Risk-Sensitive Reinforcement Learning with ϕ-Divergence-Risk

Xinyi Ni and Lifeng Lai

Abstract

Standard reinforcement learning (RL) algorithms primarily focus on minimizing the expected sum of costs, which can be insufficient in contexts where risk sensitivity is crucial. This paper explores the application of a class of coherent risk measures, termed ϕ-Divergence-Risk (PhiD-R) in risk-sensitive RL. This class of risk measures not only includes established measures such as Conditional Value-at-Risk (CVaR) as special cases but also broadens the horizon for exploring new risk measures. We propose a trajectory-based policy gradient method specifically tailored for PhiD-R, applicable across all forms of risk measures formed by different ϕ-divergence. We prove the asymptotic convergence of our algorithm towards locally optimal policies using multi-time stochastic approximation techniques. Extensive simulation experiments validate the effectiveness and practicality of our approach.

Index Terms

risk-sensitive, ϕ-divergence, multi-time scale, stochastic approximation

I. Introduction

In the standard risk-neutral reinforcement learning (RL), the goal is to find an optimal policy that minimizes the expected sum of (discounted) cost [1], [2]. However, in many applications, decision-makers exhibit a preference for *risk-sensitive* optimizations, acknowledging the significance of events with small probability but severe consequences. Rather than focusing solely on the expected value of the sum of costs, risk-sensitive optimization approaches aim to incorporate risk measures into the objective functions [3]–[14].

A multitude of risk measures have been studied in the literature and successfully applied to RL, such as Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), Entropic risk measure and Entropic Value-at-Risk (EVaR) et al [15]–[25]. One widely used framework for capturing risk is through *coherent* risk measures, which are a class of risk measures that satisfy a set of natural and desirable properties: *1) monotonicity, 2) translation invariance, 3) subadditivity, 4) positive homogeneity*, ensuring rationality and reliability in capturing risk preferences [18], [26]. One important property of coherent risk measures is the *dual representation*, where they could be interpreted as the mean of a random variable under a probability distribution that is in an uncertainty set defined in the neighborhood of the true probability distribution [26].

The extensive exploration of risk measures in decision-making contexts often requires adopting specific algorithms tailored to each measure, potentially reducing decision-making efficiency. While some risk measures incorporate a risk-tolerance parameter that reflects decision-makers' preferences to an extent, their varied methodologies might not capture these preferences accurately due to different risk quantification approaches. [18] introduce a policy gradient method applicable to a wide range of coherent risk measures. However, this method assumes a structured form of the measures' envelope sets in their dual representation. Although the approach is comprehensive, it involves significant computation complexity, especially in identifying saddle points across four parameters. This complexity, arising from the specific constraints within the dual representation, represents a trade-off between generality and computational efficiency in risk-sensitive reinforcement learning. Realizing these challenges in existing work on risk senstive RL, we are prompted to explore a critical question:

X. Ni and L. Lai are with the Department of Electrical and Computer Engineering, University of California, Davis, CA. Email:{xyni, lflai}@ucdavis.edu. This work was supported by National Science Foundation under grant CCF-2232907.

Is it possible to develop a class of coherent risk measures that cover popular risk measures and to design an accompanying algorithm that not only offers decision-makers greater flexibility in selecting risk measures but also ensures efficiency and robustness?

To address this question, we adopt a new class of risk measures named *ϕ-Divergence-Risk (PhiD-R)*, whose uncertainty sets in their dual representations are defined by using ϕ-divergence [27], to RL problems. Our choice of this class of risk measures is motiviated by several factors: 1) PhiD-R is coherent and includes many widely adopted risk measures such as CVaR and EVaR as special cases [27]. 2) ϕ-divergence has been thoroughly explored in the machine learning domain, particularly in policy optimization and robust RL [28]–[33]. This extensive research supports the potential of ϕ-divergence to foster innovative developments in risk measures for risk-sensitive RL. 3) Previous work by [16] illustrated that solving CVaR RL was equivalent to tackling risk-neutral RL when uncertainties in transition probabilities are defined by specific divergence measures. This finding motivates further exploration into the equivalence of PhiD-R RL and robust RL, aiming to address the robustness concerns identified. 4) The explicit generalized representation of PhiD-R in [27] allows for the development of a generalized policy gradient method applicable to all forms of ϕ-divergence, which ensures both flexibility and efficiency of the approach.

In this study, we introduce a trajectory-based policy gradient method tailored to solve RL problems under this new class of risk measures, PhiD. The explicit representation of PhiD-R allows for efficient gradient estimation. Based on these gradient estimates, we propose specific update rule for each parameter. By using multi-time stochastic approximation technique [19], [22], [34], we demonstrate that our proposed method asymptotically converges to locally optimal policies. This approach is highly versatile, applying to the entire spectrum of ϕ-divergence, thereby broadening the scope beyond traditional risk measures such as CVaR. This extension also offers a new approach to address CVaR RL and explores novel approaches within risk-sensitive RL. Our approach benefits from the coherence property of PhiD-R and the dual presentation theorem, which ensures that solving PhiD-R RL is equivalent to solving robust RL when uncertainties in transition probabilities are defined by ϕ-divergence. This connection between risk and robustness is particular valuable when decision-makers face scenarios with inherent uncertainty and wish to incorporate their risk preferences.

Several works are closely related to our studies. [18] proposes a generalized method for solving RL problems with coherent risk measures, which aligns with PhiD-R, and demonstrates convergence to local optimality. Our approach also guarantees near-optimality, while providing a simplified solution for PhiD-R, requiring fewer assumptions and optimized parameters. We build on the well-established representation of PhiD-R from [27], offering a more efficient and practical method tailored to these risk measures. In particular, when applied to CVaR RL, our algorithm reduces the number of parameters without compromising local optimality. Compared to policy gradient-based CVaR RL approaches that extend the likelihood-ratio method for demonstrating local optimality [17], [35], our work estimates gradients directly using the explicit representation of PhiD-R. While our methodology and objectives differ from those of [17], [35], all algorithms achieve convergence to a locally optimal policy. Furthermore, our approach contrasts with existing policy gradient research on CVaR [19], [36], [37], which is typically limited to the constrained RL framework. Our method also diverges from [22], which focuses solely on EVaR.

The remainder of this paper is organized as follows. Section II provides background on risk measures, ϕ-divergence and the new risk measure class PhiD-R, detailing their definition and drawing upon exiting properties from [27]. Section III outlines the notations and problem formulation of this work. In Section IV, we introduce the proposed trajectory-based policy gradient algorithm and establishes its asymptotic convergence towards local optima, utilizing the multi-time stochastic approximation technique from [34]. Section V presents empirical validation through various experimental setups. Finally, Section VI offers concluding remarks.

II. PRELIMINARIES

A. Risk Measures

A risk measure ρ is a mapping from a random variable $Z \in \mathcal{Z}$ with distribution P to a real value, providing a mean to assess and quantify the risk associated with the random variable Z. We will particularly focus on coherent risk measures [26], which satisfies the following properties:

- (P1) *Translation invariance*: $\rho(Z + c) = \rho(Z) + c$ for any $Z \in \mathcal{Z}$ and $c \in \mathbb{R}$;
- (P2) *Subadditivity*: $\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2)$ for all $Z_1, Z_2 \in \mathcal{Z}$;
- (P3) *Monotonicity*: If $Z_1(w) \leq Z_2(w)$ for all $w \in \Omega$, then $\rho(Z_1) \leq \rho(Z_2)$;
- (P4) *Positive homogeneity*: $\rho(\lambda Z) = \lambda\rho(Z)$ for all $Z \in \mathcal{Z}$ and $\lambda \geq 0$.

