{O}(\epsilon)$ under our Bayesian formulation.

2 Problem Formulation

Consider an infinite-horizon Markov Decision Process (MDP) over a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. For each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, a transition to the next state s' follows the transition kernel P^*, i.e. $s' \sim P^*(\cdot|s,a)$. A stationary policy π is defined as a function mapping from the state space to a probability simplex $\Delta(\cdot)$ over the action space. Given any transition probability P, define $\lambda^{\pi,P}$ to be the discounted state-action occupancy measure under policy π :

$$\lambda_{sa}^{\pi,P} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \mid \pi, s_0 \sim \tau, P\right) \tag{1}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where τ is the initial distribution, $\gamma \in (0,1)$ is the discount factor. We consider a general loss function $F(\lambda, P)$ defined over the occupancy measure λ and transition kernel P, which is assumed to be convex in λ. In practice, the true distribution P^* is usually unknown and needs to be estimated. In this work, we take a Bayesian approach to estimate the environment. We assume that the transition kernel $P^* \equiv P_{\theta^*}$ is parameterized by θ^*, where $\theta^* \in \Theta$ is the true but unknown parameter value, $\Theta \subseteq \mathbb{R}^p$ is the parameter space, and p is the dimension of Θ. Under the parametric assumption, we assume we have access to some data which are state transitions $\zeta = (s,a,s')$, where s' follows the distribution $P_{\theta^*}(\cdot|s,a)$ and define $P_{\theta^*}(\zeta) := P_{\theta^*}(s'|s,a)$. Now given a fixed batch of data $\zeta^{(N)}$ of N samples, we can update the posterior distribution (denoted by μ_N) on the parameter θ using the Bayes rule: $\mu_N(\theta) = \frac{P_\theta(\zeta^{(N)}))\mu_0(\theta)}{\int_{\theta'} P_{\theta'}(\zeta^{(N)})\mu_0(\theta')d\theta'}$, where μ_0 is a prior distribution of θ we assume. Furthermore, model mis-specification caused by the lack of data could lead to sub-optimality of the learned policy when it is implemented in a real-world setting. Hence, we further impose a risk functional on the objective with respect to (w.r.t.) the Bayesian posterior to account for the epistemic uncertainty, which results in the following composed formulation:

$$\min_{\pi} \rho_{\theta\sim\mu_N}(F(\lambda^{\pi,P_\theta}, P_\theta)) \tag{2}$$

where ρ is a general coherent risk measure w.r.t. the posterior μ_N. We aim to solve problem (2) in this paper. By this formulation, we look for a policy that minimizes a performance measure taking into account the epistemic uncertainty caused by lack of data for a general convex loss function.

3 Policy Gradient Algorithm: Derivation and Estimation

We adopt the policy gradient algorithm, which directly optimizes parameterized policies. Consider a stochastic parameterized policy $\pi_\alpha : \mathcal{S} \to \Delta(\mathcal{A})$, parameterized by $\alpha \in W \subset \mathbb{R}^d$. To directly work on the parameterized policy, we denote $F(\lambda^{\pi_\alpha,P_\theta}, P_\theta)$ by $C(\alpha,\theta)$. The policy optimization problem (2) then becomes:

$$\min_{\alpha} G(\alpha) := \rho_{\theta\sim\mu_N}(C(\alpha,\theta)). \tag{3}$$

As shown in (Shapiro et al., 2021), a coherent risk measure has a well-known dual representation. According to the dual representation, we can write the coherent risk measure as a maximization

problem, where the decision variable is ξ and the objective is a linear functional of ξ, and define the Lagrangian function $L_\alpha\left(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}\right)$ for it. Using the Lagrangian relaxation, we derive the policy gradient in Theorem 1. Detailed derivations can be found in Appendix A.1.

Theorem 1. *Assume that Assumption A.1 and A.2 hold, and ρ satisfies Definition A.1. Assume that μ_N is a Radon measure (see Appendix H.1 for definition). Define $\xi^* \in \arg\max_{\xi\geq 0, \|\xi\|_q \leq B_q} \min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}} L_\alpha\left(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}\right)$. Then we have the policy gradient*

$$g(\alpha) := \nabla_\alpha \rho_{\theta\sim\mu_N}(C(\alpha,\theta)) = \int_{\theta\in\Theta} \xi^*_\alpha(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)d\theta. \tag{4}$$

To estimate $\nabla_\alpha C(\alpha,\theta)$, any plug-in estimation method satisfies our demand. Here, we adopt the variational policy gradient theorem in (Zhang et al., 2020), which considers the policy gradient for a concave function defined on the occupancy measure for a RL problem. Details are offered in Appendix A.1. To evaluate the integral $\int_{\theta\in\Theta} \xi^*_\alpha(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)d\theta$, we use samples $\{\theta_k\}_{k=1}^r$ from the posterior distribution μ_N to construct the policy gradient estimator

$$\nabla_\alpha \rho_{\theta\sim\mu_N}(C(\alpha,\theta)) \approx \widehat{g}(\alpha) := \frac{1}{r}\sum_{k=1}^{r} \xi^*(\theta_k)\nabla_\alpha C\left(\alpha,\theta_k\right). \tag{5}$$

To perform policy gradient optimization, we iteratively use the gradient descent step (6), where η_t is the step size, and $\mathrm{Proj}_W(x) = \arg\min_{y\in W}\|y-x\|_2^2$ is the projection into the parameter space W. We summarize the full algorithm in Algorithm 1 in Appendix G.

$$\alpha_{t+1} = \arg\min_{\alpha\in W} \langle \widehat{g}(\alpha_t), \alpha-\alpha_t\rangle + \frac{\eta_t}{2}\|\alpha-\alpha_t\|^2 = \mathrm{Proj}_W\left(\alpha_t - \frac{1}{\eta_t}\widehat{g}(\alpha_t)\right). \tag{6}$$

So far we have considered the offline setting with a fixed batch of data, but in many application problems data can be collected periodically. The collected data can be then used to learn about the environment and update the policy. This process can be repeated iteratively. Thus, we extend our approach to an episodic setting as described above. A potential approach is to use Algorithm 1 to make policy updates during each episode, as detailed in Algorithm 2 in Appendix G.

4 Convergence Analysis

In this section, we analyze the convergence properties of Algorithm 1 and Algorithm 2. We demonstrate the finite-time first-order convergence rate is $\mathcal{O}(T^{-1/2}+r^{-1/2})$, where T is the number of policy gradient iterations and r is the sample number of the gradient estimator.

Theorem 2. *(Stationary convergence) Suppose that Assumption A.1, A.2, B.1 and H.1 hold. By choosing $\eta_t = 2L_G$ in Algorithm 1, it holds that $\mathbb{E}\|\nabla G(\alpha_{out})\| \leq \mathcal{O}\left(T^{-1/2}+r^{-1/2}\right)$.*

Theorem 2 shows that the gradient bound of the output policy consists of two parts: an asymptotically diminishing error bound $T^{-1/2}$ in the exact setting and an estimation error bound $r^{-1/2}$ of the policy gradient. The total sample complexity from the posterior μ_N to achieve accuracy $\mathcal{O}(\epsilon)$ is $\mathcal{O}(\epsilon^{-4})$ by choosing $T=\epsilon^{-2}$ and $r=\epsilon^{-2}$. Because of the intrinsic non-convex structure under a fixed posterior, only convergence to a stationary point can be achieved under a fixed posterior (or in other words, in a fixed episode), which is discussed in Appendix D. However, global convergence can be achieved in the episodic setting as the posterior updates and converges. This is shown in Corollary 1 later.

Theorem 3. *(Consistency of episodic optimal policy) Suppose that Assumption A.1, A.2, B.1, H.1 and H.2 hold. Define $G_i(\alpha) := \rho_{\theta\sim\mu_i}(C(\alpha,\theta))$, which is the objective for the posterior μ_i with data size i. Then we have $D_i := \sup_{\alpha\in W}|G_i(\alpha)-C(\alpha,\theta^*)|$ and $E_i := \sup_{\alpha\in W}\|\nabla_\alpha G_i(\alpha)-\nabla_\alpha C(\alpha,\theta^*)\|_2$ tend to 0 with probability 1 as $i\to\infty$, where the probability is w.r.t. the data-generating distribution. Moreover, $C(\alpha_i^*,\theta^*)-C(\alpha^*,\theta^*)\to 0$ with probability 1 as $i\to\infty$, where α_i^* is a global minimizer of $G_i(\alpha)$ and α^* is a global minimizer of $C(\alpha,\theta^*)$.*

As the data size N increases, the posterior distribution converges to a Dirac measure, which is a point mass at the true parameter θ^*. Consequently, the performance of the optimal policy for the posterior μ_N converges to the optimal policy under the true environment θ^*, as demonstrated in Theorem 3. In the episodic setting, we iteratively use the current policy for data collection and posterior updates,

and perform policy updates based on the updated posterior, as described in Algorithm 2. Notably, the inner map $\lambda(\alpha, \theta^*)$ from policy parameter to occupancy measure is not necessarily convex in α, though $F(\lambda, \theta^*)$ is convex in λ. However, the hidden convexity can be utilized to get the global optimality under the true environment, regardless of the gradient estimation method.

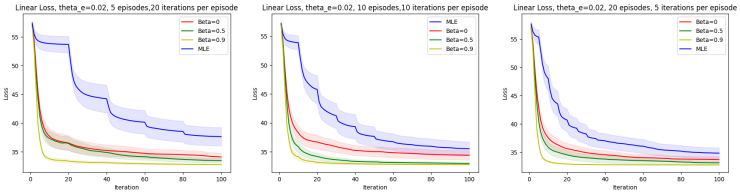
Theorem 4. *(Finite-episode error bound) Suppose that Assumption A.1, A.2, B.1, H.1, H.2 and H.3 hold. Assume that $G_i(\alpha)$ defined in Theorem 3 has $L_{G,i}$-Lipschitz continuous gradient. Let $\{\alpha_{i,j}\}_{i=1}^{N}{}_{j=0}^{t_i}$ be the generated policy parameter sequences for N episodes by Algorithm 2. For any $\epsilon > 0$, if we choose $t_i = \Theta(L_{G,i}(E_{i-1}+D_i)\epsilon^{-2})$, $r = \Theta(\epsilon^{-2})$, then we can keep a constant gradient bound $\mathbb{E}\left[\left(\sum_{j=0}^{t_i-1} \|\nabla G_i(\alpha_{i,j})\|_2\right)/t_i\right] \leq \epsilon$ for each i. Furthermore, $\mathbb{E}C(\alpha_{out}, \theta^*) - C(\alpha^*, \theta^*) \leq \mathcal{O}(\epsilon + E_N)$, where D_i, E_i are defined in Theorem 3.*

Corollary 1. *(Global convergence) Using the same assumptions and notations in Theorem 4, $\mathbb{E}C(\alpha_{out}, \theta^*) - C(\alpha^*, \theta^*) \leq \mathcal{O}(\epsilon)$ for any $\epsilon > 0$ as $N \to \infty$.*

Theorem 4 offers theoretical advice on how to choose the iteration number in each episode. Generally speaking, we need fewer iterations to keep the gradient bound when i grows since D_i, E_i approaches 0. Corollary 1 is a direct result of Theorem 3 and Theorem 4, and Corollary 1 implies global convergence to the true optimal policy in the episodic setting (when the data size N increases to infinity). Detailed proofs can be found in Appendix H.

5 Numerical Experiments

We evaluate our proposed formulation and algorithm on the offline Frozen Lake problem (Ravichandiran, 2018), an OpenAI benchmark. We refer readers to Appendix L for a detailed description of the problem, experiment design and results. We consider different convex loss functions, including the mean and Kullback-Leibler (KL) divergence, for various tasks. We compare the Bayesian Risk Policy Gradient (BR-PG) algorithm with CVaR risk measure under different risk levels $\beta = 0, 0.5, 0.9$, respectively, with two other methods. For example, Figure 1 shows that the loss of our algorithm decreases quickly in spite of few data. In a nutshell, our method outperforms other two methods and show robustness in the case of few data.



(a) 5×20 total iterations (b) 10×10 total iterations (c) 20×5 total iterations

Figure 1: Results for episodic case with different episode numbers and iterations per episode under the same escape probability $\theta_e = 0.02$ and 50 replications. Here the loss function is still chosen to be the linear loss. 95% confidence intervals are reported by the shaded bands.

6 Conclusions

In this paper, we develop a Bayesian risk approach to jointly address both epistemic and intrinsic uncertainty in the infinite-horizon MDP. For a general coherent risk measure and a general convex loss function, we design a policy gradient algorithm for the proposed formulation and demonstrate the algorithm's convergence at a rate of $\mathcal{O}(T^{-1/2} + r^{-1/2})$. Furthermore, we establish the consistency of an online episodic extension and provide bounds on the number of iterations required to maintain a constant gradient bound $\mathcal{O}(\epsilon)$ for each episode. The numerical experiments confirm the stationary analysis of the proposed algorithm and demonstrate the robustness of the formulation under various loss functions.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76, 2021.

Altman, E. *Constrained Markov decision processes*. Routledge, 2021.

Bai, Q., Bedi, A. S., Agarwal, M., Koppel, A., and Aggarwal, V. Achieving zero constraint violation for concave utility constrained reinforcement learning via primal-dual approach. *Journal of Artificial Intelligence Research*, 78:975–1016, 2023.

Balasubramanian, K. and Ghadimi, S. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pp. 1–42, 2022.

Barakat, A., Fatkhullin, I., and He, N. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pp. 1753–1800. PMLR, 2023.

Brezis, H. and Brézis, H. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

Hong, L. J. and Liu, G. Simulating sensitivities of conditional value at risk. *Management Science*, 55 (2):281–293, 2009.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022.

Milgrom, P. and Segal, I. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Ravichandiran, S. *Hands-on reinforcement learning with Python: master reinforcement and deep reinforcement learning using OpenAI gym and tensorFlow*. Packt Publishing Ltd, 2018.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Shapiro, A., Zhou, E., and Lin, Y. Bayesian distributionally robust optimization. *SIAM Journal on Optimization*, 33(2):1279–1304, 2023.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Solla, S. A., Leen, T. K., and Müller, K.-R. (eds.), *Advances in Neural Information Processing Systems*, pp. 1057–1063, 1999.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

Ying, D., Guo, M. A., Ding, Y., Lavaei, J., and Shen, Z.-J. Policy-based primal-dual methods for convex constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10963–10971, 2023.

Ying, D., Zhang, Y., Ding, Y., Koppel, A., and Lavaei, J. Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, pp. 4572–4583, 2020.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34: 2228–2240, 2021.

Zhang, J., Bedi, A. S., Wang, M., and Koppel, A. Multi-agent reinforcement learning with general utilities via decentralized shadow reward actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9031–9039, 2022.