Examples of coherent measures include *Conditional Value-at-Risk* (CVaR) [38] and *Entropic Value-at-Risk* (EVaR) [27], while non-coherent measures include *Variance*, *Mean-Standard-Deviation* (MSD), and *Value-at-Risk* (VaR) [26].

A key feature of coherent risk measures is the dual representation theorem [26], which states that a coherent risk measure can be expressed as the maximum expected value over a probability distribution Q within an uncertainty set $\mathcal{U}$ around the true distribution P, i.e.,

$$\rho(Z) = \max_{Q \in \mathcal{U}} \mathbb{E}_Q[Z].$$

Different coherent risk measures correspond to different sets $\mathcal{U}$.

We now present several risk measures relevant to this work. Consider a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω represents the sample space, $\mathcal{F}$ is a σ-algebra over Ω, and $\mathcal{P}$ is a probability measure over $\mathcal{F}$. Let $\mathcal{Z}$ denote the space of bounded random variables $Z : \Omega \to \mathbb{R}$ defined on this probability space. In this work, we focus on the case where the random variable Z is non-negative and bounded within the interval $[Z_{\min}, Z_{\max}]$. Let Q and P be two probability measures within this probability space.

CVaR, also referred to as expected shortfall or tail conditional expectation, is defined at a confidence level $\alpha \in (0, 1]$ as follows [38]:

$$\text{CVaR}_\alpha(Z) = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-\alpha} \mathbb{E}_P \left[(Z - t)^+ \right] \right\},$$

where $(z)^+ = \max(z, 0)$. CVaR captures the expected value of losses that exceed the VaR threshold, which provides a more comprehensive assessment of tail risks in comparison to VaR alone. An important property of CVaR is its coherency, and its dual representation is:

$$\text{CVaR}_\alpha(Z) = \sup_{Q \in \mathcal{U}_{\text{CVaR}}} \mathbb{E}_Q[Z],$$

where $\mathcal{U}_{\text{CVaR}} = \left\{ Q \ll P : D_{\text{RN}}(Q, P) \in \left[0, \frac{1}{1-\alpha}\right] \right\}$ with the Radon-Nikodym derivative $D_{\text{RN}}(Q, P) := \frac{Q(\omega)}{P(\omega)}$.

We also introduce another significant risk measure, known as EVaR. Suppose that the moment generating function $M_Z(t) = \mathbb{E}_P \left[e^{tZ} \right]$ exists for all $t \in \mathbb{R}^+$ for the random variable Z. The EVaR at a given confidence level α is defined as follows [27]:

$$\text{EVaR}_\alpha(Z) = \inf_{t > 0} \left\{ t^{-1} \ln(M_Z(t)) - t^{-1} \ln(1 - \alpha) \right\}.$$

It is noteworthy that EVaR is also a coherent risk measure, and its dual representation is:

$$\text{EVaR}_\alpha(Z) = \sup_{Q \in \mathcal{U}_{\text{EVaR}}} \mathbb{E}_Q[Z],$$

where $\mathcal{U}_{\text{EVaR}} = \{Q \ll P : D_{\text{KL}}(Q, P) \leq -\ln(1 - \alpha)\}$, and the KL divergence is defined as $D_{\text{KL}}(Q, P) := \sum_\omega Q(\omega) \log \frac{Q(\omega)}{P(\omega)}$.

B. ϕ-Divergence-Risk (PhiD-R)

Building upon the dual representation theorem, which links different choices of the uncertainty sets $\mathcal{U}$ to various risk measures, we extend this framework by constructing risk measures using ϕ-divergence.

For two probability measures Q and P within the probability space, the ϕ-divergence is defined as:

$$D_\phi(Q,P) = \sum_{z\in\Omega} P(z)\phi\left(\frac{Q(z)}{P(z)}\right), \tag{1}$$

where ϕ is a closed and convex function satisfying $\phi(1) = 0$. The choice of the function ϕ directly determines the type of divergence, allowing for various risk measures to be modeled. Below, we present some common choices of ϕ and their corresponding divergences:

1). Total variation distance: $\phi(x) = \frac{1}{2}|x-1|$.

2). KL divergence: $\phi(x) = x\log x$ for $x \geq 0$.

3). χ^2-divergence: $\phi(x) = (x-1)^2$.

We now define the ϕ-Divergence-Risk (PhiD-R), in which the uncertainty sets $\mathcal{U}$ are constructed based on ϕ-divergence, following the framework described in [27].

Definition 1. (ϕ-Divergence-Risk) Let ϕ be a closed and convex function with $\phi(1) = 0$, and $\beta > 0$. The ϕ-divergence risk measure with divergence level β for a random variable $Z \in \mathcal{Z}$ is defined as

$$\text{PhiD-R}_{\phi,\beta}[Z] := \sup_{Q\in\mathcal{U}} \mathbb{E}_Q[Z],$$

where $\mathcal{U} = \{Q \ll P : D_\phi(Q,P) \leq \beta\}$ with D_ϕ being defined in (1).

The definition via dual representation ensures two key outcomes: (1) PhiD-R is a coherent risk measure, as validated by Theorem 3.2 in [27]; and (2) building on insights from [15], solving PhiD-R RL aligns with robust RL, where uncertainties in transition probabilities are characterized by ϕ-divergence. Furthermore, Theorem 5.1 in [27] provides an explicit representation of PhiD-R, which plays a crucial role in developing the policy gradient method discussed in the following sections.

Theorem 1 (Theorem 5.1 of [27])**.** *For any $Z \in \mathcal{Z}$, the ϕ-divergence risk measure has the following representation:*

$$\textit{PhiD-R}_{\phi,\beta}[Z] = \inf_{\nu>0,\omega\in\mathbb{R}} \left\{ \nu\left[\omega + \mathbb{E}_P\left(\phi^*\left(\frac{Z}{\nu} - \omega + \beta\right)\right)\right]\right\}, \tag{2}$$

where ϕ^* is the conjugate of ϕ (the Legendre–Fenchel transform).

It is important to note that the class of ϕ-divergence risk measures encompasses widely used risk measures in risk-sensitive RL, such as CVaR and EVaR, as special cases. For instance, by selecting $\phi(x) = 0$ for $0 \leq x \leq \frac{1}{1-\alpha}$ and $+\infty$ otherwise, we recover CVaR, with $\phi^*(x) = \frac{1}{1-\alpha}\max\{0,x\}$. Additionally, by setting $\beta = 0$, we obtain

$$\text{PhiD-R}_{\phi,\beta}[Z] = \inf_{t\in\mathbb{R}}\left\{t + \frac{1}{1-\alpha}\mathbb{E}_P\left[(Z-t)^+\right]\right\},$$

which exactly corresponds to the definition of CVaR as mentioned earlier.

Similarly, by selecting $\phi(x) = x\log x$ for $x \geq 0$, we recover EVaR, with $\phi^*(x) = e^{x-1}$. By setting $\beta = -\ln(1-\alpha)$ [27], we derive

$$\text{PhiD-R}_{\phi,\beta}[Z] = \inf_{\nu>0}\left\{\nu\ln\mathbb{E}_P\left(e^{\frac{Z}{\nu}}\right) - \nu\ln(1-\alpha)\right\},$$

which corresponds to the representation formula for EVaR.

III. Problem Statement

We consider a Markov decision process (MDP) defined by a tuple $(\mathcal{X}, \mathcal{A}, C, P, P_0, \gamma)$. Here, $\mathcal{X}$ represents the state space, $\mathcal{A}$ denotes the action space, and $C(x, a) \in [0, C_{\max}]$ represents a bounded deterministic cost. The transition probability distribution is denoted as $P(\cdot|x, a)$, and P_0 represents the initial state probability distribution, where x_0 is set deterministically as $P_0(x) = \mathbb{I}\{x_0 = x\}$. The discounting factor is denoted as $\gamma \in [0, 1]$. Each state $x \in \mathcal{X}$ is associated with an action set $\mathcal{A}(x)$. We consider a stationary policy $\pi(\cdot|x)$, parameterized by a κ-dimensional vector θ in the policy gradient method. The policy space is defined as $\pi(\cdot|x, \theta), x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R}^\kappa$, where Θ is assumed to be a convex compact set.

The total discounted costs incurred by an agent following policy π, starting at state x, is denoted as $J^\theta(x)$. It is defined as the sum of the discounted costs encountered over a time horizon T, with γ representing the discount factor and $C(x_k, a_k)$ denoting the cost at state x_k when action a_k is taken

$$J^\theta(x) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k) | x_0 = x, \pi(\cdot|\cdot, \theta).$$

Our goal is to solve PhiD-R RL by minimizing the objective function

$$\min_\theta \quad \text{PhiD-R}_{\phi,\beta}\left(J^\theta(x_0)\right) \tag{3}$$

for a given divergence level $\beta \geq 0$. This optimization problem seeks to find the optimal policy θ^* that minimizes the risk-sensitive objective. By incorporating the representation (2) of the ϕ-divergence risk measure, the optimization problem can be reformulated as follows:

$$\min_{\theta,\nu,\omega} L(\nu, \omega, \theta) := \nu \left[\omega + \mathbb{E}_P\left(\phi^*\left(\frac{J^\theta(x_0)}{\nu} - \omega + \beta\right)\right)\right]. \tag{4}$$

While similar formulations have been explored in the literature on risk measures and optimization, our application of this reformulation to the context of RL is novel. Our main idea to solve the optimization problem (4) is to adopt a gradient descent method, which will be detailed in the subsequent section. Before presenting our algorithm and providing convergence analysis, we first state Assumption 1, which is a typical assumption found in the literature related to policy gradient methods.

Assumption 1. For any $x \in \mathcal{X}$ and $a \in \mathcal{A}(x)$, $\pi(a|x, \theta)$ is continuously differentiable in Θ and $\nabla_\theta \pi(a|x, \theta)$ is a Lipschitz function in θ for every $(x, a) \in \mathcal{X} \times \mathcal{A}$. Moreover, the ratio $\nabla_\theta \pi(a|x, \theta)/\pi(a|x, \theta)$ is bounded for all $\theta \in \mathbb{R}^\kappa$ and every $(x, a) \in \mathcal{X} \times \mathcal{A}$.

IV. Trajectory-based Policy Gradient

In this section, we introduce a trajectory-based policy gradient algorithm that descends in ν, μ, and θ based on the gradients of $L(\nu, \omega, \theta)$ with respect to ν, ω, and θ, respectively.

During each iteration, the algorithm generates N trajectories by executing the current policy π. Subsequently, these trajectories are utilized to estimate the gradients. Using these gradient estimates, the parameters ν, ω, and θ are updated with stepsizes that satisfy specific conditions to be discussed in the sequel.

Let $\xi = \{x_0, a_0, \ldots, x_{T-1}, a_{T-1}, x_T\}$ represent a single trajectory, where x_T denotes the terminal state. The corresponding cost function is given by

$$J(\xi) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k)$$

and the probability of generating such a trajectory is

$$\mathbb{P}_\theta(\xi) = P_0(x_0) \prod_{k=0}^{T-1} \pi(a_k|x_k, \theta) P(x_{k+1}|x_k, a_k).$$

We also have

$$\nabla_\theta \log \mathbb{P}_\theta(\xi) = \sum_{k=0}^{T-1} \nabla_\theta \log \pi(a_k|x_k,\theta) = \sum_{k=0}^{T-1} \nabla_\theta \pi(a_k|x_k,\theta)/\pi(a_k|x_k,\theta),$$

whenever $\mathbb{P}_\theta(\xi) \neq 0$ and $\pi(a_k|x_k,\theta) \in (0,1]$.

With the trajectory representation established, we can now proceed to derive the estimated form of these gradients. The derivation details could be found in Appendix A.

Gradient estimate w.r.t ν

$$\widehat{\nabla}_\nu L(\nu,\omega,\theta) = \omega + \sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right) - \sum_\xi \mathbb{P}_\theta(\xi)\frac{J(\xi)}{\nu}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.$$

Gradient estimate w.r.t ω

$$\widehat{\nabla}_\omega L(\nu,\omega,\theta) = \nu - \nu\sum_\xi \mathbb{P}_\theta(\xi)\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.$$

Gradient estimate w.r.t θ

$$\widehat{\nabla}_\theta L(\nu,\omega,\theta) = \nu\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right).$$

However, these estimates are not immediately usable due to the presence of the unknown transition probability $P(x_{k+1}|x_k,a_k)$ in the expression of $\mathbb{P}_\theta(\xi)$. To address this, we use empirical mean to estimate the sample mean. Moreover, it is important to note that when $\mathbb{P}_\theta(\xi) \neq 0$, the gradients $\nabla_\theta \mathbb{P}_\theta(\xi)$ and $\nabla_\theta \log \mathbb{P}_\theta(\xi)$ can be expressed as $\mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)$, and the latter is only dependent on π without any reliance on the unknown transition probability $P(x_{k+1}|x_k,a_k)$. By utilizing these insights and generating N trajectories per iteration, we obtain the gradient estimates as:

Gradient estimate w.r.t ν

$$\widetilde{\nabla}_\nu L(\nu,\omega,\theta) = \omega + \sum_\xi \frac{1}{N}\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right) - \sum_\xi \frac{1}{N}\frac{J(\xi)}{\nu}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}. \tag{5}$$

Gradient estimate w.r.t ω

$$\widetilde{\nabla}_\omega L(\nu,\omega,\theta) = \nu - \nu\sum_\xi \frac{1}{N}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}. \tag{6}$$

Gradient estimate w.r.t θ

$$\widetilde{\nabla}_\theta L(\nu,\omega,\theta) = \nu\sum_\xi \frac{1}{N}\nabla_\theta \log \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right). \tag{7}$$

Based on these gradient estimates and let $\xi_{j,k}$ denote the j-th trajectory generated at iteration k and properly chosen step sizes $\zeta_1(k)$, $\zeta_2(k)$ and $\zeta_3(k)$, we design the following update rules for parameter ν,ω,θ that will be utilized in our algorithm.