A Policy Gradient Algorithm: Derivation and Estimation

As discussed in the introduction, the dynamic programming type of algorithms may not be applicable to a general convex loss function $F(\cdot)$. So we adopt the policy gradient algorithm, which directly optimizes parameterized policies. Consider a stochastic parameterized policy $\pi_\alpha : \mathcal{S} \to \Delta(\mathcal{A})$, parameterized by $\alpha \in W \subset \mathbb{R}^d$. To directly work on the parameterized policy, we denote $F(\lambda^{\pi_\alpha, P_\theta}, P_\theta)$ by $C(\alpha, \theta)$. The policy optimization problem (2) then becomes:

$$\min_{\alpha} G(\alpha) := \rho_{\theta \sim \mu_N}(C(\alpha, \theta)). \tag{7}$$

It is worth mentioning that $G(\alpha)$ is not necessarily a convex function though F is convex w.r.t. λ. So we can only reach a stationary point of $G(\alpha)$ by the policy gradient descent method. In the rest of the section, we derive the policy gradient to the proposed formulation (3) using the envelope theorem, and construct the policy gradient estimator. It should be noted that our proposed formulation allows for flexible methods to estimate the policy gradient, including the variational approach such as in (Zhang et al., 2020), and the zeroth-order method such as in (Balasubramanian & Ghadimi, 2022).

A.1 Preliminaries

To ensure the objective $G(\alpha)$ is well defined, we first assume that $C(\alpha, \theta) \in \mathcal{Z} := L_p(\Theta, \mu_N)$.

Assumption A.1. $C(\alpha, \theta) \in \mathcal{Z} = \{f : \|f\|_p := \left(\int_\Theta |f(\theta)|^p d\mu_N(\theta)\right)^{1/p} < \infty\}, \forall \alpha \in W$, *for some* $p \geq 1$.

The choice of p depends on the specific coherent risk measure. For example, p can be chosen as 1 for CVaR introduced in Example 1. Let $\mathcal{B} := \{\xi \in \mathcal{Z}^* : \int_\Theta \xi(\theta)\mu_N(\theta)d\theta = 1, \xi \succeq 0\}$, where $\mathcal{Z}^* := L_q(\Theta, \mu_N)$ is the dual space of $\mathcal{Z}$ with $1/p + 1/q = 1$. As shown in (Shapiro et al., 2021), a coherent risk measure has a well-known dual representation.

Theorem 5. *(Theorem 6.6 in (Shapiro et al., 2021).) A risk measure $\rho : \mathcal{Z} \to \mathbb{R}$ is coherent if and only if there exists a convex bounded and closed set (also known as risk envelope) $\mathcal{U} = \mathcal{U}(\mu_N) \subset \mathcal{B}$ such that $\rho(Z) = \max_{\xi : \xi \in \mathcal{U}(\mu_N)} \mathbb{E}_\xi[Z]$, where $\mathbb{E}_\xi[Z] := \int_{\theta \in \Theta} Z(\theta)\xi(\theta)\mu_N(\theta)d\theta$.*

Note ξ could be viewed as perturbation on the posterior μ_N that satisfies certain conditions, and the risk measure can be understood as the extreme performance for these perturbations. Theorem 5 implies that a functional ρ defined by $\rho(Z) = \max_{\xi : \xi \in \mathcal{U}} \mathbb{E}_\xi[Z]$ is a coherent risk measure if $\mathcal{U} \subset \mathcal{B}$ is convex, bounded and closed. In this paper we only focus on a class of coherent risk measures ρ following Definition A.1 throughout the paper.

Definition A.1. *For each given policy parameter $\theta \in \mathbb{R}^K$, there exists an expression for the risk envelope $\mathcal{U}$ of the coherent risk measure ρ in the following form:*

$$\mathcal{U}(\mu_N) = \{\xi \in \mathcal{Z}^* : g_e(\xi, \mu_N) = 0, \forall e \in \mathcal{E}, f_i(\xi, \mu_N) \leq 0,$$
$$\forall i \in \mathcal{I}, \int_{\theta \in \Theta} \xi(\theta)\mu_N(\theta)d\theta = 1, \xi(\theta) \geq 0, \|\xi\|_q \leq B_q\},$$

where $g_e(\xi, \mu_N)$ is an affine function in ξ, $f_i(\xi, \mu_N)$ is a convex function in ξ, $\|\cdot\|_q$ is the L_q norm in $\mathcal{Z}^*$, and there exists a strictly feasible point $\bar{\xi}$. $\mathcal{E}$ and $\mathcal{I}$ here denote the sets of equality and inequality constraints, respectively. Furthermore, for any given $\xi \in \mathcal{B}$, $f_i(\xi, \mu_N)$ and $g_e(\xi, \mu_N)$ are twice differentiable in μ_N, and there exists a $M > 0$ such that $\forall \omega \in \Omega$:

$$\max\left\{\max_{i \in \mathcal{I}} \left|\frac{df_i(\xi, \mu_N)}{d\mu_N(\theta)}\right|, \max_{e \in \mathcal{E}} \left|\frac{dg_e(\xi, \mu_N)}{d\mu_N(\theta)}\right|\right\} \leq M.$$

The conditions on g_e and f_i ensure that risk envelope $\mathcal{U}(\mu_N)$ is a convex closed set, and the condition $\|\xi\|_q \leq B_q$ makes $\mathcal{U}(\mu_N)$ bounded. A similar assumption is considered in Assumption 2.2(Tamar et al., 2015). The assumption about bounded derivatives can be easily satisfied if Θ is compact. While (Tamar et al., 2015) only consider the case where Θ is finite, we extend it to the continuous case, leading to a functional problem over an infinite dimensional space instead of a finite-dimensional case. Therefore, we extend the result in (Tamar et al., 2015) to the infinite dimensional space, which is shown in Theorem 1. Notably, the function forms of $g_e(\cdot)$ and $f_i(\cdot)$ can be exactly specified for a given coherent risk measure. We refer the readers to Appendix J and Section 6.3.2 (Shapiro et al., 2021) for some examples of the envelope set for coherent risk measures, which cover most common coherent risk measures.

A.2 Derivation of Policy Gradient

According to Theorem 5, we can write the coherent risk measure as a maximization problem (8), where the decision variable is ξ and the objective is a linear functional of ξ, and define the Lagrangian function (9) for problem (8):

$$\rho_{\theta\sim\mu_N}(C(\alpha,\theta)) = \max_{\xi:\xi\in\mathcal{U}(\mu_N)} \mathbb{E}_\xi[C(\alpha,\theta)] = \max_{\xi:\xi\in\mathcal{U}(\mu_N)} \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta)C(\alpha,\theta)d\theta. \tag{8}$$

$$\begin{aligned} L_\alpha(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}) &= \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta)C(\alpha,\theta)d\theta - \lambda^{\mathcal{P}}\left(\int_{\theta\in\Theta}\xi(\theta)\mu_N(\theta)d\theta - 1\right) \\ &- \sum_{e\in\mathcal{E}} \lambda^{\mathcal{E}}(e) g_e\left(\xi,\mu_N\right) - \sum_{i\in\mathcal{I}} \lambda^{\mathcal{I}}(i) f_i\left(\xi,\mu_N\right). \end{aligned} \tag{9}$$

Using the Lagrangian relaxation (9), we derive the policy gradient for (8) in Theorem 1. For this purpose, we need some mild assumptions on the objective function.

Assumption A.2. *(1) $\nabla_\lambda F(\lambda, P)$ is uniformly bounded by $L_{F,\infty}$ for any λ and P w.r.t. $\|\cdot\|_\infty$; (2) $\nabla_\alpha C(\alpha,\theta)$ is $L_{\theta,2}$-Lipschitz continuous w.r.t. $\theta\in\Theta$ and $\|\cdot\|_2$ for any $\alpha\in W$; (3) $\nabla C(\alpha,\theta)$ is uniformly bounded by B for any $\alpha\in W$ and $\theta\in\Theta$ w.r.t. $\|\cdot\|_2$; (4) $\Theta\subseteq\mathbb{R}^p$ is compact and convex; (5) W, the domain of α, is bounded by B_W.*

Assumption A.2 requires the uniform boundedness and Lipschitz continuity of ∇C and ∇F, where $C(\alpha,\theta) = F(\lambda^{\pi_\alpha,P_\theta}, P_\theta)$. One sufficient condition easy to verify for Assumption A.2 to hold is: each component in the composed function $F(\lambda^{\pi_\alpha,P_\theta}, P_\theta)$ is (somewhere) twice continuously differentiable w.r.t parameters α,θ, and the domains of two parameters are compact convex sets.

Theorem 6. *Assume that Assumption A.1 and A.2 hold, and ρ satisfies Definition A.1. Assume that μ_N is a Radon measure (see Appendix H.1 for definition). Define $\xi^* \in \arg\max_{\xi\geq 0,\|\xi\|_q\leq B_q} \min_{\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}}} L_\alpha\left(\xi,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}\right)$. Then we have the policy gradient*

$$g(\alpha) := \nabla_\alpha \rho_{\theta\sim\mu_N}(C(\alpha,\theta)) = \int_{\theta\in\Theta} \xi^*_\alpha(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)d\theta. \tag{10}$$

Proof details of Theorem 1 can be found in Appendix H.2. Theorem 1 implies that we can plug in a saddle point of Lagrangian (9) into (4) to get the policy gradient. However, (4) still involves ∇C, the gradient of the loss function, and the integration w.r.t. the posterior μ_N. In the next subsection, we show how to estimate the policy gradient in (4).

A.3 Construction of the Policy Gradient Estimator

In this section, we focus on how to estimate the policy gradient $g(\alpha)$ and denote its estimator by $\widehat{g}(\alpha)$. We first need to find ξ^* in Theorem 1. For some coherent risk measures, the closed-form expression of ξ^* is known. For CVaR with risk level $\beta\in(0,1)$, $\xi^*(\theta) = \frac{1}{1-\beta}$ if $C(\alpha,\theta)\geq v_\beta$ and 0 otherwise, where v_β is the β-quantile of $C(\alpha,\theta)$. For a general coherent risk measure, we can use the approach sample average approximation (SAA). We first sample $\theta_k, k=1,\ldots,r$, from μ_N, and then solve the following SAA problem for the solution $\xi^*(\theta_k)$ for each k:

$$\begin{aligned} &\max_{\substack{\xi\geq 0,\\ (\sum_{k=1}^r |\xi(\theta_k)|^q)/r\leq B_q}} \min_{\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{P}},\lambda^{\mathcal{E}}} \frac{1}{r}\sum_{k=1}^r \xi(\theta_k)C(\alpha,\theta_k) - \lambda^{\mathcal{P}}\left(\frac{1}{r}\sum_{k=1}^r \xi(\theta_k) - 1\right) \\ &- \sum_{e\in\mathcal{E}} \lambda^{\mathcal{E}}(e) g_e\left(\xi^{(r)},\mu_N(r)\right) - \sum_{id\in\mathcal{I}} \lambda^{\mathcal{I}}(k) f_i\left(\xi^{(r)},\mu_N(r)\right) \end{aligned} \tag{11}$$

Notice the objective in (11) is linear w.r.t. $\lambda^{\mathcal{I}},\lambda^{\mathcal{E}}$ and concave w.r.t ξ, and the domain of ξ is a convex bounded set in $\mathbb{R}^r$. Thus, (11) can be solved by any max-min optimization algorithm for a concave-convex function, such as alternating gradient descent ascent. Here we assume that we can solve (11) to derive $\xi^*(\theta_k)$ accurately for each k. Apart from ξ^*, we need to estimate $\nabla_\alpha C(\alpha,\theta)$ and the integral $\int_{\theta\in\Theta}\xi^*_\alpha(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)$. To estimate $\nabla_\alpha C(\alpha,\theta)$, any plug-in estimation method satisfies our demand. Here, we adopt the variational policy gradient theorem in (Zhang et al., 2020), which considers the policy gradient for a concave function defined on the occupancy measure for a RL

problem. Different from our Bayesian-risk problem with a general loss function, (Zhang et al., 2020) only considers the inner-layer F of our objective (2) in the online setting. It should also be noticed that their method can be replaced by other methods such as the zeroth-order estimation method in (Balasubramanian & Ghadimi, 2022). While the variational policy gradient theorem require access to the conjugate function F^*, which may be difficult to calculate in some cases, zeroth-order method only requires function evaluation of F but leads to higher computational cost in general cases.

Lemma A.1. *(Theorem 3.1 in (Zhang et al., 2020)) Suppose F is convex and continuously differentiable in an open neighborhood of $\lambda^{\pi_\alpha, P_\theta}$. Fix the transition kernel P_θ and denote $V(\alpha; z)$ to be the expected cumulative cost of policy π_α when the cost function is z, and assume $\nabla_\alpha V(\alpha; z)$ always exists. Then we have*

$$\nabla_\alpha C(\alpha, \theta) = -\lim_{\delta \to 0_+} \operatorname*{argmin}_{x \in \mathbb{R}^{SA}} \sup_{z \in \mathbb{R}^{SA}} \{V(\alpha; z) + \delta \nabla_\alpha V(\alpha; z)^\top x - F^*(z) + \frac{\delta}{2}\|x\|^2\}, \tag{12}$$

where $V(\alpha; z) = \langle z, \lambda(\alpha, \theta)\rangle$, $\nabla_\alpha V(\alpha; z)^\top x = \langle z, \nabla_\alpha \lambda(\alpha, \theta) x\rangle$, $F^*(z) := \sup_{x \in \mathbb{R}^{SA}} x^\top z - F(x)$ is the Fenchel conjugate of F.

It may have a high computational cost if we directly estimate each part at a specific α in $\nabla_\alpha C = \nabla_\lambda F \cdot \nabla_\alpha \lambda$. The variational policy gradient method bypasses this issue by changing this problem into a problem of calculating some linear functions and the conjugate function at any z, shown in (12). (Zhang et al., 2020) considers an online setting and thus they need to interact with the environment to estimate $\nabla_\alpha C$. In our offline setting, we can directly solve (12) to get $\nabla_\alpha C$. An example algorithm to solve (12) is given in Appendix H.3. To evaluate the integral $\int_{\theta \in \Theta} \xi_\alpha^*(\theta) \mu_N(\theta) \nabla_\alpha C(\alpha, \theta) d\theta$, we use samples θ_k to construct the policy gradient estimator

$$\nabla_\alpha \rho_{\theta \sim \mu_N}(C(\alpha, \theta)) \approx \widehat{g}(\alpha) := \frac{1}{r} \sum_{k=1}^{r} \xi^*(\theta_k) \nabla_\alpha C(\alpha, \theta_k). \tag{13}$$

In this paper, we assume the access to samples from the posterior distribution μ_N. While computing the posterior often requires costly methods like Markov Chain Monte Carlo (MCMC), utilizing a conjugate prior yields closed-form updates for the posterior parameters, improving efficiency. Computing the posterior typically requires expensive methods such as Markov Chain Monte Carlo (MCMC). However, by utilizing a conjugate prior, we obtain a closed-form expression for the posterior parameters, making the calculation more efficient.