ν-update

$$\begin{aligned}\nu_{k+1} &= \Gamma_{\mathcal{N}}\left[\nu_k - \zeta_1(k)\widetilde{\nabla}_\nu L(\nu,\omega,\theta)\big|_{\nu=\nu_k,\omega=\omega_k,\theta=\theta_k}\right] \\ &= \Gamma_{\mathcal{N}}\left[\nu_k - \zeta_1(k)\left(\omega_k + \sum_{j=1}^N \frac{1}{N}\phi^*\left(\frac{J(\xi_{j,k})}{\nu_k} - \omega_k + \beta\right) - \sum_{j=1}^N \frac{1}{N}\frac{J(\xi_{j,k})}{\nu_k}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega_k+\beta}\right)\right].\end{aligned} \tag{8}$$

ω-update

$$\begin{aligned}\omega_{k+1} &= \Gamma_{\mathcal{R}}\left[\omega_k - \zeta_2(k)\widetilde{\nabla}_{\omega} L(\nu,\omega,\theta)\big|_{\nu=\nu_k,\omega=\omega_k,\theta=\theta_k}\right] \\ &= \Gamma_{\mathcal{R}}\left[\omega_k - \zeta_2(k)\cdot\left(\nu_k - \nu_k \sum_{j=1}^{N}\frac{1}{N}\frac{\partial \phi^*}{\partial u}\Big|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega_k+\beta}\right)\right].\end{aligned} \tag{9}$$

θ-update

$$\begin{aligned}\theta_{k+1} &= \Gamma_{\Theta}\left[\theta_k - \zeta_3(k)\widetilde{\nabla}_{\theta} L(\nu,\omega,\theta)\big|_{\nu=\nu_k,\omega=\omega_k,\theta=\theta_k}\right] \\ &= \Gamma_{\Theta}\left[\theta_k - \zeta_3(k)\left(\nu_k \sum_{j=1}^{N}\frac{1}{N}\nabla_{\theta}\log \mathbb{P}_{\theta}(\xi_{j,k})\cdot\phi^*\left(\frac{J(\xi_{j,k})}{\nu_k}-\omega_k+\beta\right)\right)\right].\end{aligned} \tag{10}$$

The projections introduced in the update rules, i.e., $\Gamma_{\mathcal{N}}(\nu) = \text{argmin}_{\nu\in[V_{\min},V_{\max}]}||\nu-\hat{\nu}||_2^2$, $\Gamma_{\mathcal{R}}(\omega) = \text{argmin}_{\omega\in[W_{\min},W_{\max}]}||\omega-\hat{\omega}||_2^2$, $\Gamma_{\Theta}(\theta) = \text{argmin}_{\theta\in\Theta}||\theta-\hat{\theta}||_2^2$, are employed to enforce the updated values to remain within specified bounds, thereby ensuring the convergence of the policy gradient algorithm for PhiD-R. Additionally, we adopt a common assumption regarding the stepsizes utilized in the update rules (8) (9) (10).

Assumption 2. The stepsizes $\zeta_1(k)$, $\zeta_2(k)$ and $\zeta_3(k)$satisfy

$$\sum_k \zeta_1(k) = \sum_k \zeta_2(k) = \sum_k \zeta_3(k) = \infty, \tag{11}$$

$$\sum_k \zeta_1^2(k), \sum_k \zeta_2^2(k), \sum_k \zeta_3^2(k) < \infty, \tag{12}$$

$$\zeta_1(k) = o(\zeta_2(k)), \zeta_2(k) = o(\zeta_3(k)). \tag{13}$$

The first two conditions in Assumption 2 are common in RL problems. The third condition assumes that the stepsizes satisfy the standard requirements of stepsizes in multi-scale stochastic approximation algorithms. Moreover, from Eq (13), we observe that the update frequencies for ν, ω, and θ occur at different timescales, with ν updating at the fastest timescale $\zeta_1(k)$, ω updating at a second fast timescale $\zeta_2(k)$, and θ updating at the slowest timescale $\zeta_3(k)$.

Algorithm 1 outlines the proposed trajectory-based policy gradient method for PhiD-R. Line 5 details the collection of N trajectories by following the current parameterized policy with θ_k and line 5 updates the parameters. Lines 9 to 13 describe adjustments to the selected ranges for ν and ω. If no adjustments are needed, the iteration ceases, resulting in the local optimal θ.

Theorem 2 provides theoretical guarantees for Algorithm 1, establishing its convergence to a locally optimal policy for the optimization problem (4).

Theorem 2. *(Local Optimality) Under Assumptions 1 and 2, as $k \to \infty$, the policy sequence generated by Algorithm 1 converges almost surely to a locally optimal policy θ^*.*

Proof Sketch. Our proof is inspired by [15]. Initially, we treat the updates $(\nu_k, \omega_k, \theta_k)$ as a multi-time scale discrete stochastic approximation, under the condition that the stepsizes satisfy Assumption 2. We prove that the sequences $(\nu_k, \omega_k, \theta_k)$ converge to the solutions of the corresponding continuous-time systems, each with varying convergence rates. Subsequently, we apply Lyapunov analysis to demonstrate that the sequences $(\nu_k, \omega_k, \theta_k)$ further converge to local asymptotically stable points denoted as $(\nu^*, \omega^*, \theta^*)$. Finally, we establish that the attained points $(\nu^*, \omega^*, \theta^*)$ serve as local optimal solutions for the optimization problem (3). More details can be found in Appendix B.

Algorithm 1 PhiD-R RL: A Trajectory-based Policy Gradient Method

Given: divergence level β, parameterized policy $\pi(\cdot|\cdot,\theta)$, tolerance parameters $\epsilon_\nu, \epsilon_\omega$.
Initialization: choose $\nu = \nu_0$, $\omega = \omega_0$, $\theta = \theta_0$ and initial state x_0.
while TRUE **do**
  for $k = 0, 1, 2, \dots$ **do**
    Generate N trajectories $\{\xi_{j,k}\}_{j=1}^N$ by following policy π_{θ_k} starting from the initial state x_0.
    Update (ν, ω, θ) by

$$\nu_{k+1} = \Gamma_{\mathcal{N}}\Bigg[\nu_k - \zeta_1(k)\Bigg(\omega_k + \sum_{j=1}^N \frac{1}{N}\phi^*\Bigg(\frac{J(\xi_{j,k})}{\nu_k} - \omega_k + \beta\Bigg) - \sum_{j=1}^N \frac{1}{N}\frac{J(\xi_{j,k})}{\nu_k}\frac{\partial \phi^*}{\partial u}\Bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega_k+\beta}\Bigg)\Bigg],$$

$$\omega_{k+1} = \Gamma_{\mathcal{R}}\Bigg[\omega_k - \zeta_2(k)\cdot\Bigg(\nu_k - \nu_k\sum_{j=1}^N \frac{1}{N}\frac{\partial \phi^*}{\partial u}\Bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega_k+\beta}\Bigg)\Bigg],$$

$$\theta_{k+1} = \Gamma_{\Theta}\Bigg[\theta_k - \zeta_3(k)\Bigg(\nu_k\sum_{j=1}^N \frac{1}{N}\nabla_\theta \log \mathbb{P}_\theta(\xi_{j,k})\cdot\phi^*\Bigg(\frac{J(\xi_{j,k})}{\nu_k} - \omega_k + \beta\Bigg)\Bigg)\Bigg].$$

  end for
  if ν_k lies within the ϵ_ν-neighborhood of the boundary **then**
    Extend the boundary for ν
  else if ω_k lies within the ϵ_ω-neighborhood of the boundary **then**
    Extend the boundary for ω
  else
    Return (ν, ω, θ) and **terminate**
  end if
end while

More specifically, we use the multi-time scale stochastic approximation approach discussed in [34]. In two-time scale stochastic approximation, for sequences (x_i, y_i), consider the following update process:

$$x_{n+1} = x_n + a(n)\left[h(x_n, y_n) + M_{n+1}^{(1)}\right], \tag{14}$$

$$y_{n+1} = y_n + b(n)\left[g(x_n, y_n) + M_{n+1}^{(2)}\right], \tag{15}$$

where h and g are two Lipschitz continuous functions, $M_{n+1}^{(1)}$, $M_{n+1}^{(2)}$ are two Martingale differences w.r.t the increasingσ-field $\mathcal{F}_n = \sigma(x_m, y_m, M_m^{(1)}, M_m^{(2)}, m \le n)$ for $n \ge 0$. The martingale differences satisfy $\mathbb{E}[||M_{n+1}^{(i)}||^2|\mathcal{F}_n] \le K(1 + ||x_n|| + ||y_n||)^2$ for $i = 1, 2$. Additionally, the step sizes $a(n)$ and $b(n)$ are positive scalars satisfying $\sum_n a_(n) = \sum_n b(n) = \infty$, $\sum_n a^2(n), \sum_n b^2(n)$, and $b(n) = o(a(n))$. Notably, the equation (15) can be equivalently expressed as: $y_{n+1} = y_n + a(n)\frac{b(n)}{a(n)}\left[g(x_n, y_n) + M_{n+1}^{(2)}\right]$.

Given that $b(n)$ exhibits a faster convergence to zero compared to $a(n)$, it is natural to consider the ordinary differential equations (ODEs) $\dot{x}(t) = h(x, y)$ and $\dot{y}(t) = 0$. By applying Theorem 2 in Chapter 6 of [34], it can be demonstrated that as $n \to \infty$, (x_n, y_n) converges to $(\lambda(y^*), y^*)$ almost surely to the point $\lambda(y^*)$ represents a globally asymptotically stable equilibrium of the o.d.e. $\dot{x} = h(x, y)$, where λ is a Lipschitz continuous function, and y^* is a globally asymptotically equilibrium of the o.d.e $\dot{y} = g(\lambda(y), y)$.

In our work, the sequence $(\nu_n, \omega_n, \theta_n)$ can be represented as a three-time scale stochastic approximation problem, which can be decomposed into two two-time scale problems. The chosen stepsizes adhere to the conditions outlined in Assumption 2. Consequently, the convergence behavior of the sequence can be characterized as follows:
1). The ω and θ can be viewed as constants when analyze convergence for ν_n.
2). In the analysis of ω_n's convergence, θ can be treated as a constant, and ν can be seen as the converged value $\nu^*(\theta)$.
3). Similarly, in the convergence analysis of θ_n, both ν and ω are represented by their respective converged values $\nu^*(\theta)$ and $\omega^*(\theta)$.

By leveraging the insights derived from the two-time scale stochastic approximation, we obtain that (ν, ω) converges to (ν^*, ω^*). Utilizing contraction arguments, we subsequently prove that this point corresponds to a local minimum for the objective function $L(\nu, \omega, \theta$. Following a similar approach, we further establish the convergence of θ_n, ultimately leading to the overall convergence of the sequence (ν, ω, θ) towards a local optimal point $(\nu^*, \omega^*, \theta^*)$.

V. Experiments

In this section, we present numerical examples to demonstrate the practicality and efficiency of the proposed algorithms. We first validate our approach using an investment problem and the optimal stopping problem, as utilized in related work [15], [18], [22], highlighting comparison over existing methods. Additionally, we conduct a more comprehensive evaluation using OpenAI's Gym environment to further demonstrate the generalizability of our algorithms.

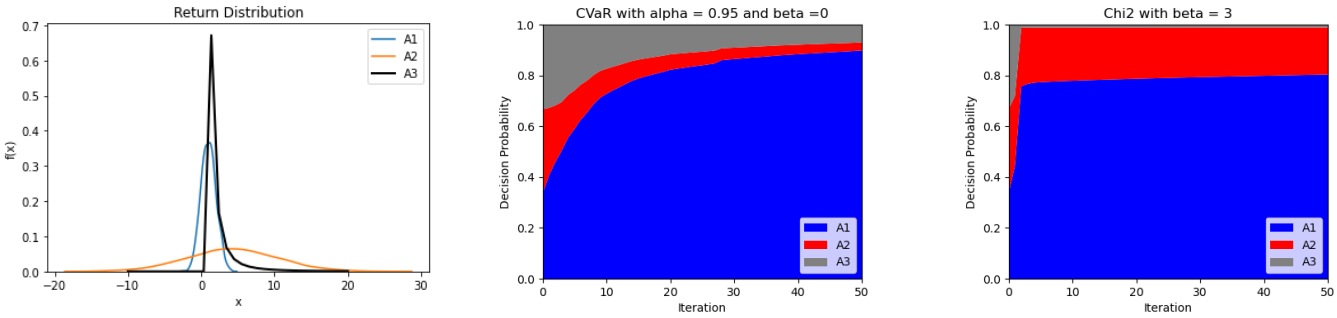
A. Investment Problem

We conduct a validation of our method using the same experimental setup as [18]. We examine a scenario involving a trading agent with options to invest in one of three assets. The returns of the first two assets, A_1 and A_2, follow normal distributions: A_1 is distributed as $\mathcal{N}(1, 1)$, and A_2 as $\mathcal{N}(4, 6)$. The third asset, A_3, exhibits a Pareto distribution characterized by $f(x) = \frac{\alpha}{x^{\alpha+1}}$ for $x > 1$ with a parameter $\alpha = 1.5$. This distribution results in a mean return of 3 for A_3, but with an infinite variance, reflecting the heavy-tailed distributions commonly employed in financial modeling [39]. The agent's investment decisions are randomized, with the probability of choosing asset A_i denoted as $P(A_i) \sim \exp(\theta_i)$, where $\theta \in \mathbb{R}^3$ represents the policy parameters. Here we plot the results of running 50 iterations, with $10,000$ trajectories to estimate gradients in each iteration.

In the experiment, we choose the Radon-Nikodym derivative and χ^2-divergence as examples. Figure 1 illustrates how the probabilities of choosing A_1, A_2, and A_3 change over iterations. For Radon-Nikodym derivative, the agent is highly risk-averse at $\alpha = 0.95$, favoring A_1 and the agent is less risk-averse at$\alpha = 0.05$, resulting in shifts in probabilities. For χ^2, $P(A_i)$ also changes with different β. Notably, different ϕ-divergence reflects different risk preferences as the probability distribution differs. These results align with our theoretical analysis. Moreover, in comparison to the experimental results in [18], our method exhibits enhanced efficiency, achieving convergence with a small number of iterations, even when applied to more complex forms of risk measures.

B. Optimal Stopping Problem

In this section, we consider a more complex setup similar to the CVaR and EVaR policy gradient work [15] [22]. The environment is designed as an optimal stopping problem, where the state at each time step k is represented by $x = [k, c_k]$. Here, c_k denotes the cost at time k. The cost sequence $\{c_k\}_{k=0}^T$ is generated as follows: at each time step, the cost at the next time step either increases by a constant factor $f_u > 1$ (i.e., $c_{k+1} = f_u c_k$) with probability p, or decreases by a constant factor $f_d < 1$ (i.e., $c_{k+1} = f_d c_k$) with probability $1 - p$. The agent's task is to decide whether to accept the current cost



(a) Distribution of asset return. (b) PhiD-R with RN derivative (c) PhiD-R with χ^2-divergence

Fig. 1: Probability of selecting each asset versus training iterations, for policies generated by solving PhiD-R RL based on Radon-Nikodym derivative and χ^2-divergence.