A.4 Full Algorithm and Episodic Setting

To perform policy gradient optimization, we iteratively use the gradient descent step (6), where η_t is the step size, and $\mathrm{Proj}_W(x) = \arg\min_{y \in W} \|y - x\|_2^2$ is the projection into the parameter space W. We summarize the full algorithm in Algorithm 1 in Appendix G.

$$\alpha_{t+1} = \arg\min_{\alpha \in W} \langle \widehat{g}(\alpha_t), \alpha - \alpha_t \rangle + \frac{\eta_t}{2}\|\alpha - \alpha_t\|^2 = \mathrm{Proj}_W\left(\alpha_t - \frac{1}{\eta_t}\widehat{g}(\alpha_t)\right). \tag{14}$$

So far we have considered the offline setting with a fixed batch of data, but in many application problems data can be collected periodically. Again, consider a self-driving car as an example: the car is trained in an offline setting and then deployed to a real environment for a test drive while collecting more data from the environment. The collected data can be then used to learn about the environment and update the policy. This process can be repeated iteratively. Thus, we extend our approach to an episodic setting as described above. A potential approach is to use Algorithm 1 to make policy updates during each episode, as detailed in Algorithm 2 in Appendix G.

B Convergence Analysis

In this section, we analyze the convergence properties of Algorithm 1 and Algorithm 2. We begin by establishing the error bound for the policy gradient estimator. Next, we demonstrate the finite-time first-order convergence rate is $\mathcal{O}(T^{-1/2} + r^{-1/2})$, where T is the number of policy gradient iterations and r is the sample number of the gradient estimator. Furthermore, we prove the consistency of the proposed Bayesian risk formulation, meaning that the optimal policy obtained through this formulation converges to the one obtained by solving the true problem as the number of initial data

points N approaches infinity. Lastly, for the episodic setting we show the number of iterations required in any episode to maintain an $\mathcal{O}(\epsilon)$-error bound over all episodes, which implies the global convergence of Algorithm 2 as N goes to infinity.

B.1 Estimation Error of the Policy Gradient

Assumption B.1. *Assume ξ^* in Theorem 1 satisfies* $\sup_{\alpha\in W} \mathrm{Var}_{\theta\sim\mu_N}[\xi^*(\theta)\nabla C(\alpha,\theta)] = \sigma_\xi < \infty$.

Assumption B.1 requires the uniformly bounded variance of $\xi^*\nabla C$. It is hard to show some property of ξ^* in a general case as the envelope set is given in a general form. One sufficient condition for Assumption 4.1 to hold is that ξ^* is bounded on Θ. As Θ is a compact and convex set, it is not a strong condition.

Theorem 7. *Assume Assumption A.1, A.2 and B.1 hold. By using r samples for gradient estimator in* (5)*, the gradient estimation error is* $\mathbb{E}\left[\|\widehat{g}(\alpha) - g(\alpha)\|_2^2\right] \leq \frac{\sigma_\xi}{r}, \forall \alpha \in W$.

Theorem 7 implies that the sample complexity of $\Theta(1/\epsilon)$ is required to achieve the estimation error $\mathcal{O}(\epsilon)$. Please refer to Appendix H.4 for the detailed proof.

B.2 Convergence Analysis

First we make an assumption about the Lipschitz continuity of $g(\alpha)$ in Assumption H.1.

Theorem 8. *(Stationary convergence) Suppose that Assumption A.1, A.2, B.1 and H.1 hold. By choosing $\eta_t = 2L_G$ in Algorithm 1, it holds that* $\mathbb{E}\|\nabla G(\alpha_{out})\| \leq \mathcal{O}\left(T^{-1/2} + r^{-1/2}\right)$.

Theorem 2 shows that the gradient bound of the output policy consists of two parts: an asymptotically diminishing error bound $T^{-1/2}$ in the exact setting and an estimation error bound $r^{-1/2}$ of the policy gradient. The total sample complexity from the posterior μ_N to achieve accuracy $\mathcal{O}(\epsilon)$ is $\mathcal{O}(\epsilon^{-4})$ by choosing $T = \epsilon^{-2}$ and $r = \epsilon^{-2}$. The proof and assumptions are shown in Appendix H.5. Because of the intrinsic non-convex structure under a fixed posterior, only convergence to a stationary point can be achieved under a fixed posterior (or in other words, in a fixed episode). Detailed discussion about the non-convex structure is provided in Appendix D. However, global convergence can be achieved in the episodic setting as the posterior updates and converges. This is shown in Corollary 1 later.

Theorem 9. *(Consistency of episodic optimal policy) Suppose that Assumption A.1, A.2, B.1, H.1 and H.2 hold. Define $G_i(\alpha) := \rho_{\theta\sim\mu_i}(C(\alpha,\theta))$, which is the objective for the posterior μ_i with data size i. Then we have $D_i := \sup_{\alpha\in W} |G_i(\alpha) - C(\alpha,\theta^*)|$ and $E_i := \sup_{\alpha\in W} \|\nabla_\alpha G_i(\alpha) - \nabla_\alpha C(\alpha,\theta^*)\|_2$ tend to 0 with probability 1 as $i \to \infty$, where the probability is w.r.t. the data-generating distribution. Moreover, $C(\alpha_i^*,\theta^*) - C(\alpha^*,\theta^*) \to 0$ with probability 1 as $i \to \infty$, where α_i^* is a global minimizer of $G_i(\alpha)$ and α^* is a global minimizer of $C(\alpha,\theta^*)$.*

As the data size N increases, the posterior distribution converges to a Dirac measure, which is a point mass at the true parameter θ^*. Consequently, the performance of the optimal policy for the posterior μ_N converges to the optimal policy under the true environment θ^*, as demonstrated in Theorem 3. Additional assumptions are required to ensure the convergence of a series of posteriors. Broadly speaking, it is necessary that all parameters and all data points have positive probabilities of being sampled under both the prior and posterior distributions, and that the interchangeability of limits and integrals is satisfied. Detailed proof and assumptions for Theorem 3 are provided in Appendix H.6. In the episodic setting, we iteratively use the current policy for data collection and posterior updates, and perform policy updates based on the updated posterior, as described in Algorithm 2. A natural question arises: how many iterations are required within a given episode to achieve a certain level of accuracy? This is addressed in Theorem 4. Notably, the inner map $\lambda(\alpha,\theta^*)$ from policy parameter to occupancy measure is not necessarily convex in α, though $F(\lambda,\theta^*)$ is convex in λ. However, the hidden convexity can be utilized to get the global optimality under the true environment, regardless of the gradient estimation method. By utilizing the local bijection assumption of $\lambda(\cdot,\theta^*)$, a stationary point is still globally optimal, shown by Theorem 5.13 in (Zhang et al., 2021), which requires Assumption H.3. Assumption H.3 can be satisfied when λ is a locally differentiable bijection on a compact convex set W.

Theorem 10. *(Finite-episode error bound) Suppose that Assumption A.1, A.2, B.1, H.1, H.2 and H.3 hold. Assume that $G_i(\alpha)$ defined in Theorem 3 has $L_{G,i}$-Lipschitz continuous gradient. Let*

$\{\alpha_{i,j}\}_{i=1}^{N}{}_{j=0}^{t_i}$ *be the generated policy parameter sequences for* N *episodes by Algorithm 2. For any* $\epsilon > 0$*, if we choose* $t_i = \Theta(L_{G,i}(E_{i-1}+D_i)\epsilon^{-2})$*,* $r = \Theta(\epsilon^{-2})$*, then we can keep a constant gradient bound* $\mathbb{E}\left[\left(\sum_{j=0}^{t_i-1} \|\nabla G_i(\alpha_{i,j})\|_2\right)/t_i\right] \leq \epsilon$ *for each* i*. Furthermore,* $\mathbb{E}C(\alpha_{out}, \theta^*) - C(\alpha^*, \theta^*) \leq \mathcal{O}(\epsilon + E_N)$*, where* D_i, E_i *are defined in Theorem 3.*

Corollary 2. *(Global convergence) Using the same assumptions and notations in Theorem 4,* $\mathbb{E}C(\alpha_{out}, \theta^*) - C(\alpha^*, \theta^*) \leq \mathcal{O}(\epsilon)$ *for any* $\epsilon > 0$ *as* $N \to \infty$*.*

Theorem 4 offers theoretical advice on how to choose the iteration number in each episode. Generally speaking, we need fewer iterations to keep the gradient bound when i grows since D_i, E_i approaches 0. Corollary 1 is a direct result of Theorem 3 and Theorem 4, and Corollary 1 implies global convergence to the true optimal policy in the episodic setting (when the data size N increases to infinity). Detailed proof can be found in Appendix H.7.

C Review on convex RL

Our problem is highly relevant to convex RL, which generalizes cumulative reward on a convex general-utility objective instead of cumulative reward. Specifically, our problem is closely tied to convex RL, which extends the traditional cumulative reward framework to a convex general-utility objective. Prior research has explored policy gradient methods to address convex RL. For instance, (Zhang et al., 2020) demonstrates that the policy gradient of convex RL can be formulated as a min-max optimization problem. To reduce estimator variance, (Zhang et al., 2021) introduces an off-policy policy gradient estimator that leverages mini-batch techniques and truncation mechanisms, while (Barakat et al., 2023) employs a recursive approach to handle large state-action spaces. In the domain of multi-agent convex RL, (Zhang et al., 2022) assumes global state observability and proposes a trajectory-based actor-critic method. Recent studies have also focused on safe convex RL, where the objective is to maximize a convex utility function under convex safety constraints. For example, (Ying et al., 2023) develops a primal-dual algorithm with strong guarantees on the optimality gap and constraint violations, achieving an $\mathcal{O}(1/\epsilon^3)$ bound in the convex-concave case with zero constraint violation. Building on this, (Bai et al., 2023) improves the bound to $\mathcal{O}(1/\epsilon^2)$. Furthermore, (Ying et al., 2024) extends the primal-dual framework to multi-agent convex safe RL.

D Discussion on Non-convex Structure

For any fixed environment parameter θ, the set of occupancy measure is convex, which makes global convergence achievable. But the global convergence cannot be achieved under the Bayesian setting because the set of occupancy measure is nonconvex! Consider the simple case that $\Theta = \{\theta_1, \theta_2\}$ and the state space and action space are finite. Then for any fixed policy parameter α and environment parameter θ_i, the occupancy measure $\lambda(\alpha, \theta_i)$ is a $|S||A|$-dimensional vector. Under any fixed θ_i, for any $t \in [0, 1]$ and policy parameters α_1, α_2, the convex combination $t\lambda(\alpha_1, \theta_i) + (1-t)\lambda(\alpha_2, \theta_i)$ is also an occupancy measure corresponding to some policy α_3 under θ_i. This hidden convexity can be utilized to achieve global convergence for a fixed θ. However, the occupancy measure is a $2|S||A|$-dimensional vector under the Bayesian setting, and we need an additional constraint to guarantee that the occupancy measures under different environments correspond to the same policy α:

$$\frac{\lambda(\alpha, \theta_1)_{s,a}}{\sum_{a'} \lambda(\alpha, \theta_1)_{s,a'}} = \frac{\lambda(\alpha, \theta_2)_{s,a}}{\sum_{a'} \lambda(\alpha, \theta_2)_{s,a'}}, \forall s, a.$$

This constraint means that the agent chooses the action a at state s with the same probability under two environments, which makes the set of occupancy measure nonconvex. In other words, for any $t \in [0, 1], \alpha_1, \alpha_2$, there may not exist a policy α_3 such that $t(\lambda(\alpha_1, \theta_1), \lambda(\alpha_1, \theta_2)) + (1-t)(\lambda(\alpha_2, \theta_1), \lambda(\alpha_2, \theta_2)) = (\lambda(\alpha_3, \theta_1), \lambda(\alpha_3, \theta_2))$. Thus, the global convergence cannot be achieved due to the lack of intrinsic convexity under the Bayesian setting with a fixed posterior. However, as the data size N increases, the posterior distribution converges to a Dirac measure, which is a point mass at the true parameter θ^*. Consequently, the performance of the optimal policy for the posterior μ_N converges to the optimal policy under the true environment, as demonstrated in Theorem 3. What's more, the global optimality gap will converge to any accuracy ϵ when the data size N increases, as

shown in Theorem 4. In other words, global convergence can be achieved in our episodic setting (when the data size N increases to infinity).

E Discussion on Assumption A.2

Recall that $C(\alpha, \theta) := F\left(\lambda^{\pi_\alpha, P_\theta}, P_\theta\right)$. By chain rule, $\nabla_\alpha C(\alpha, \theta) = \nabla_\lambda F \cdot \nabla_\alpha \lambda^{\pi_\alpha}$. Similar to classical Policy Gradient Theorem, it holds that $\nabla_\alpha \lambda^{\pi_\alpha}(s, a) = \lambda^{\pi_\alpha}(s, a) \nabla_\alpha \log \pi_\alpha(a \mid s)$. Thus, the behavior of $\nabla_\alpha C(\alpha, \theta)$ depends on the regularity of both $\nabla_\lambda F$ and $\nabla_\alpha \log \pi_\alpha(a \mid s)$. In classical RL analysis like Agarwal et al. (2021); Papini et al. (2018), one typically assumes that $\nabla_\alpha \log \pi_\alpha(a \mid s)$ is either Lipschitz continuous or uniformly bounded, and per-step rewards $r(s, a)$ are assumed to be bounded for all (s, a) pairs, which together implies the smoothness of the expected return with respect to the policy parameter. In our setting, we replace the usual expected return with a general convex loss function F. Consequently, we must impose smoothness conditions on F. Specifically, if $F(\lambda)$ is Lipschitz continuous with Lipschitz-continuous gradient, and $\nabla_\alpha \log \pi_\alpha(a \mid s)$ is Lipschitz continuous and bounded, then Assumption 3.2 is satisfied.

We can demonstrate that the classic Linear-Quadratic Regulator (LQR), perhaps the simplest continuous-action benchmark, satisfies all of these assumptions.

The dynamics is

$$s_{t+1} = As_t + Ba_t + w_t, \quad w_t \sim \mathcal{N}(0, \Sigma_w)$$

and the policy is Gausssian

$$a_t \sim \pi_\alpha(\cdot \mid s_t) = \mathcal{N}(Ks_t, \Sigma_a)$$

with parameter $\alpha = \mathrm{vec}(K)$. Define the loss function F to be any convex function in λ. Since both the transition and the policy are affine and Gaussian, the joint law (s_t, a_t) is Gaussian at every t. Then the occupancy measure is a discounted summation of Gaussian distributions. Recall that the behavior of $\nabla_\alpha C(\alpha, \theta)$ depends on the regularity of both $\nabla_\lambda F$ and $\nabla_\alpha \log \pi_\alpha(a \mid s)$. In the example of LQR, the first and second derivatives of $\log \pi_\alpha(a \mid s)$ are bounded since the policy is Gaussian. Specifically, if $F(\lambda)$ is Lipschitz continuous with Lipschitz-continuous gradient, and since $\nabla_\alpha \log \pi_\alpha(a \mid s)$ is Lipschitz continuous and bounded in this LQR problem, then Assumptions 3.2 (2)(3) are satisfied.

In the example of LQR, Θ is a set containing possible A, B, Σ_w. Assumptions 3.2 (4) is about the compactness and convexity of Θ, which is not strong. One sufficient condition for Assumption 4.1 to hold is that ξ^* is bounded on Θ. As Θ is a compact and convex set, it is not a strong condition.