($a_k = 1$) or wait ($a_k = 0$) at each time step. If the agent chooses to accept the cost or the time step reaches $k = T$, the cost is set to $\min(K, c_k)$, where K represents the cost threshold. However, if the agent chooses to wait, an additional cost of p_h is incurred. Hence, the discounted cost can be expressed as $J^\theta(x) = \sum_{k=0}^{T} \gamma^k \left(\mathbf{1}\{a_k = 1\} \min(K, c_k) + \mathbf{1}\{a_k = 0\} p_h\right)$.

Here we choose $x_0 = [1; 0]$, $p_h = 0.1$, $T = 20$, $K = 5$, $\gamma = 0.95$, $f_u = 2$, $f_d = 0.5$, $p = 0.65$, $N = 500,000$ and $\Theta = [-20, 20]^{\kappa_1}$, where the dimension of the basis function is $\kappa_1 = 64$. Furthermore, we implement radial basis functions (RBFs) to extract the features for each state and search over the class of Boltzmann policies

$$\left\{\theta : \{\theta_{x,a}\}_{x\in\mathcal{X}, a\in\mathcal{A}},\ \mu_\theta(a|x) = \frac{\exp(\theta_{x,a}^{\mathrm{T}} x_f(x))}{\sum_{a\in\mathcal{A}} \exp(\theta_{x,a}^{\mathrm{T}} x_f(x))}\right\},$$

where $x_f(x)$ is the feature chosen by RBF at state x.

We evaluate the effectiveness of our algorithm using various ϕ-divergences. First, we employ the Radon-Nikodym derivative as the ϕ-divergence, corresponding to the widely-used CVaR measure. Next, we consider the KL divergence, corresponding to the EVaR, a relatively recent risk measure adopted in risk-sensitive RL [9]. These first two choices demonstrate our approach's efficiency with popular risk measures, offering fresh perspectives on tackling these risk measures in risk-sensitive RL. Furthermore, we explore the χ^2 divergence, a common divergence in RL, yet without a designated risk measure defined by this divergence. This experiment highlights our algorithm's potential in addressing less clear or undefined risk measures, potentially inspiring new research on innovative risk measures. Finally, we utilize the squared Hellinger distance to underscore our algorithm's necessity and advantages over other policy gradient methods. The frequency distribution of costs under PhiD-R with different choices of ϕ-divergence is presented in Figure 2.

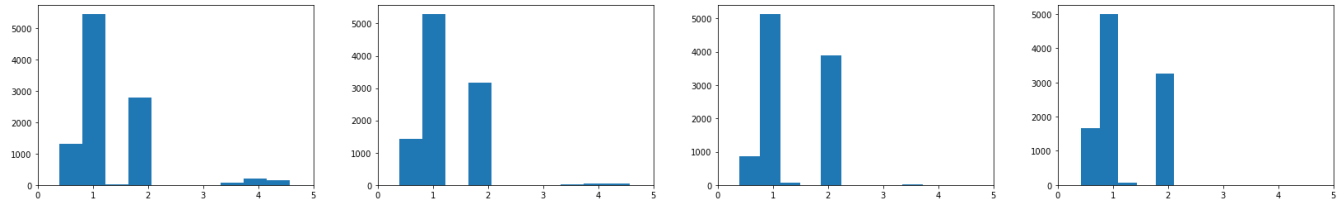
1) Radon-Nikodym Derivative (CVaR): We begin by selecting the ϕ-divergence as the Radon-Nikodym derivative, where $\phi(x) = 0$ for $0 \le x \le \frac{1}{1-\alpha}$ and $+\infty$ otherwise. In this case, the conjugate function $\phi^*(x)$ is

$$\frac{1}{1-\alpha}\max\{0, x\} = \frac{1}{\alpha}(0, x)^+.$$

By setting $\beta = 0$, the corresponding ϕ-divergence risk measure is CVaR and we obtain the following expressions.

$$\mathrm{CVaR}_\alpha(Z) = \inf_{\nu>0, \omega\in\mathbb{R}} \left\{\nu\omega + \frac{1}{1-\alpha}\mathbb{E}_P\left((Z - \nu\omega)^+\right)\right\}.$$

Notice that $\frac{d}{dx}\phi^*(x) = \frac{1}{1-\alpha}\mathbb{I}\{x > 0\}$, where $\mathbb{I}$ is the indicator function and $\frac{d^2}{dx^2}\phi^*(x) = 0$.

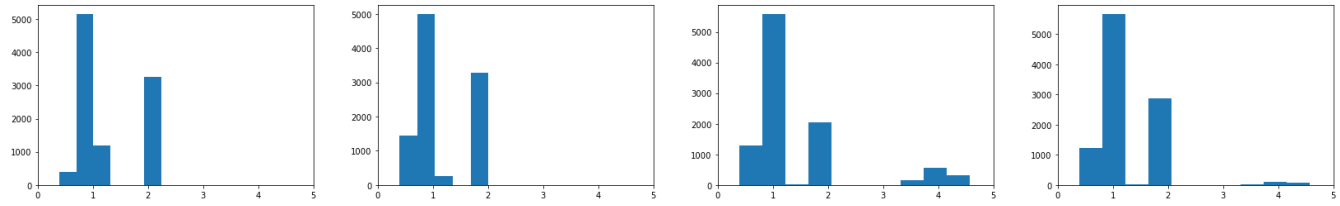


(a) $\alpha = 0.05$, $\beta = 0$, $\mathbb{E}[Z] = 1.35$ and CVaR $= 1.42$

(b) $\alpha = 0.95$, $\beta = 0$, $\mathbb{E}[Z] = 1.29$ and CVaR $= 2.49$

(c) $\alpha = 0.05$, $\beta = 3$, $\mathbb{E}[Z] = 1.36$ and EVaR $= 1.54$

(d) $\alpha = 0.95$, $\beta = 0.05$, $\mathbb{E}[Z] = 1.25$ and EVaR $= 2.67$



(e) $\beta = 0.95$, $\mathbb{E}[Z] = 1.31$ and $\text{PR}_{\chi^2,\beta} = 1.79$

(f) $\beta = 3$, $\mathbb{E}[Z] = 1.27$ and $\text{PR}_{\chi^2,\beta} = 1.83$

(g) $\beta = 0.95$, with $\mathbb{E}[Z] = 1.46$ and $\text{PR}_{\text{shd},\beta} = 1.54$

(h) $\beta = 3$, with $\mathbb{E}[Z] = 1.29$ and $\text{PR}_{\text{shd},\beta} = 1.97$

Fig. 2: Frequency distribution of costs under PhiD-R defined by: 1). Radon-Nikodym derivative (CVaR); 2) KL divergence (EVaR); 3) χ^2 divergence and 4) Squared Hellinger Distance with different choices of parameters.

By employing Algorithm 1 with CVaR update rules at various confidence levels α, we obtain the results in Figures 2a and 2b. The mean of discounted costs generated by following the optimal policy at $\alpha = 0.05$ exceeds the mean at $\alpha = 0.95$, whereas the opposite holds true for the CVaR value. The observed results align with the theoretical properties of CVaR. Specifically, when the risk aversion parameter (α) is set to 0.05, the agent exhibits a risk-averse behavior, opting for a safer strategy that results in higher costs but reduced risk exposure. Conversely, for $\alpha = 0.95$, the agent demonstrates risk-seeking tendencies, prioritizing lower costs despite the associated higher level of risk.

2) KL Divergence (EVaR): In this case, we choose the ϕ-divergence to be the KL divergence, denoted as $\phi(x) = x \log x$ for $x \geq 0$. Consequently, we have

$$\phi^*(x) = e^{x-1}$$

and $\beta = -\ln(1-\alpha)$ according to [27]. With this selection, the resulting ϕ-divergence risk measure corresponds to EVaR, given by

$$\text{EVaR}_\alpha(Z) = \inf_{\nu>0,\omega\in\mathbb{R}} \left\{ \nu \left[\omega + \mathbb{E}_P \left(e^{\frac{Z}{\nu} - \omega + \beta} \right) \right] \right\}.$$

By employing Algorithm 1 and incorporating EVaR update rules, we obtained results for two specific risk parameter settings: $\alpha = 0.05$ ($\beta = 3$) and $\alpha = 0.95$ ($\beta = 0.05$). For the case where $\alpha = 0.05$, the agent demonstrates a risk-averse preference by selecting higher costs to mitigate potential high risks. Conversely, for $\alpha = 0.95$, the agent exhibits a more aggressive behavior, seeking to minimize costs even in the presence of higher risks. Furthermore, the observation that EVaR consistently exceeds CVaR under the same distribution of a random variable aligns with the theoretical facts in [27].

3) χ^2 Divergence: In this case, we utilize the χ^2 divergence and set $\phi(x) = (x-1)^2$. Consequently, we obtain

$$\phi^*(x) = \frac{x^2}{4} + x$$

and $\beta > 0$. Thus, we have

$$\text{PR}_{\chi^2,\beta}(Z) = \inf_{\nu>0,\omega\in\mathbb{R}} \left\{ \nu \mathbb{E}_P \left(\frac{\left(\frac{Z}{\nu} - \omega + \beta\right)^2}{4} + \frac{Z}{\nu} + \beta \right) \right\}.$$

Applying Algorithm 1 with corresponding update rules for χ^2 divergence, we obtain the following results. Figure 2e and 2f illustrate that when $\beta = 0.95$, the mean of discounted costs exceeds the mean in the case where $\beta = 3$, while the risk value is lower. The selection of β indeed reflects the decision-maker's attitude towards risk.

4) Squared Hellinger Distance: In [22], the authors propose a two-update-rules trajectory-based policy gradient method to solve EVaR in risk-sensitive RL. However, the two-update-rules algorithm is not applicable to the entire class of ϕ-divergences. Here, we illustrate the necessity and practicality of our Algorithm 1 by employing the squared Hellinger distance as an example.

We choose $\phi(x) = (\sqrt{x} - 1)^2$ and the conjugate function is

$$\phi^*(x) = \frac{1}{1-x} - 1$$

for $x > 0$. Hence, this ϕ-divergence risk measure is given by

$$\text{PR}_{\phi,\beta}(Z) = \inf \left\{ \nu \left[\omega + \mathbb{E}_P \left(\frac{1}{1 - \frac{Z}{\nu} + \omega - \beta} - 1 \right) \right] \right\},$$

where $'\inf'$ is taken over the set $\left\{\nu, \omega : \nu > 0, \omega \in \mathbb{R}, \frac{Z}{\nu} - \omega + \beta > 0\right\}$.

Applying Algorithm 1 for the square Hellinger distance with varying values of β, we obtained the results shown in Figures 2g and 2h. The figures clearly indicate that when $\beta = 0.95$, the sample mean of the discounted cost is 1.46, which exceeds the mean in the case where $\beta = 3$, while the risk is lower. The selection of the parameter β directly reflects the risk preference exhibited by the agent.

We present a summary of numerical results for PhiD-R using various ϕ-divergences under different parameter settings in Table I, wher RN derivative means Radon-Nikodym derivative and SH distance means squared Hellinger distance. The data shows that all risk values exceed the mean and vary with parameter choices, validating the algorithm and demonstrating its alignment with established risk concerns. Additionally, these results offer insights into interpreting new risk measures, such as PhiD-R with χ^2-divergence and squared Hellinger distance, especially when supported by extensive simulations across diverse parameters. This adaptability in parameter selection highlights the flexibility of our approach, allowing decision-makers to align with their risk preferences while maintaining local optimality and efficiency.

TABLE I: Numerical results of different choices of ϕ-divergence.

ϕ-divergence	Parameters	Mean	PhiD-R
RN derivative	$\alpha = 0.05, \beta = 0$	1.35	1.42
	$\alpha = 0.95, \beta = 0$	1.29	2.49
KL-divergence	$\alpha = 0.05, \beta = 3$	1.36	1.54
	$\alpha = 0.95, \beta = 0.05$	1.25	2.67
χ^2-divergence	$\beta = 0.95$	1.31	1.79
	$\beta = 3$	1.27	1.83
SH distance	$\beta = 0.95$	1.46	1.54
	$\beta = 3$	1.29	1.79

C. Experiments on Gym

In this section, we validate our approach using OpenAI's Gym [40]. Specifically, we choose the `CartPole-v1` environment, which involves a pole attached to a cart moving along a frictionless track. The goal is to prevent the pole from falling over by applying force to the cart. The action space is discrete, with two possible actions: pushing the cart to the left or to the right. Since we consider cost in this work, we design the environment such that the agent receives a cost of 0 for every time step the pole remains upright and a cost of $+1$ for failing to keep the pole upright. The goal is to keep the pole balanced for as many time steps as possible, up to a maximum of T steps. The episode terminates if the pole angle exceeds $\pm 15^\circ$ or the cart moves more than 2.4 units from the center. We run both risk-neutral RL policy gradient and our approach with different choices of ϕ-divergence. For these experiments, we set $\gamma = 0.99$ and run $N = 10,000$ episodes with a time step of $T = 1,000$.

As shown in Figure 3, the results are plotted with episode length on the y-axis and episodes on the x-axis. Longer episode lengths indicate better performance, and changes in episode length over time illustrate the convergence speed of the algorithm. The upper panel presents the episode length for each individual episode, while the lower panel shows the mean episode length over the past 50 episodes using a sliding window. This results in a smoother blue curve (mean) in the lower panel, with the yellow shaded area around the blue curve providing a visual indication of the variability around the moving average. A smaller shaded area suggests more consistent and robust algorithm performance, while a larger shaded area indicates greater variability and less consistency. This visualization aids in understanding the stability of the training process over time. Figure 3 shows the result of running policy gradient method for risk-neutral RL.

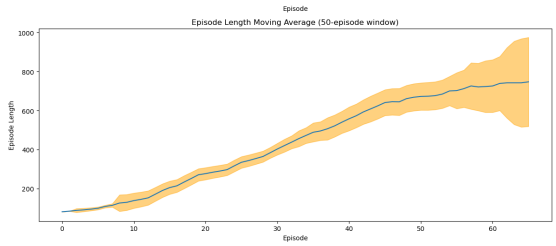
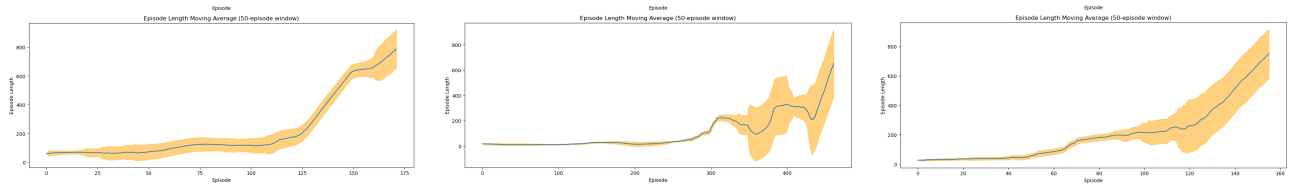


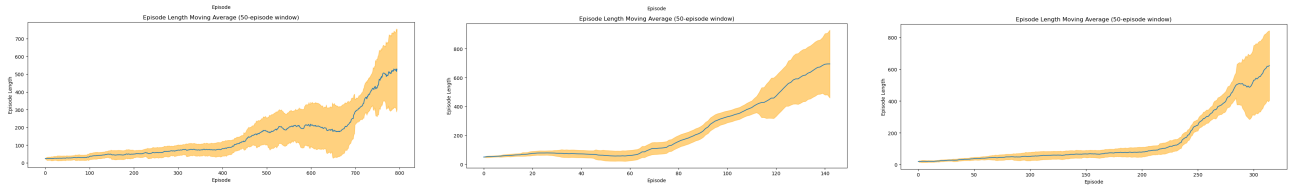
Fig. 3: Episode length versus episodes for risk-neutral RL.