F Discussion on Differences between Our Method and Tamar et al. (2015)

We want to obtain $\nabla_\alpha \left[\rho_{\theta \sim \mu_N}(C(\alpha, \theta))\right]$, where the derivative is taken with respect to the policy parameter α in the general loss function $C(\alpha, \theta)$. On the other hand, the problem in Tamar et al. (2015) is how to get $\nabla_\alpha \left[\rho_{\theta \sim \mu_N(\alpha)}(D(\theta))\right]$ for a random variable $D(\theta)$, where the derivative is taken with respect to α in the distribution μ_N. The difference in settings essentially leads to different forms of Lagrangians and causes the failure to apply their results to our setting. What's more, Θ is a finite set in their setting, but Θ can be an uncountable continuous subset of some $\mathbb{R}^d$ in our setting. As a result, when we use the Envelope Theorem to prove the result, we are facing an infinite-dimensional optimization problem over functions, which is a much harder problem than their finite-dimensional optimization problem over vectors. Briefly speaking, in our proof we ensure differentiability and integrability conditions hold uniformly, and construct a separable set of disturbation functions as the domain for Lagrangians.

The original envelope theorem, as presented by Milgrom & Segal (2002), primarily addresses finite-dimensional parameter spaces. Tamar et al. (2015) utilize a discrete-parameter space framework for policy gradients in risk-sensitive Markov Decision Processes (MDPs), restricting their applicability to problems with finite and discrete parameter settings. Our extension generalizes the envelope theorem to handle continuous parameter spaces, significantly broadening the applicability of policy gradient methods to a wider class of problems, such as those involving continuous uncertainty sets or continuous Bayesian posterior distributions over model parameters. Specifically, we address the additional complexities introduced by functional optimization in infinite-dimensional spaces, which involves ensuring differentiability and integrability conditions hold uniformly.

This generalization is nontrivial as it requires overcoming challenges associated with infinite-dimensional optimization, such as ensuring the boundedness and continuity of gradients and validating strong duality under more complex integrative constraints. As such, our contribution facilitates the development of theoretically sound policy gradient methods capable of addressing a broader and more practical range of MDP formulations where uncertainty is represented continuously. This makes our methodology independently interesting to researchers in stochastic control, reinforcement learning, and risk-sensitive optimization.

G Algorithms

Algorithm 1 Bayesian Risk Policy Gradient (BR-PG)

input: Initial α_0, data $\zeta^{(N)}$ of size N, prior distribution $\mu_0(\theta)$, iteration number T;
Calculate the posterior $\mu_N(\theta) = \frac{P_\theta(\zeta^{(N)})\mu_0(\theta)}{\int_\Theta P_{\theta'}(\zeta^{(N)})\mu_0(\theta')d\theta'}$;
for $t = 0$ to $T-1$ **do**
  Sample $\{\theta_k^t\}_{k=1}^r$ from $\mu_N(\theta)$;
  Use the closed-form expression or solve (11) to get $\xi^*(\theta_k^t)$;
  Solve (12) to get $\nabla_\alpha C(\alpha_t, \theta_k^t)$ for $k = 1\ldots, r$;
  $\widehat{g}_t := \frac{1}{r}\sum_{k=1}^r \xi^*(\theta_k^t)\nabla_\alpha C(\alpha, \theta_k^t)$;
  $\alpha_{t+1} = \mathrm{Proj}_W\left(\alpha_t - \frac{1}{\eta_t}\widehat{g}_t\right)$;
end for
output: Choose α_{out} uniformly from $\alpha_0, \ldots, \alpha_{T-1}$.

Algorithm 2 Episodic BR-PG

input: Initial α_0, prior distribution $\mu_0(\theta)$, total episode number N.
Deploy policy $\pi(\alpha_0)$ to gain the initial data set $\zeta^{(1)}$.
for $i = 1$ to N **do**
  if $i = 1$ **then**
    $\alpha_{i,0} = \alpha_0$
  else
    Let $\alpha_{i,0}$ to be uniformly chosen from $\alpha_{i-1,0}, \ldots, \alpha_{i-1,t_{i-1}-1}$;
  end if
  Deploy policy $\pi(\alpha_{i,0})$ to gain a data set $\zeta^{(i)}$.
  Calculate the posterior $\mu_i(\theta) = \frac{P_\theta(\zeta^{(i)})\mu_{i-1}(\theta)}{\int_{\theta'} P_{\theta'}(\zeta^{(i)})\mu_{i-1}(\theta')}$;
  Use Algorithm 1 with t_i iterations and initial point $\alpha_{i,0}$ to generate the policy parameter sequence $\alpha_{i,1}, \cdots, \alpha_{i,t_i}$.
end for
output: Let α_{out} to be randomly and uniformly chosen from $\alpha_{N,0}, \ldots, \alpha_{N,t_N-1}$.

H Proof Details

H.1 Definition of Radon Measure

Definition H.1. *μ_N is a Radon measure on Θ if (i) $\mu_N(\Theta)$ is finite, (ii) for all Borel set $E \subseteq \Theta$, we have $\mu_N(E) = \inf\{\mu_N(U) : E \subseteq U, U$ is open$\}$ and $\mu_N(E) = \sup\{\mu_N(K) : K \subseteq E, K$ is compact$\}$.*

For a continuous parameter space Θ, if the prior is a continuous distribution and the likelihood function is continuous in θ, then the posterior is Radon. And it always holds for discrete cases. Thus it hold in most cases that we may care about, and most common probability distributions are Radon Measures.

H.2 Proof of Theorem 1

Proof.

$$\mathcal{U}(\mu_N) = \{\xi : g_e(\xi, \mu_N) = 0, \forall e \in \mathcal{E}, f_i(\xi, \mu_N) \leq 0, \forall i \in \mathcal{I}, \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta) = 1, \xi(\theta) \geq 0, \|\xi\|_q \leq B_q\}.$$

Define the Lagrangian:

$$L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}}) = \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta)C(\alpha, \theta) - \sum_{i\in\mathcal{I}} \lambda^{\mathcal{I}}(i) f_i(\xi, \mu_N) - \sum_{e\in\mathcal{E}} \lambda^{\mathcal{E}}(e) g_e(\xi, \mu_N), \quad (15)$$

and a subtly relaxed envelope

$$\mathcal{U}'(\mu_N) = \{\xi : \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta) = 1, \xi(\theta) \geq 0, , \|\xi\|_q \leq B_q\}.$$

As mentioned before, we can rewrite the objective as the value of a max-min problem in (8)

$$\rho_{\theta\sim\mu_N}(C(\alpha, \theta)) = \max_{\xi\in\mathcal{U}'(\mu_N)} \min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{E}}} L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}}).$$

Two things deserved to be noticed: (i) Slater's condition holds in the primal optimization problem (8) by Definition A.1. (ii) $L_\theta\left(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}}\right)$ is concave in ξ and convex in $(\lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$. Then strong duality holds for (8).

$$\begin{aligned} \rho_{\theta\sim\mu_N}(C(\alpha, \theta)) &= \max_{\xi\in\mathcal{U}'(\mu_N)} \min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{E}}} L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}}) \\ &= \min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{E}}} \max_{\xi\in\mathcal{U}'(\mu_N)} L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}}) \end{aligned} \quad (16)$$

As $\nabla_\alpha C(\alpha, \theta)$ is uniformly bounded for all θ and α, we have $\nabla_\alpha L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$ is uniformly bounded w.r.t α and continuous at all$(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$. Then we have $L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$ is absolutely continuous w.r.t α for all $(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$. Since $\nabla^2_\alpha C(\alpha, \theta)$ is uniformly bounded for all θ and α, we have $\{L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})\}_{(\xi,\lambda^{\mathcal{I}},\lambda^{\mathcal{E}})}$ is equi-differentiable in α.

As Θ is compact and convex, Θ is a separable metric space with Euclidean metric and its Borel sigma algebra. Then (Θ, μ_N) is a separable metric measure space. By Theorem 4.13 (Brezis & Brézis, 2011), $L^q(\Theta, \mu_N)$ is separable. Then $\mathcal{U}'(\mu_N) = \{\xi \in L^q(\Theta, \mu_N) : \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta) = 1, \xi(\theta) \geq 0, , \|\xi\|_q \leq B_q\}$ is separable.

Define the set of saddle point for (16) by $X^* = \arg\max_{\xi\in\mathcal{U}'(\mu_N)} \min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{E}}} L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$ and $Y^* = \arg\min_{\lambda^{\mathcal{I}}\geq 0, \lambda^{\mathcal{E}}} \max_{\xi\in\mathcal{U}'(\mu_N)} L_\alpha(\xi, \lambda^{\mathcal{I}}, \lambda^{\mathcal{E}})$.

Then for every selection of saddle point $\left(\xi^*_\alpha, \lambda^{*,\mathcal{E}}_\alpha, \lambda^{*,\mathcal{I}}_\alpha\right) \in X^* \times Y^*$, the Envelope theorem for saddle-point problems (Theorem 4(Milgrom & Segal, 2002)) shows that

$$\begin{aligned}\nabla_\alpha \rho_{\theta\sim\mu_N}(C(\alpha,\theta)) =& \nabla_\alpha \max_{\xi\in\mathcal{U}'(\mu_N)} \min_{\lambda^{\mathcal{I}}\geq 0,\lambda^{\mathcal{E}}} L_\alpha(\xi,\lambda^{\mathcal{I}},\lambda^{\mathcal{E}}) \\ =& \nabla_\alpha L_\alpha(\xi,\lambda^{\mathcal{I}},\lambda^{\mathcal{E}})\big|_{(\xi^*_\alpha,\lambda^{*,\mathcal{E}}_\alpha,\lambda^{*,\mathcal{I}}_\alpha)} \\ =& \int_{\theta\in\Theta} \xi^*_\alpha(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)\end{aligned} \tag{17}$$

□

H.3 Proof of Lemma A.1

Proof. Here is a brief proof sketch, and the full proof can be found in the proof of Theorem 3.1(Zhang et al., 2020). For a convex function, the conjugate of the conjugate is itself. Notice that $V(\alpha;z)+\delta\nabla_\alpha V(\alpha;z)^\top x - F^*(z) = \langle z, \lambda(\alpha,\theta)+\delta\nabla_\alpha\lambda(\alpha,\theta)x\rangle - F^*(z)$. Then we have $\sup_{z\in\mathbb{R}^{SA}} V(\alpha;z)+\delta\nabla_\alpha V(\alpha;z)^\top x - F^*(z) = F(\lambda(\alpha,\theta)+\delta\nabla_\alpha\lambda(\alpha,\theta)x)$. By the first-order condition, we have

$$\operatorname*{argmin}_{x\in\mathbb{R}^{SA}} F(\lambda(\alpha,\theta)+\delta\nabla_\alpha\lambda(\alpha,\theta)x) + \frac{\delta}{2}\|x\|_2^2 = -\nabla F\left(\lambda(\alpha,\theta)+\delta\nabla_\alpha\lambda(\alpha,\theta)x\right)\nabla_\alpha\lambda(\alpha,\theta)x).$$

By letting $\delta\to 0^+$ and using the chain rule, we get the result (12). □

H.3.1 Algorithm for solving Theorem A.1

Estimate $V(\alpha,z)$: Recall that we consider an offline setting where the transition kernel P_θ is assumed to be known for any given θ. For any fixed transition kernel P_θ and policy π_α, we can estimate the occupancy measure by making a truncation K in the definition of occupancy measure in (1):

$$\widehat{\lambda^{\pi,P}_{sa}} = \sum_{t=0}^{K}\gamma^t\cdot\mathbb{P}\left(s_t=s,a_t=a\mid\pi,s_0\sim\tau,P\right)$$

with the error $\|\widehat{\lambda}-\lambda\|_1\leq\epsilon_\lambda := \gamma^K/(1-\gamma)$. This error can be made arbitrarily small by increasing K, thus we assume that we can exactly compute occupancy measure. After computing the occupancy measure, $V(\alpha;z)=\langle z,\lambda\rangle$.

Estimate $\nabla_\alpha V(\alpha,z)$: The policy gradient theorem (Sutton et al., 1999) shows that

$$\nabla_\alpha V(\alpha;z) = \mathbb{E}^{\pi_\alpha}\left[\sum_{t=0}^{\infty}\gamma^t Q^{\pi_\alpha}\left(s_t,a_t;z\right)\cdot\nabla_\alpha\log\pi_\alpha\left(a_t\mid s_t\right)\right]$$

where $Q^\pi(s,a;z) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty}\gamma^t z\left(s_t,a_t\right)\mid s_0=s,a_0=a,a_t\sim\pi\left(\cdot\mid s_t\right)\right]$ satisfying the Bellman equation

$$Q^\pi(s,a;z) = \mathbb{E}[z(s,a)] + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} P_\theta(s'|s,a)\pi(a'|s')Q^\pi(s',a';z). \tag{18}$$

For any given θ, policy π and cost function z, we can solve the Bellman equation (18) exactly to get $Q(\cdot,\cdot)$. It can be seen that $\nabla_\alpha V(\alpha;z)$ is a linear function of λ:

$$\nabla_\alpha V(\alpha;z) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} Q(s,a)\dot{\nabla}_\alpha\log\pi_\alpha\left(a\mid s\right)\lambda(s,a).$$

For any θ, policy π and cost function z, we can calculate the Q value function by solving the Bellman equation:

$$Q^\pi(s,a;z) = \mathbb{E}[z(s,a)] + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} P_\theta(s'|s,a)\pi(a'|s')Q^\pi(s',a';z)$$

Then we can use Algorithm 3 to solve (12) in Lemma A.1. It should be noticed that $\delta\nabla_\alpha V(\alpha;z)^\top x = \mathcal{O}(\delta)$ is omitted when calculating the gradient for z as $\delta\to 0$. Thus we omit this term when calculating

the gradient for z. To evaluate the integral $\int_{\theta\in\Theta}\xi_\alpha^*(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)$, we sample i.i.d θ_k from μ_N for $k=1,\ldots,r$, then we can construct the policy gradient estimator

$$\nabla_\alpha \rho_{\theta\sim\mu_N}(C(\alpha,\theta)) \approx \widehat{g}(\alpha) := \frac{1}{r}\sum_{k=1}^{r}\xi^*(\theta_k)\nabla_\alpha C\left(\alpha,\theta_k\right).$$

Algorithm 3 Alternative Gradient Descent Method

input: initial z_0, x_0, step sizes a_t, b_t, iteration number T, transition kernel parameter θ, policy parameter α;
for $t=0$ to $T-1$ **do**
$z_{t+1} = z_t + a_t[\lambda(\alpha,\theta) - \nabla F^*(z_t)]$
$x_{t+1} = x_t - b_t[\nabla_\alpha V(\alpha;z) + x_t]$, where $\nabla_\alpha V(\alpha;z) = \sum_{s,a} Q(s,a)\dot{\nabla}_\alpha \log\pi_\alpha\left(a\mid s\right)\lambda(s,a)$
end for
output: $-x_T$.

H.4 Proof of Theorem 7

Proof. By Theorem 1, the true gradient is

$$g(\alpha) = \int_{\theta\in\Theta}\xi_\alpha^*(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta).$$

And our gradient estimator is

$$\widehat{g}(\alpha) := \frac{1}{r}\sum_{k=1}^{r}\xi^*(\theta_k)\nabla_\alpha C\left(\alpha,\theta_k\right).$$

Then we have

$$\mathbb{E}\|\widehat{g}-g\|_2^2 \le \frac{1}{r}\mathbb{E}\|\xi^*(\theta_1)\nabla_\alpha C\left(\alpha,\theta_1\right) - \int_\Theta \xi^*(\theta)\mu_N(\theta)\nabla_\alpha C(\alpha,\theta)d\theta\|_2^2 \le \frac{\sigma_\xi}{r}.$$

□

H.5 Proof of Theorem 2

First, we make an assumption about G.

Assumption H.1. *There exists some $L_G > 0$ s.t. $g(\alpha)$ is L_G-Lipschitz continuous in α.*

Proof. For ease of notation, denote $g(\alpha_t)$ as g_t and $\widehat{g}(\alpha_t)$ as $\widehat{g}_t$. By Assumption H.1, we have

$$\begin{aligned} G(\alpha) &\le G\left(\alpha_t\right) + \langle g_t, \alpha-\alpha_t\rangle + \frac{L_G}{2}\left\|\alpha-\alpha_t\right\|_2^2 \\ &\le G(\alpha) + L_G\left\|\alpha-\alpha_t\right\|_2^2. \end{aligned} \tag{19}$$

Then we have

$$
\begin{aligned}
G(\alpha_{t+1}) &\leq G\left(\alpha_t\right)+\langle \widehat{g}_t, \alpha_{t+1}-\alpha_t\rangle+\langle g_t-\widehat{g}_t, \alpha_{t+1}-\alpha_t\rangle+\frac{L_G}{2}\left\|\alpha_{t+1}-\alpha_t\right\|_2^2 \\
&\leq G\left(\alpha_t\right)+\langle \widehat{g}_t, \alpha_{t+1}-\alpha_t\rangle+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2+\frac{L_G}{2}\|\alpha_{t+1}-\alpha_t\|_2^2+\frac{L_G}{2}\left\|\alpha_{t+1}-\alpha_t\right\|_2^2 \\
&= G\left(\alpha_t\right)+\langle \widehat{g}_t, \alpha_{t+1}-\alpha_t\rangle+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2+L_G\|\alpha_{t+1}-\alpha_t\|_2^2 \\
&= \min_{\alpha\in W} G(\alpha_t)+\langle \widehat{g}_t, \alpha-\alpha_t\rangle+L_G\|\alpha-\alpha_t\|_2^2+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2 \\
&= \min_{\alpha\in W} G(\alpha_t)+\langle g_t, \alpha-\alpha_t\rangle+L_G\|\alpha-\alpha_t\|_2^2+\langle \widehat{g}_t-g_t, \alpha-\alpha_t\rangle+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2 \\
&\leq \min_{\alpha\in W} G(\alpha_t)+\langle g_t, \alpha-\alpha_t\rangle+L_G\|\alpha-\alpha_t\|_2^2+\frac{L_G}{2}\|\alpha-\alpha_t\|_2^2+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2+\frac{1}{2L_G}\|g_t-\widehat{g}_t\|_2^2 \\
&= \min_{\alpha\in W} G(\alpha_t)+\langle g_t, \alpha-\alpha_t\rangle+\frac{3L_G}{2}\|\alpha-\alpha_t\|_2^2+\frac{1}{L_G}\|g_t-\widehat{g}_t\|_2^2 \\
&= G(\alpha_t)-\frac{1}{6L_G}\|g_t\|_2^2+\frac{1}{L_G}\|g_t-\widehat{g}_t\|_2^2
\end{aligned}
$$

where the first inequality comes from (19), the second inequality comes from Cauchy–Schwarz inequality, the second equality holds because the definition of α_{t+1}, the third inequality holds because of Cauchy–Schwarz inequality again, and the fifth equality holds by taking $\alpha=\alpha_t-\frac{1}{3L_g}g_t$. Telescoping over t, we have

$$
\frac{\sum_{t=0}^{T-1}\|g_t\|_2^2}{T} \leq \frac{6L_G}{T}(G(\alpha_0)-G(\alpha_T))+6\sum_{t=0}^{T-1}\frac{\|g_t-\widehat{g}_t\|_2^2}{T}
$$

Since

$$
\mathbb{E}\left[\|g_t-\widehat{g}_t\|_2^2\right] \leq \frac{\sigma_\xi}{r}
$$

Then we have

$$
\mathbb{E}\|g_{\text{out}}\|_2^2 \leq \frac{6L_G}{T}(G(\alpha_0)-\mathbb{E}G(\alpha_T))+6\frac{\sigma_\xi}{r}
$$

Let $T=\epsilon^{-2}, r=\epsilon^{-2}$, then

$$
\mathbb{E}\|g_{\text{out}}\|_2^2=\mathcal{O}(\epsilon^2)
$$

and then

$$
\mathbb{E}\|g_{\text{out}}\|_2=\mathcal{O}(\epsilon)
$$

□

H.6 Proof of Theorem 3

Assumption H.2. *(Assumption 3.1 in (Shapiro et al., 2023))*

(1) The set Θ is convex compact with nonempty interior.

(2) $\ln\mu_0(\theta)$ is bounded on Θ, i.e., there are constants $c_1>c_2>0$ such that $c_1\geq\mu_0(\theta)\geq c_2$ for all $\theta\in\Theta$.

(3) $P^*(\zeta)>0$ for any $\zeta\in\Xi$.

(4) $P_\theta(\zeta)>0$, and hence $\mu_N(\theta)>0$, for all $\xi\in\Xi$ and $\theta\in\Theta$.

(5) $P_\zeta(\xi)$ is continuous in $\theta\in\Theta$.

(6) $\ln P_\theta(\zeta), \theta\in\Theta$, is dominated by an integrable (w.r.t. P_*) function.

Assumption H.2 (1), (2) are used to guarantee the uniform convergence of posterior. Assumption H.2 (3), (4) require that all data points has positive probability to be sampled under the prior and posterior. Assumption H.2 (5), (6) are used to exchange the order of limit and integral.

With Assumption H.2, we are now ready to prove Theorem 3. Define a function $\psi(\theta) = \mathbb{E}_{P^*}[\ln P_\theta(\xi)]$ and let $\Theta^* := \{\theta' \in \Theta : \psi(\theta') = \inf_{\theta\in\Theta} \psi(\theta)\}$. For $\epsilon > 0$, define sets

$$V_\epsilon := \{\theta \in \Theta : \psi(\theta^*) - \psi(\theta) \geq \epsilon\}, U_\epsilon := \Theta \backslash V_\epsilon = \{\theta \in \Theta : \psi(\theta^*) - \psi(\theta) < \epsilon\}.$$

First we need to show two intermediate lemmas.

Lemma H.1. *(Lemma 3.1. (Shapiro et al., 2023)) Suppose that Assumption H.2 holds. Then for $0 < \epsilon_2 < \epsilon_1 < \epsilon_0$, it follows that w.p. 1 for N large enough*

$$\sup_{\theta\in V_{\epsilon_0}} \mu_N(\theta) \leq \kappa(\epsilon_2)^{-1} e^{-N(\epsilon_1-\epsilon_2)},$$

where V_{ϵ_0} and U_{ϵ_0} are defined in (3.2), and $\kappa(\epsilon_2) := \int_{U_{\epsilon_2}} d\theta$.

Lemma H.2. *Suppose that Assumption H.2 holds. $\forall \delta > 0$, $\exists \epsilon > 0$ such that $d(\theta, \Theta^*) < \delta$ for all $\theta \in U_\epsilon$.*

Proof. We prove this lemma by contradiction. Suppose that $\exists \delta_0 > 0$ such that $\forall \epsilon > 0$, there exists $\theta \in \Theta$ satisfying $\psi(\theta^*) - \psi(\theta) < \epsilon$ and $d(\theta, \Theta^*) \geq \delta_0$.

Choose $\epsilon = \frac{1}{n}$ and then get a sequence $\{\theta_n\}_{n=1}^\infty$. As Θ is compact, there exists a subsequence of $\{\theta_n\}_{n=1}^\infty$ that converge to a point $\theta' \in \Theta$ satisfying $d(\theta', \Theta^*) \geq \delta_0$. As ψ is continuous, $\psi(\theta') = \psi(\theta^*)$. Contradiction! □

Then we can prove Theorem 3

Proof. For any $\delta > 0$, we can choose ϵ_0 such that $d(\theta, \Theta^*) \leq \delta$ for $\theta \in U_{\epsilon_0}$. Then we have

$$\begin{aligned}
&|\rho_{\theta\sim\mu_N}(C(\alpha,\theta)) - C(\alpha,\theta^*)| \\
&= |\max_{\xi:\xi\in\mathcal{U}(\mu_N)} \int_{\theta\in\Theta} \xi(\theta)\mu_N(\theta)[C(\alpha,\theta) - C(\alpha,\theta^*)]d\theta| \\
&\leq \max_{\xi:\xi\in\mathcal{U}(\mu_N)} \int_{U_{\epsilon_0}} \xi(\theta)\mu_N(\theta)|C(\alpha,\theta) - C(\alpha,\theta^*)|d\theta + \max_{\xi:\xi\in\mathcal{U}(\mu_N)} \int_{V_{\epsilon_0}} \xi(\theta)\mu_N(\theta)|C(\alpha,\theta) - C(\alpha,\theta^*)|d\theta \\
&\leq \sup_{\|\theta-\theta^*\|\leq\delta} |C(\alpha,\theta) - C(\alpha,\theta^*)| + 2\sup_{\alpha\in W,\theta\in\Theta} |C(\alpha,\theta)| \max_{\xi:\xi\mu_N\in\mathcal{U}(\mu_N)} \int_{V_\epsilon} \xi(\theta)\mu_N(\theta)d\theta
\end{aligned}$$

By Holder's Inequality, we have

$$\begin{aligned}
\int_{V_\epsilon} \xi(\theta)\mu_N(\theta)d\theta &= \int_{V_\epsilon} \xi(\theta)\mu_N(\theta)^{1/q}\mu_N(\theta)^{1/p}d\theta \\
&\leq \left[\int_{V_\epsilon} \xi(\theta)^q\mu_N(\theta)d\theta\right]^{1/q} \left[\int_{V_\epsilon} \mu_N(\theta)d\theta\right]^{1/p} \\
&\leq B_q\kappa(\epsilon_2)^{-1/p} e^{-N(\epsilon_1-\epsilon_2)/p} \cdot Vol(\Theta)^{1/p}
\end{aligned}$$

Thus

$$|\rho_{\theta\sim\mu_N}(C(\alpha,\theta)) - C(\alpha,\theta^*)| \leq \delta L_\theta + 2B_q\kappa(\epsilon_2)^{-1/p} e^{-N(\epsilon_1-\epsilon_2)/p} Vol(\Theta)^{1/p} \sup_{\alpha\in W,\theta\in\Theta} |C(\alpha,\theta)|,$$

where the inequality holds because $C(\alpha,\theta)$ is $B-$ Lipschitz continuous w.r.t. θ. This implies $D_N \to 0$ as $N \to \infty$ since δ is arbitrary. Then we have

$$\begin{aligned}
&C(\alpha_N^*,\theta^*) - C(\alpha^*,\theta^*) \\
&= C(\alpha_N^*,\theta^*) - \rho_{\theta\sim\mu_N}(C(\alpha_N^*,\theta)) + \rho_{\theta\sim\mu_N}(C(\alpha_N^*,\theta)) - \rho_{\theta\sim\mu_N}(C(\alpha^*,\theta)) + \rho_{\theta\sim\mu_N}(C(\alpha^*,\theta)) - C(\alpha^*,\theta^*) \\
&\leq 2\delta B + 4B_q\kappa(\epsilon_2)^{-1/p} e^{-N(\epsilon_1-\epsilon_2)/p} Vol(\Theta)^{1/p} \sup_{\alpha\in W,\theta\in\Theta} |C(\alpha,\theta)|,
\end{aligned}$$

Let $N \to \infty$ and recall that δ is arbitary, we get the result.

Define

$$E_N := \sup_{\alpha \in W} \|\nabla_\alpha \rho_{\theta \sim \mu_N}(C(\alpha, \theta)) - \nabla_\alpha C(\alpha, \theta^*)\|_2 .$$

Similar to D_N, we have

$$\begin{aligned}
&\|\nabla_\alpha \rho_{\theta \sim \mu_N}(C(\alpha, \theta)) - \nabla_\alpha C(\alpha, \theta^*)\|_2 \\
&= \left\| \int_\Theta \xi_\alpha^*(\theta) \mu_N(\theta) [\nabla_\alpha C(\alpha, \theta) - \nabla_\alpha C(\alpha, \theta^*)] d\theta \right\|_2 \\
&\leq \int_{U_{\epsilon_0}} \xi_\alpha^*(\theta) \mu_N(\theta) \|\nabla_\alpha C(\alpha, \theta) - \nabla_\alpha C(\alpha, \theta^*)\|_2 d\theta + \int_{V_{\epsilon_0}} \xi_\alpha^*(\theta) \mu_N(\theta) \|\nabla_\alpha C(\alpha, \theta) - \nabla_\alpha C(\alpha, \theta^*)\|_2 d\theta \\
&\leq \delta L_{\theta,2} + 2B_q \kappa(\epsilon_2)^{-1/p} e^{-N(\epsilon_1 - \epsilon_2)/p} Vol(\Theta)^{1/p} \sup_{\alpha \in W, \theta \in \Theta} \|\nabla_\alpha C(\alpha, \theta)\|,
\end{aligned}$$

which implies $E_N \to 0$ as $N \to \infty$ since δ is arbitrary.

□

H.7 Proof of Theorem 4 and Corollary 1

Assumption H.3. *(Assumption 5.11 in (Zhang et al., 2021)) For policy parameterization π_α, α overparametrizes the set of policies in the following sense. (i). For any α and $\lambda(\alpha)$ under the true environment P_{θ^*}, there exist (relative) neighourhoods $\alpha \in \mathcal{U}_\alpha \subset W$ and $\lambda(\alpha) \in \mathcal{V}_{\lambda(\alpha)} \subset \lambda(W, \theta^*)$ s.t. $(\lambda|_{\mathcal{U}_\alpha})(\cdot)$ forms a bijection between $\mathcal{U}_\alpha$ and $\mathcal{V}_{\lambda(\alpha)}$, where $(\lambda \mid \mathcal{U}_\alpha)(\cdot)$ is the confinement of λ onto $\mathcal{U}_\alpha$. We assume $(\lambda \mid \mathcal{U}_\alpha)^{-1}(\cdot)$ is ℓ_α-Lipschitz continuous and $(\lambda \mid \mathcal{U}_\alpha)(\cdot)$ is L_λ-Lipschitz smooth for any α . (ii). Let π_{α^*} be the optimal policy under the true environment. Assume there exists $\bar{\epsilon}$ small enough, s.t. $(1-\epsilon)\lambda(\alpha) + \epsilon\lambda(\alpha^*) \in \mathcal{V}_{\lambda(\alpha)}$ for $\forall \epsilon \leq \bar{\epsilon}, \forall \alpha$.*

Under the true environment P_{θ^*}, the set of all occupancy measures is a convex set, and there is an bijection between all policies and all occupancy measures. More discussions can be found in Section 5.2 in (Zhang et al., 2021). Based on this observation, we have $\min_\pi F(\lambda^\pi, \theta^*) = \min_\lambda F(\lambda, \theta^*)$, which turns the non-convex policy optimization problem into a convex optimization problem over occupancy measure. Then any stationary point is also globally optimal, which is shown in the following lemma.

Lemma H.3. *Assume that Assumption H.3 holds. Then $C(\bar{\alpha}, \theta^*) - C(\alpha^*, \theta^*) \leq \mathcal{O}(\|\nabla_\alpha C(\bar{\alpha}, \theta^*)\|_2), \forall \bar{\alpha} \in W$.*

Proof. Notice that

$$\partial(F \circ \lambda)(\alpha) = \nabla_\alpha \lambda(\alpha)^\top \partial F(\lambda), \forall \alpha \in W.$$

So there exists $\bar{w} \in \partial F(\bar{\lambda}, \theta^*)$ s.t. $\nabla_\alpha C(\bar{\alpha}, \theta^*) = \nabla_\alpha \lambda(\bar{\alpha})^\top \bar{w}$. Then for any $\lambda(\alpha) \in \mathcal{V}_{\lambda(\bar{\alpha})}$, we have

$$\begin{aligned}
&\langle \bar{w}, \lambda - \bar{\lambda} \rangle \\
&= \langle \bar{w}, \nabla_\alpha \lambda(\bar{\alpha})(\alpha - \bar{\alpha}) \rangle + \langle \bar{w}, \lambda - \bar{\lambda} - \nabla_\alpha \lambda(\bar{\alpha})(\alpha - \bar{\alpha}) \rangle \\
&= I_1 + I_2
\end{aligned} \tag{20}$$

For the first term, we have

$$I_1 = \langle \nabla_\alpha C(\bar{\alpha}, \theta^*), \alpha - \bar{\alpha} \rangle \geq -\|\nabla_\alpha C(\bar{\alpha}, \theta^*)\|_2 \|\alpha - \bar{\alpha}\|_2 \geq -\|\nabla_\alpha C(\bar{\alpha}, \theta^*)\|_2 \ell_\alpha \|\lambda - \bar{\lambda}\|_2.$$

For the second term, we have

$$I_2 \geq -\|\bar{w}\|_2 \cdot \frac{L_\lambda}{2} \|\alpha - \bar{\alpha}\|_2^2 \geq -\frac{L_\lambda \ell_\alpha^2}{2} \|\bar{w}\|_2 \|\lambda - \bar{\lambda}\|_2^2$$

Then we have

$$\langle \bar{w}, \lambda - \bar{\lambda} \rangle \geq -\|\nabla_\alpha C(\bar{\alpha}, \theta^*)\|_2 \ell_\alpha \|\lambda - \bar{\lambda}\|_2 - \frac{L_\lambda \ell_\alpha^2}{2} \|\bar{w}\|_2 \|\lambda - \bar{\lambda}\|_2^2.$$

Replace λ by $(1-\epsilon)\lambda(\bar{\alpha})+\epsilon\lambda(\alpha^*)$ for any $\epsilon\in(0,\bar{\epsilon}]$ and then it holds

$$\epsilon\langle\bar{w},\lambda(\alpha^*)-\lambda(\bar{\alpha})\rangle\geq-\epsilon\|\nabla_\alpha C(\bar{\alpha},\theta^*)\|_2\ell_\alpha\|\lambda(\alpha^*)-\lambda(\bar{\alpha})\|_2-\frac{L_\lambda\epsilon^2\ell_\alpha^2}{2}\|\bar{w}\|_2\|\lambda(\alpha^*)-\lambda(\bar{\alpha})\|_2^2.$$

Divide both sides by ϵ and let $\epsilon\to 0$, we have

$$\langle\bar{w},\lambda(\alpha^*)-\lambda(\bar{\alpha})\rangle\geq-\|\nabla_\alpha C(\bar{\alpha},\theta^*)\|_2\ell_\alpha\|\lambda(\alpha^*)-\lambda(\bar{\alpha})\|_2.$$

Finally,

$C(\bar{\lambda},\theta^*)-C(\lambda^*,\theta^*)\leq-\langle\bar{w},\lambda(\alpha^*)-\lambda(\bar{\alpha})\rangle\leq\ell_\alpha\|\nabla_\alpha C(\bar{\alpha},\theta^*)\|_2\|\lambda(\alpha^*)-\lambda(\bar{\alpha})\|_2\leq\ell_\alpha D_\lambda\|\nabla_\alpha C(\bar{\alpha},\theta^*)\|_2$,
where $D_\lambda=\sup_{\alpha,\alpha'\in W}\|\lambda(\alpha)-\lambda(\alpha')\|_2$. □

Then we can turn to prove Theorem 4.

Proof. If $\mathbb{E}\left[\frac{\sum_{j=0}^{t_i-1}\|\nabla G_i(\alpha_{i,j})\|}{t_i}\right]\leq\epsilon$, choose $\alpha_{i+1,0}$ uniformly from $\alpha_{i,0},\ldots,\alpha_{i,t_i-1}$. Then $\mathbb{E}\|\nabla G_i(\alpha_{i+1,0})\|_2\leq\epsilon$ and then

$$\mathbb{E}\|\nabla C(\alpha_{i+1,0},\theta^*)\|_2\leq\epsilon+E_i.$$

By Lemma H.3 we have

$$\mathbb{E}C(\alpha_{i+1,0},\theta^*)-C(\alpha^*,\theta^*)\leq(\epsilon+E_i)\ell_\alpha D_\lambda,$$

and then

$$\mathbb{E}G_{i+1}(\alpha_{i+1,0})-G_{i+1}(\alpha_{i+1}^*)\leq(\epsilon+E_i)\ell_\alpha D_\lambda+2D_{i+1}.$$

By Theorem 2 we have

$$\mathbb{E}\left[\frac{\sum_{j=0}^{t_{i+1}-1}\|\nabla G_{i+1}(\alpha_{i+1,j})\|_2^2}{t_{i+1}}\right]\leq\frac{6L_{G,i+1}}{t_{i+1}}[(\epsilon+E_i)\ell_\alpha D_\lambda+2D_{i+1}]+6\frac{\sigma_\xi}{r}$$

Then we can choose

$$t_{i+1}=12L_{G,i+1}[(\epsilon+E_i)\ell_\alpha D_\lambda+2D_{i+1}]\epsilon^{-2}$$
$$r=12\sigma_\xi\epsilon^{-2}$$

to make

$$\mathbb{E}\left[\frac{\sum_{j=0}^{t_{i+1}-1}\|\nabla G_{i+1}(\alpha_{i+1,j})\|_2^2}{t_{i+1}}\right]\leq\epsilon^2.$$

Then by Jensen's inequality we have

$$\mathbb{E}\left[\frac{\sum_{j=0}^{t_{i+1}-1}\|\nabla G_{i+1}(\alpha_{i+1,j})\|_2}{t_{i+1}}\right]\leq\epsilon.$$

When $\mathbb{E}\|\nabla G_N(\alpha_{\text{out}})\|_2\leq\epsilon$, we have $\|\nabla C(\alpha_{\text{out}},\theta^*)\|_2\leq\epsilon+E_N$. Then by Lemma H.3 we complete the proof of Theorem 4.

As Theorem 3 shows that $E_N\to\infty$ when $N\to\infty$, then we complete the proof of Corollary 1. □

I Examples of Loss Function

Example 1 (Risk-Averse Constrained MDP)**.** *In safe RL problems, one usually considers a constrained MDP (Altman, 2021), where the goal is to minimize the total expected discounted cost under a risk-averse constraint. Given a random vector penalty d, the risk-averse constraint is to control a risk measure of the total expected discounted penalty. This leads to the following constrained MDP formulation:*

$$\min_\pi\ \mathbb{E}[\sum_{t=0}^\infty\gamma^t c(s_t,a_t)\mid\pi,s_0\sim\tau]\quad\textit{s.t. }\rho\left(\mathbb{E}[\sum_{t=0}^\infty\gamma^t d(s_t,a_t)\mid\pi,s_0\sim\tau]\right)\leq D,$$

where ρ is a coherent risk measure, such as Conditional Value-at-Risk (CVaR)[1] *Using Lagrangian relaxation, we can choose F to be a convex function of λ, i.e., $F(\lambda,P)=\langle\lambda,c\rangle+\ell(\rho(\langle\lambda,d\rangle)-D)$, where ℓ is the Lagrange multiplier.*

[1] $\text{CVaR}(X)=\mathbb{E}[X|X\geq v_\beta(X)]$, where $v_\beta(X)$ is a β-quantile of X, i.e. $\mathbb{P}(X\geq v_\beta(X))=1-\beta$

Example 2 (Imitation Learning)**.** *During imitation learning, the agent learn through some demonstrations to behave similarly to an expert. One formulation is minimize the $f-$divergence between the occupancy measure of the current policy and the target occupancy measure:*

$$\min_{\pi} D_f(\lambda^{\pi}, q) = \sum_{s,a} q(s,a) f\left(\frac{\lambda^{\pi}(s,a)}{q(s,a)}\right)$$

J Examples of Risk Envelop

Example 3. *[Conditional Value at Risk] First, Value-at-risk $\mathrm{VaR}_{\beta}(X)$ is defined as the β-quantile of X, i.e., $\mathrm{VaR}_{\beta}(X) := \inf\{t : \mathbb{P}(X \leq t) \geq \beta\}$, where the confidence level $\beta \in (0,1)$. Assuming there is no probability atom at $\mathrm{VaR}_{\beta}(X)$, CVaR at confidence level β is defined as the mean of the β-tail distribution of X, i.e., $\mathrm{CVaR}_{\beta}(X) = \mathbb{E}\left[X \mid X \geq \mathrm{VaR}_{\beta}(X)\right]$. The envelope set is*

$$\mathcal{U}(\mu_N) = \{\xi \in \mathcal{Z}^* : \int_{\Theta} \xi(\theta)\mu_N(\theta)d\theta = 1, \xi(\theta) \in \left[0, \frac{1}{1-\beta}\right] a.s. \theta \in \Theta\}$$

Example 4. *(Mean-Upper-Semideviation of Order $\mathbf{p}$). For $\mathcal{Z} := \mathcal{L}_p(\Theta, \mathcal{F}, \mu_N)$ and $\mathcal{Z}^* := \mathcal{L}_q(\Theta, \mathcal{F}, \mu_N)$, with $p \in [1, +\infty)$, $c \in [0,1]$ and $\mathcal{F}$ to be a σ-field on Θ, consider*

$$\rho(Z) := \mathbb{E}[Z] + c\left(\mathbb{E}\left[[Z - \mathbb{E}[Z]]_+^p\right]\right)^{1/p},$$

where $[a]_+^p = \max\{0, a\}^p$. Then the envelope set is

$$\mathcal{U}(\mu_N) = \{\xi' \in \mathcal{Z}^* : \xi' = 1 + \xi - \mathbb{E}[\zeta], \|\xi\|_q \leq c, \xi \succeq 0\}\}.$$

More examples can be found in Section 6.3.2(Shapiro et al., 2021).

K Policy Gradient for MDP with CVaR Risk Measure : A Special Case Study

Here we offer an example of gradient estimator with a common coherent risk measure Conditional Value at Risk(CVaR), the definition of which can be found in Example 3. For the considered CVaR risk functional, (Hong & Liu, 2009) shows that the gradient of the CVaR risk functional can be expressed as

$$\nabla \mathrm{CVaR}_{\beta}(X(\alpha)) = \mathbb{E}[\nabla X(\alpha) | X(\alpha) \geq v_{\beta}(\alpha)]$$

where $v_{\beta} = v_{\beta}(\alpha) := \mathrm{VaR}_{\beta}(X(\alpha))$ for a random parameterized variable $X(\alpha)$ satisfying Assumption K.1. Unless otherwise specified, the derivative is assumed to be taken w.r.t. α.

Assumption K.1. *(Assumption 1, 2, 3 (Hong & Liu, 2009)) (i) There exists a random variable L with $\mathbb{E}(K) < \infty$ such that $|X(\alpha_2) - X(\alpha_1)| \leq K \|\alpha_2 - \alpha_1\|_2$ for all $\alpha_1, \alpha_2 \in W$, and $\nabla_{\alpha} X(\alpha)$ exists almost surely for all $\alpha \in W$.*

(ii) VaR function $v_{\beta}(\alpha)$ is differentiable for any $\alpha \in W$.

(iii) For any $\alpha \in W, \mathbb{P}(X(\alpha) = v_{\beta}(\alpha)) = 0$.

Assumption K.1 (i) is commonly used in path-wise derivative estimation; (ii) shows that VaR function is locally Lipschitz; (iii) requires that there is no probability atom at $VaR(X)$ and implies that $\mathbb{P}(X(\alpha) \geq v_{\beta}(\alpha)) = 1 - \beta$.

Theorem 11. *Suppose that Assumption K.1 holds. Then, for any $\alpha \in W$ and $\beta \in (0,1)$, the policy gradient to the objective function in* (3) *is given by:*

$$\begin{aligned} g(\alpha) &= \mathbb{E}_{\theta \sim \mu_N}\left[\nabla C(\alpha, \theta) \mid C(\alpha, \theta) \geq v_{\beta}(\alpha)\right] \\ &= \frac{1}{1-\beta} \mathbb{E}_{\theta \sim \mu_N}\left[\nabla C(\alpha, \theta) \mathbb{1}_{\{C(\alpha,\theta) \geq v_{\beta}\}}\right] \end{aligned} \tag{21}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

If we apply Theorem 1 to CVaR, we will get the same result as Theorem11. To compute the gradient $g(\alpha)$, we require the cumulative value $C(\alpha,\theta)$ of policy π_α and its gradient $\nabla C(\alpha,\theta)$, value-at-risk v_β, as well as the evaluation of the expectation taken w.r.t. the posterior distribution μ_N. Here we show how to use zeroth-order method instead of variational approach to estimate $\nabla_\alpha C(\alpha,\theta)$. Since there is no closed-form expression for the expectation, we estimate the gradient $g(\alpha)$ with samples $\{\theta^i\}_{i=1}^n$ generated from μ_N. We construct the gradient estimator as follows:

$$\widehat{g}(\alpha) = \frac{1}{n(1-\beta)} \sum_{i=1}^{n} \widehat{\nabla C}(\alpha,\theta^i) \mathbb{1}_{\{\widehat{C}(\alpha,\theta^i) \geq \widehat{v}_\beta\}}. \tag{22}$$

For a fixed α and θ^i, we first estimate the occupancy measure λ^i by making a truncation of horizon K in (1) with error

$$\|\widehat{\lambda}^i - \lambda^i\|_\infty \leq \epsilon_\lambda := \gamma^K/(1-\gamma) \tag{23}$$

for some $K > 0$. The cumulative value with the truncated occupancy measure $\widehat{\lambda}^i$ is denoted by $\widehat{C}(\alpha,\theta^i) = F(\widehat{\lambda}, P_{\theta^i})$. The value-at-risk estimate is $\widehat{v}_\beta := \widehat{C}(\alpha,\theta)_{\lceil n\beta \rceil:n}$, where $\widehat{C}(\alpha,\theta)_{\lceil n\beta \rceil:n}$ is the $\lceil n\beta \rceil$-th smallest quantity in $\{\widehat{C}(\alpha,\theta^i)\}_{i=1}^n$.

Here we adopt the Gaussian smoothing approach of estimating gradients from function evaluations (Nesterov & Spokoiny, 2017; Balasubramanian & Ghadimi, 2022). When there is no oracle to the first-order information or it is not efficient to calculate the gradient directly, Gaussian smoothing approach is a useful technique in zeroth-order method. Compared with finite difference method, Gaussian smoothing approach requires weaker smoothness condition of objective function. For a fixed α and θ^i, generate $\{u^{i,j}\}_{j=1}^{m_i}$, where $u^{i,j} \sim \mathcal{N}(0, I_d)$. Then $\widehat{\nabla C}$ can be constructed as:

$$\widehat{\nabla C}(\alpha,\theta^i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{\widehat{C}\left(\alpha + \nu u^{i,j}, \theta^i\right) - \widehat{C}\left(\alpha,\theta^i\right)}{\nu} u^{i,j} \tag{24}$$

where $\nu > 0$ is the smoothing parameter.

For ease of notation, let $\widehat{G}(\alpha)$ denote the sample estimate of $\rho_{\theta\sim\mu_N}(C(\alpha,\theta))$. We use the following gradient descent step in the t-th iteration:

$$\begin{aligned} \alpha_{t+1} &= \arg\min_{\alpha\in W} \widehat{G}(\alpha_t) + \langle \widehat{g}(\alpha_t), \alpha - \alpha_t \rangle + \frac{\eta_t}{2}\|\alpha - \alpha_t\|^2 \\ &= \mathrm{Proj}_W\left(\alpha_t - \frac{1}{\eta_t}\widehat{g}(\alpha_t)\right) \end{aligned} \tag{25}$$

where η_t is the stepsize and $\mathrm{Proj}_W(x) = \arg\min_{y\in W} \|y - x\|_2^2$ projects x into the parameter space W. We summarize the full algorithm in Algorithm 4.

K.1 Convergence Analysis for CVaR Risk Measure

Here we only show the estimation error of the policy gradient. To get a finite-step convergence result similar to Theorem 2, we only need to substitute $\mathcal{O}(r^{-1/4})$ in Theorem 2 with $\mathcal{O}(R^{1/2})$, where $R^2 = \mathcal{O}\left(dn^{-1} + \epsilon_\lambda + \frac{d\epsilon_\lambda^2}{\nu^2} + \frac{d+\nu^2 d^3}{m}\right)$ is the bound for $\mathbb{E}\|[g - \widehat{g}]\|_2^2]$ in Theorem 12.

Here we still adopt the Assumption A.2 about the smoothness for the considered loss functions, which are commonly used in gradient descent analysis. The error bound for the zeroth-order estimation for ∇C is then shown in the next lemma.

Lemma K.1. *Suppose Assumption K.1 and Assumption A.2 hold. Then we have for each $i \in [n]$*

$$\begin{aligned} \mathbb{E}\|\widehat{\nabla C}(\alpha,\theta_i) - \nabla C(\alpha,\theta_i)\|_2^2 &\leq \frac{8d}{\nu^2} L_{F,\infty}^2 \epsilon_\lambda^2 \\ &+ \frac{8(d+5)B^2}{m_i} + \frac{2\nu^2 L_{C,2}^2 (d+6)^3}{m_i}, \end{aligned} \tag{26}$$

where $L_{F,\infty}, L_{C,2}, B$ are constants in Assumption A.2, ϵ_λ is the truncation error defined in (23)*, d is the dimension of the policy parameter α, m_i is the number of samples used to construct the zeroth-order estimator in* (24)*.*

Algorithm 4 BR-PG: Bayesian Risk Policy Gradient for CVaR

input: initial α_0, data $\zeta^{(N)}$ of size N, prior distribution $\mu_0(\theta)$, iteration number T, truncation horizon K;
calculate the posterior $\mu_N(\theta) = \frac{P_\theta(\zeta^{(N)})\mu_0(\theta)}{\int_{\theta'} P_{\theta'}(\zeta^{(N)})\mu_0(\theta')}$;
for $t = 0$ to $T-1$ **do**
  sample $\{\theta_t^i\}_{i=1}^n$ from $\mu_N(\theta)$;
  for $i = 1$ to n **do**
    calculate $\widehat{\lambda}_t^i$ using the truncation of horizon K specified in (1);
    calculate $\widehat{C}(\alpha_t, \theta_t^i) := F(\widehat{\lambda}_t^i, P_{\theta_t^i})$;
    generate $\{u^{i,j}\}_{j=1}^{m_i}$, where $u^{i,j} \sim \mathcal{N}(0, I_d)$;
    calculate $\widehat{\nabla C}(\alpha_t, \theta_t^i)$ by (24);
  end for
  calculate $\widehat{v}_\beta(\alpha_t) := \widehat{C}(\alpha_t, \theta_t^i)_{\lceil n\beta \rceil : n}$.
  calculate $\widehat{g}(\alpha_t)$ by (22);
  update α_{t+1} by (6).
end for
output: α_T.

Assumption K.2. *(Assumptions 4 and 5 in (Hong & Liu, 2009)*

(1) For all $\alpha \in W$, $C(\alpha, \theta)$ is a continuous random variable with a density function $f_{C,\alpha}(y)$. Furthermore, $f_{C,\alpha}(y)$ and $g_{C,\alpha}(y) := \mathbb{E}_\theta[\nabla C(\alpha, \theta) \mid C(\alpha, \theta) = y]$ are continuous at $y = v_\alpha$, and $f_{C,\alpha}(v_\alpha) > 0$.

(2) $\mathbb{E}_\theta\left[C(\alpha, \theta)^2\right] < \infty$ for all $\alpha \in W$.

Now we are ready to show the error for our gradient estimator given in (22).

Theorem 12. *Suppose that Assumption K.1, Assumption A.2 and Assumption K.2 hold. Also assume that the cumulative distribution function of $C(\alpha, \theta)$ w.r.t θ is ℓ_C- Lipschitz continuous for each $\alpha \in W$. Let $m_i = m \; \forall i \in [n]$. Then for each $\alpha \in W$,*

$$\mathbb{E}\|[g - \widehat{g}]\|_2^2 \leq \mathcal{O}\left(dn^{-1} + \epsilon_\lambda + \frac{d\epsilon_\lambda^2}{\nu^2} + \frac{d + \nu^2 d^3}{m}\right),$$

where n is the number of samples of θ.

Proof. First recall that the true gradient and our gradient estimator are $g = \frac{1}{1-\beta}\mathbb{E}\left[\nabla C(\alpha, \theta)\mathbb{1}_{\{C(\alpha,\theta) \geq v_\beta\}}\right]$ and $\widehat{g} = \frac{1}{n(1-\beta)}\sum_{i=1}^n \widehat{\nabla C}(\alpha, \theta_i)\mathbb{1}_{\{\widehat{C}(\alpha,\theta_i) \geq \widehat{v}_\beta\}}$. Let

$$\tilde{g} = \frac{1}{n(1-\beta)}\sum_{i=1}^n \nabla C(\alpha, \theta_i)\mathbb{1}_{\{C(\alpha,\theta_i) \geq \tilde{v}_\beta\}},$$

and

$$\widehat{g_1} = \frac{1}{n(1-\beta)}\sum_{i=1}^n \nabla C(\alpha, \theta_i)\mathbb{1}_{\{\widehat{C}(\alpha,\theta_i) \geq \widehat{v}_\beta\}},$$

where $\tilde{v}_\beta := C(\alpha, \theta_i)_{\lceil n\beta \rceil : n}$. Then we have the decomposition $g - \widehat{g} = (g - \tilde{g}) + (\tilde{g} - \widehat{g_1}) + (\widehat{g_1} - \widehat{g}) := R_1 + R_2 + R_3$. For R_1, it is the error in the estimation of expectation taken w.r.t. θ. Suppose that Assumption K.1 and Assumption K.2 hold, Theorem 4.2 from (Hong & Liu, 2009) shows that

$$\|\mathbb{E}R_1\|_2 = \|\mathbb{E}[\tilde{g}] - g\|_2 = o(n^{-1/2}d^{-1/2}).$$

Notice that

$$\|g - \tilde{g}\|_2^2 \leq 2\|g - \mathbb{E}\tilde{g}\|_2^2 + 2\|\mathbb{E}\tilde{g} - \tilde{g}\|_2^2.$$

By Theorem 4.3 from (Hong & Liu, 2009), $Var(\tilde{g}) = \mathcal{O}(dn^{-1})$. Thus

$$\mathbb{E}\|R_1\|_2^2 = \mathcal{O}(dn^{-1}). \tag{27}$$

For R_3, it is the error in the estimation of $C(\alpha,\theta)$. By Lemma K.1, $\mathbb{E}[\|\widehat{\nabla C}(\alpha,\theta_i)-\nabla C(\alpha,\theta_i)\|_2^2] \leq \frac{8d}{\nu^2}L_{F,\infty}^2\epsilon_\lambda^2 + \frac{8(d+5)B^2}{m_i} + \frac{2\nu^2 L_{C,2}^2(d+6)^3}{m_i}$. If we choose all m_i to be the same m, then

$$\begin{aligned}\mathbb{E}[\|\widehat{g}_1-\widehat{g}\|_2^2] &\leq \frac{1}{n(1-\beta)^2}\sum_{i=1}^n \|\widehat{\nabla C}(\alpha,\theta_i)-\nabla C(\alpha,\theta_i)\|_2^2 \\ &\leq \mathcal{O}(\frac{d\epsilon_\lambda^2}{\nu^2}+\frac{d+5}{m}+\frac{\nu^2(d+6)^3}{m}).\end{aligned}$$

Thus

$$\mathbb{E}[\|R_3\|_2^2] \leq \mathcal{O}(\frac{d\epsilon_\lambda^2}{\nu^2}+\frac{d+5}{m}+\frac{\nu^2(d+6)^3}{m}). \tag{28}$$

Now we consider R_2. Define the event $A_i = \{C(\alpha,\theta_i) \geq \widetilde{v}_\beta\}, \widehat{A_i} = \{\widehat{C}(\alpha,\theta_i) \geq \widehat{v}_\beta\}$ and $A_i \Delta \widehat{A_i} := (A_i \backslash \widehat{A_i}) \cup (\widehat{A_i} \backslash A_i)$. Then

$$\begin{aligned}\|R_2\|_2 &\leq \frac{1}{n(1-\beta)}\sum_{i=1}^n \|\nabla C(\alpha,\theta_i)\|_2 \cdot \mathbb{1}_{A_i\Delta\widehat{A_i}} \\ &\leq \frac{1}{n(1-\beta)}\sum_{i=1}^n B\mathbb{1}_{A_i\Delta\widehat{A_i}},\end{aligned}$$

and

$$\begin{aligned}\|R_2\|_2^2 &\leq \frac{1}{n^2(1-\beta)^2}(\sum_{i=1}^n B\mathbb{1}_{A_i\Delta\widehat{A_i}})^2 \\ &\leq \frac{1}{n(1-\beta)^2}B^2\sum_{i=1}^n \mathbb{1}_{A_i\Delta\widehat{A_i}}.\end{aligned}$$

Notice that

$$\mathbb{P}(\mathbb{1}_{A_i\Delta\widehat{A_i}}) = \mathbb{P}(A_i\backslash\widehat{A_i}) + \mathbb{P}(\widehat{A_i}\backslash A_i).$$

As the estimation error of λ, i.e. $\|\hat{\lambda}-\lambda\|_\infty$, is bounded by ϵ_λ and F is $L_{F,\infty}$-Lipschitz continuous w.r.t $\|\cdot\|_\infty$, we have $|\widehat{C}(\alpha,\theta_i)-C(\alpha,\theta_i)| \leq L_{F,\infty}\epsilon_\lambda$. As a result, $|\widetilde{v}_\beta - \widehat{v}_\beta| \leq L_{F,\infty}\epsilon_\lambda$. Notice that $\{C(\alpha,\theta_i) \geq \widetilde{v}_\beta + 2L_{F,\infty}\epsilon_\lambda\} \subseteq \{\widehat{C}(\alpha,\theta_i) \geq \widehat{v}_\beta\} \subseteq \{C(\alpha,\theta_i) \geq \widetilde{v}_\beta - 2L_{F,\infty}\epsilon_\lambda\}$. Then we have $\mathbb{P}(A_i\backslash\widehat{A_i}) + \mathbb{P}(\widehat{A_i}\backslash A_i) \leq 4\ell_C L_{F,\infty}\epsilon_\lambda$, by the assumption on the cumulative distribution function of C, and thus

$$\mathbb{E}\|R_2\|_2^2 \leq \frac{4}{(1-\beta)^2}B^2\ell_C L_{F,\infty}\epsilon_\lambda = \mathcal{O}(\epsilon_\lambda). \tag{29}$$

Combining (27), (28) and (29), we have

$$\mathbb{E}\|[g-\widehat{g}]\|_2^2 \leq \mathcal{O}\left(dn^{-1}+\epsilon_\lambda+\frac{d\epsilon_\lambda^2}{\nu^2}+\frac{d+\nu^2 d^3}{m}\right).$$

□

Theorem 12 implies that the error of the gradient estimator can be reduced to arbitrarily small by increasing the sample size n, m or decreasing the truncation error ϵ_γ.

L Implementing Details

We evaluate our proposed formulation and algorithm on the offline Frozen Lake problem (Ravichandiran, 2018), an OpenAI benchmark. We consider different convex loss functions, including the mean and Kullback-Leibler (KL) divergence, for various tasks. We compare the Bayesian Risk Policy Gradient (BR-PG) algorithm with CVaR risk measure under different risk levels $\beta = 0, 0.5, 0.9$, respectively, with two other methods. The first is the empirical approach, which fits a maximum likelihood estimator (MLE) to the data and solves the MDP using the estimated parameters. The second is a modified offline version of distributionally robust Q-learning (DRQL) (Liu et al., 2022),

which uses Q-learning to optimize worst-case performance over a KL divergence ball centered at the MLE kernel. When the risk level β approaches 1, Bayesian-risk performance is similar as the worst-case performance. Since we are considering an offline planning problem, we modify the DRQL to interact with an offline simulator that uses the transition kernel with the MLE parameters derived from the data. For a fair comparison, we conduct DRQL experiments with different radii of the KL divergence ball. **Linear Loss.** We consider the linear loss function, which corresponds to the total discounted cost in a classical MDP problem. This is referred to as one replication, and we repeat for 50 replications using different independent data sets. **Episodic Case.** We consider the episodic setting with 50 replications where the data collection and policy update are alternatively conducted. **Mimicking a policy.** Here we consider a different problem of mimicking an expert policy still using Frozen Lake environment and 50 replications. The loss function to minimize is defined as the KL divergence between state occupancy measure under the current policy and the expert state distribution.

Frozen lake problem. Consider moving from the Start (S) to the Goal (G) on an 5×5 frozen lake with 6 holes (H). Then there are 18 ices (F) (involving Start). The agent may not move in the intended direction as the ice is slippery. The position is the row-column coordinate (i, j) with $i, j \in \{0, 1, 2, 3, 4\}$ and the state is the $5 * i + j$. The state space is $\{0, 1, \ldots, 24\}$. The action set consists of moving in four directions. The unknown slippery probability is θ_s. Before reaching the goal and standing on the ice, the agent may move in the intended direction with unknown probability $1 - \theta_s$ and move in either perpendicular direction with probability $\theta_s/2$. When falling into the hole, the agent may try to escape from the hole and move to the intended direction. Each time the agent will succeed in escaping from the hole with unknown probability θ_e. After reaching the Goal, the agent will always stay in the Goal whatever the action is. We set the cost to be 1 for each action on ice before reaching goal. Also, stronger efforts may be made when it is harder to escape from the hole. So we set the per-action cost in hole to be uniformly distributed between $[1, 1 + 2(1 - \theta_e)]$. We aim to find a policy with the minimum general loss function. The data set consists of N historical slippery movements and escapement trials.

Linear Loss. For each of the considered formulations, we obtain the corresponding optimal policy for the same data set and evaluate the actual performance of the obtained policy on the true system, i.e. MDP with the true parameter θ^*. Specifically, we use the linear loss function, which corresponds to the total discounted cost in a classical MDP problem. This is referred to as one replication, and we repeat the experiments for 50 replications using different independent data sets. As the random sampling of output policy in Algorithm 1 is for the purpose of proof, we just choose α_{t_N} as the output for convenience in implementation. Results for the frozen lake problem are presented in Table 3, with varying data size $N = 5$ and $N = 50$, slippery probability $\theta_s = 0.3$ and escape probability $\theta_e = 0.02$. Note that we report the positive-sided variance, which corresponds to the second order moment of the positive component of the difference between the actual loss and the expected loss. Intuitively, a high positive-sided variance indicates more replications with higher costs than the average, which is undesirable.

Episodic Case. We consider the episodic setting where the data collection and policy update are alternatively conducted. Similar with the previous case with fixed data size, we consider the mean loss function with slippery probability $\theta_s = 0.3$, escape probability $\theta_e = 0.02$, and 5×20, 10×10, 20×5 iterations in total. We repeat the experiments for 50 replications on different independent data sets. Figure 1 shows the decrease of the loss function by different methods. As the random sampling of output policy in Algorithm 2 is for the purpose of proof, we just choose $\alpha_{i+1,0} = \alpha_{i,t_i}$ for convenience in implementation.

Results for the frozen lake problem with escape probability $\theta_e = 0.7$ can be found in Table 1 and Table 2.

Figure 4 shows the map of the frozen lake problem with 1 Start(S), 1 Goal(G), 6 holes(H) and remaining frozen(F) parts. We design such a map so that the agent has to avoid falling in the hole when the escape probability is very small and cross the hole when the escape probability is high. Detailed parameters are set as follows. The true slippery probability is 0.3. The iteration number for gradient descent is 100, the stepsize is 0.5, and the sample number in each iteration is $r = 30$. we set the discounter factor to be $\gamma = 0.97$, the truncation horizon for occupancy measure to be $K = 130$. (24).

Table 1: Results for frozen lake problem. Expected loss and positive-sided variance at different risk levels α are reported for different algorithms. Standard errors are reported in parentheses. Escape probability $\theta_e = 0.7$ and number of data points is $N = 5$.

Approach	loss function: mean	
	expected loss	positive-sided variance
BR-PG ($\beta = 0$)	10.322 (0.0182)	0.0153
BR-PG ($\beta = 0.5$)	10.520(0.105)	0.502
BR-PG ($\beta = 0.9$)	11.718 (0.357)	4.982
Empirical	11.667 (0.0687)	0.156
DRQL (radius=0.05)	11.223(0.185)	1.283
DRQL (radius=1)	20.751(1.438)	69.514
DRQL (radius=20)	23.181(1.396)	57.495

Table 2: Results for frozen lake problem. Expected loss and positive-sided variance at different risk levels α are reported for different algorithms. Standard errors are reported in parentheses. Escape probability $\theta_e = 0.7$ and number of data points is $N = 50$.

Approach	loss function: mean	
	expected loss	positive-sided variance
BR-PG ($\beta = 0$)	10.271 (0.00227)	0.000197
BR-PG ($\beta = 0.5$)	10.256 (0.00211)	0.000188
BR-PG ($\beta = 0.9$)	10.230(0.00294)	0.000398
Empirical	11.316 (0.0235)	0.017
DRQL (radius=0.05)	10.888(0.171)	1.235
DRQL (radius=1)	20.990(1.324)	56.027
DRQL (radius=20)	23.500(1.282)	51.915

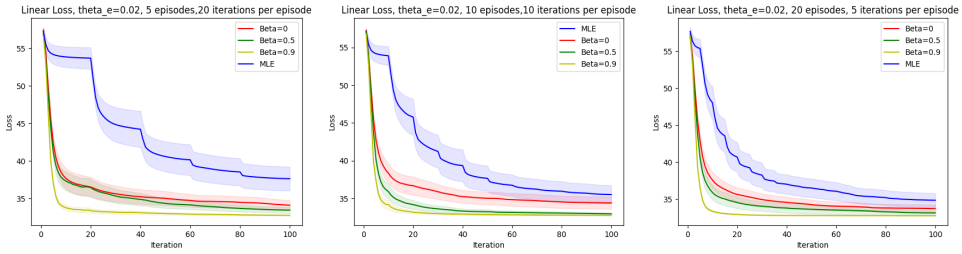
For the "mean" loss function, we use the maximum likelihood estimator (MLE) of θ as the empirical measure to be compared with BR-PG. Also, we use the distributionally robust Q-learning (DRQL)(Liu et al., 2022) with different radius for the KL divergence ball as another benchmark. We also use the MLE of θ as the parameter for the center of the KL divergence ball in DRQL with different radius. For BR-PG, the sample number from posterior in each iteration is 30, the total iteration number is 100, the step size of SGD is chosen to be 1, and the prior distributions are chosen to be Beta$(1, 1)$ for two parameters. We show the histogram of total cost over 50 replications for all methods in Figure 5 with the risk level 0.8 for CVaR over replications, which visualize the measures of dispersion.

Mimicking a policy. Here we consider a different problem of mimicking an expert policy still using Frozen Lake environment. Given an expert policy, we have access to the state distribution of the expert policy under the true environment, which is denoted by a nonnegative function J satisfying $\sum_{s\in\mathcal{S}} J(s) = 1$. The loss function we want to minimize is defined as the KL divergence between state occupancy measure under the current policy and the expert state distribution $F(\lambda) = \text{KL}\left((1-\gamma)\sum_{a\in\mathcal{A}}\lambda_a || J\right) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(1-\gamma)\lambda_{sa}\log\left(\frac{\sum_{a\in\mathcal{A}}(1-\gamma)\lambda_{sa}}{J(s)}\right)$. We compare the BR-PG algorithm with CVaR risk measure under different risk levels $\beta = 0, 0.5, 0.9$, respectively, with the benchmark empirical approach using the MLE estimator for the parameter as before. Figure 3 shows the decrease of the loss function by different methods. It should be noticed that DRQL can only be applied to the "mean" loss function, thus we don't use it as a benchmark. The performance of the 50 replications is shown in figure 6, where the shown results start from the 30-th iteration.

Conclusions. In each replication, data points are randomly sampled from the true distribution. While facing the epistemic uncertainty, BR-PG algorithm provides robustness across different loss functions. Table 3 shows that our BR-PG algorithm has lower linear loss, standard error and positive-sided variance (psv), demonstrating more robustness in the sense of balancing the mean and variability of the actual cost. In contrast, the empirical approach performs badly when the data size is small, e.g. $N = 5$, indicating that it is not robust against the epistemic uncertainty and suffers from the scarcity of data. DRQL also performs better than empirical method but worse than our algorithm in the sense

Table 3: Results for frozen lake problem. Linear loss and positive-sided variance at different risk levels α are reported for different algorithms and different data sizes with linear loss function. Standard errors are reported in parentheses. Escape probability $\theta_e = 0.02$ and number of data points is $N = 5$ and 50.

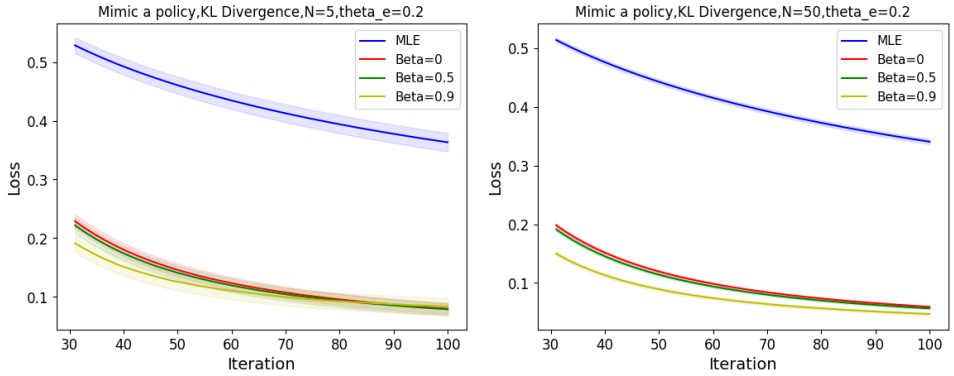
Approach	N=5		N=50	
	linear loss	positive-sided variance	linear loss	positive-sided variance
BR-PG ($\beta = 0$)	33.886(0.347)	5.212	32.784 (0.00825)	0.0026
BR-PG ($\beta = 0.5$)	33.104 (0.127)	0.710	32.757 (0.00516)	0.00119
BR-PG ($\beta = 0.9$)	32.854 (0.0641)	0.193	32.741 (0.00283)	0.000376
Empirical Method	37.057(0.927)	34.387	33.340 (0.0936)	0.380
DRQL(radius=0.05)	37.936(0.887)	26.554	34.365(0.366)	5.139
DRQL(radius=1)	35.216(0.732)	22.213	32.924(0.105)	0.519
DRQL(radius=20)	36.255(0.813)	24.622	32.855(0.063)	0.179
Optimal Policy under True Model	32.499		32.499	



(a) 5×20 total iterations (b) 10×10 total iterations (c) 20×5 total iterations

Figure 2: Results for episodic case with different episode numbers and iterations per episode under the same escape probability $\theta_e = 0.02$ and 50 replications. Here the loss function is still chosen to be the linear loss. 95% confidence intervals are reported by the shaded bands.

of having larger mean and variance of the loss. Figure 1 shows that the loss of our algorithm decreases



(a) $N = 5$ (b) $N = 50$

Figure 3: Results for loss function "KL Divergence" with data sizes $N = 5$ and 50 under $\theta_e = 0.02$. 95% confidence intervals are reported in the shaded area.

quickly in spite of few data. In the episodic case, the loss function decreases faster with more episodes (but the same total number of iterations), due to more collected data with more episodes. The loss function of our BR-PG method decreases more quickly in early episodes, which is shown by two differences between Figure 3a and Figure 3b. First, the 95% confidence interval, shown in the shaded band around each curve, is narrower for $N = 50$. Second, the absolute loss of $N = 50$ decreases by about 20% compared with $N = 5$. Figure 3 demonstrates the better performance of our proposed BR-PG algorithm compared to the empirical approach, where we achieve smaller loss and lower variability, for the policy mimicking task. From Table 3 and Figure 1, we can see when there are more data, the posterior distribution used in BR-PG algorithm and the MLE estimator used in the empirical approach converges to the true parameter as data size increases, which reduces to solving

an MDP with known transition probability, and therefore, the optimal policies and the actual costs tend to be similar.

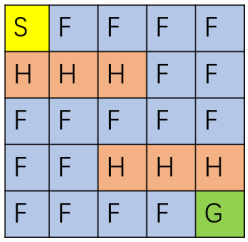
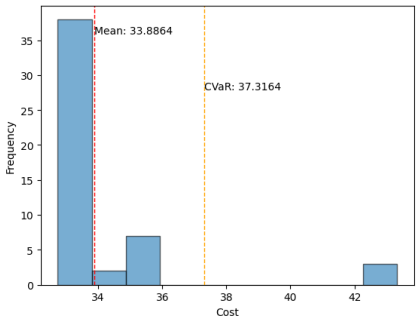
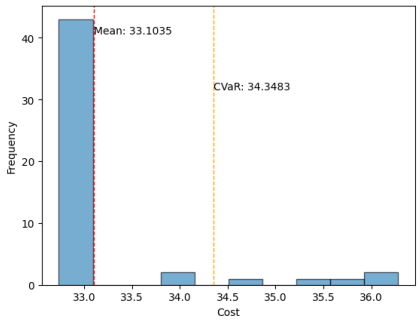


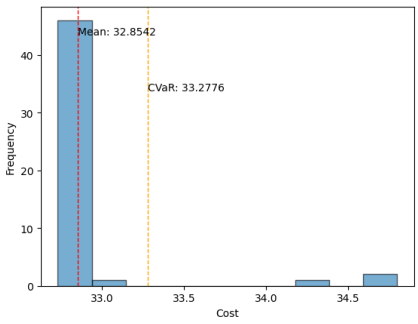
Figure 4: Map of frozen lake problem



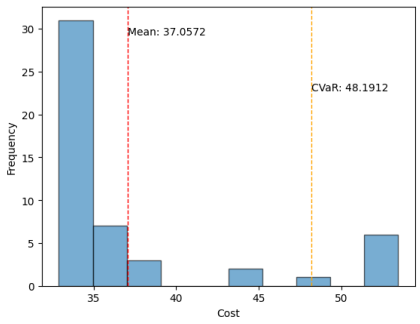
(a) BR-PG ($\beta = 0$)



(b) BR-PG ($\beta = 0.5$)

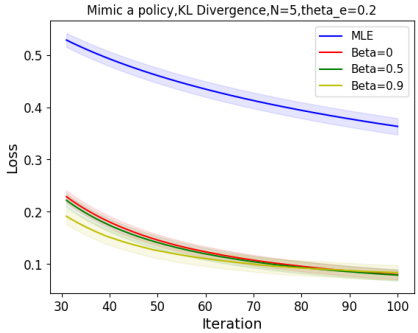


(c) BR-PG ($\beta = 0.9$)

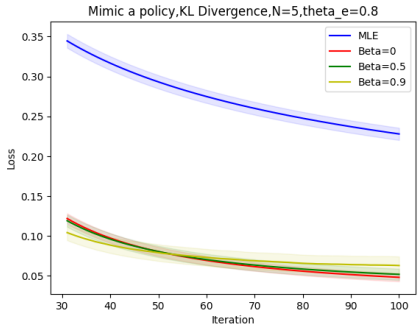


(d) empirical

Figure 5: Result for utility function "mean" with data size $N = 5$ and escape probability $\theta_e = 0.02$



(a) $\theta_e = 0.2$



(b) $\theta_e = 0.8$

Figure 6: Results for utility function "KL divergence" with data size $N = 5$ and escape probability $\theta_e = 0.2$ and $\theta_e = 0.8$