(a) RN derivative with $\beta = 0, \alpha = 0.95$.

(b) RN derivative with $\beta = 0, \alpha = 0.1$

(c) KL divergence with $\beta = 0.1$



(d) KL divergence with $\beta = 3$

(e) χ^2 with $\beta = 0.1$

(f) χ^2 with $\beta = 3$

Fig. 4: Episode length versus episodes for PhiD-R defined by: 1) Radon-Nikodym derivative (CVaR); 2) KL divergence (EVaR); 3) χ^2-divergence with different choices of parameters.

We then apply our approach to the same environment using three different choices of divergence: 1) Radon-Nikodym derivative (Figure 4a, 4b), 2) KL divergence (Figure 4c, 4d), and 3) χ^2-divergence (Figure 4e, 4f). Although the training processes vary with different divergences and parameters, the overall trends are similar. When the divergence level β is smaller, the process converges more quickly since the agent is more risk-seeking and prefers taking more aggressive actions to balance the cart pole, as shown in Figure 4a, 4c, 4e. These processes are similar to risk-neutral RL, as all agents were more risk-seeking (risk-neutral implies risk-seeking behavior). Conversely, with larger β values, the agent exhibite more risk-averse behavior, indicated by a flatter curve during the initial phase, leading to stable episode lengths compared to the smaller β case (shown as Figure 4b, 4d, 4f). This behavior also suggests that the agent is more likely to be trapped in a local optimum.

VI. Conclusion

In this paper, we have applied a new class of risk measures named PhiD-R to risk-sensitive RL. We have proposed a trajectory-based policy gradient method tailored to this class of risk measures, utilizing an explicit representation that accommodates all forms of ϕ-divergence. Our approach has extended upon previous methods targeting specific risk measures and provided a comprehensive solution that encompasses the entire range of ϕ-divergence. Furthermore, we have demonstrated the convergence of our algorithms using a multi-time stochastic approximation approach. Through simulation experiments and numerical results, we have validated the efficiency and practical utility of our algorithms.

Appendix A
Computing Gradient Estimates

In this section, we provide the details of the gradient estimate computations. From the definition of $L(\nu, \omega, \theta)$, we obtain the following:

$$\nabla_\nu L(\nu,\omega,\theta) = \omega + \mathbb{E}_P\left[\phi^*\left(\frac{J^\theta(x_0)}{\nu} - \omega + \beta\right)\right] + \nu\nabla_\nu \mathbb{E}_P\left[\phi^*\left(\frac{J^\theta(x_0)}{\nu} - \omega + \beta\right)\right], \tag{16}$$

$$\nabla_\omega L(\nu,\omega,\theta) = \nu + \nu\nabla_\omega \mathbb{E}_P\left[\phi^*\left(\frac{J^\theta(x_0)}{\nu} - \omega + \beta\right)\right], \tag{17}$$

$$\nabla_\theta L(\nu,\omega,\theta) = \nu\nabla_\theta \mathbb{E}_P\left[\phi^*\left(\frac{J^\theta(x_0)}{\nu} - \omega + \beta\right)\right]. \tag{18}$$

1. $\nabla_\nu L(\nu, \omega, \theta)$: by expanding the expectation, we have

$$L(\nu,\omega,\theta) = \nu\left[\omega + \sum_\xi \mathbb{P}_\theta(\xi)\left(\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right)\right)\right].$$

By taking the gradient w.r.t. ν, we have

$$\begin{aligned}
\widehat{\nabla_\nu} L(\nu,\omega,\theta) &= \omega + \sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right) + \nu\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\nu\left[\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right)\right] \\
&= \omega + \sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right) - \nu\sum_\xi \mathbb{P}_\theta(\xi)\frac{J(\xi)}{\nu^2}\frac{\partial\phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} \\
&= \omega + \sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu} - \omega + \beta\right) - \sum_\xi \mathbb{P}_\theta(\xi)\frac{J(\xi)}{\nu}\frac{\partial\phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.
\end{aligned}$$

2. $\nabla_\omega L(\nu,\omega,\theta)$: by taking the gradient w.r.t. ω, we have

$$\begin{aligned}\widehat{\nabla_\omega} L(\nu,\omega,\theta) &= \nu + \nu\nabla_\omega\left[\sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu}-\omega+\beta\right)\right]\\ &= \nu + \nu\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\omega\left[\phi^*\left(\frac{J(\xi)}{\nu}-\omega+\beta\right)\right]\\ &= \nu - \nu\sum_\xi \mathbb{P}_\theta(\xi)\frac{\partial\phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.\end{aligned}$$

3. $\nabla_\theta L(\nu,\omega,\theta)$: by taking the gradient w.r.t. θ, we have

$$\begin{aligned}\widehat{\nabla_\theta} L(\nu,\omega,\theta) &= \nu\nabla_\theta\left[\sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu}-\omega+\beta\right)\right]\\ &= \nu\sum_\xi \nabla_\theta\mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu}-\omega+\beta\right)\\ &= \nu\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\theta\log\mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu}-\omega+\beta\right),\end{aligned}$$

where the last equality is due to $\nabla_\theta\mathbb{P}_\theta(\xi) = \mathbb{P}_\theta(\xi)\nabla_\theta\log\mathbb{P}_\theta(\xi)$.

APPENDIX B
PROOF OF THEOREM 2: CONVERGENCE ANALYSIS

In this section, we provide the detailed proof that was outlined in the proof sketch in the main content. We will begin by analyzing the multi-time scale discrete stochastic approximation and proceed through the convergence of the sequences $(\nu_k,\omega_k,\theta_k)$ to the local optimal solutions.

A. Convergence of ν-update

Since ν converges a faster time scale than ω and θ, we can regard ω and θ as fixed in the ν-update, i.e.,

$$\nu_{k+1} = \Gamma_{\mathcal{N}}\left[\nu_k - \zeta_1(k)\left(\omega + \sum_{j=1}^N \frac{1}{N}\phi^*\left(\frac{J(\xi_{j,k})}{\nu_k}-\omega+\beta\right) - \sum_{j=1}^N \frac{1}{N}\frac{J(\xi_{j,k})}{\nu_k}\frac{\partial\phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega+\beta}\right)\right].$$

Consider the continuous time dynamics of ν defined using differential inclusion

$$\dot{\nu} \in \Upsilon_\nu\left[-\widehat{\nabla_\nu} L(\nu,\omega,\theta)\right], \tag{19}$$

where

$$\Upsilon_\nu[G(\nu)] := \lim_{0<\eta\to 0}\frac{\Gamma_N(\nu+\eta G(\nu)) - \Gamma_N(\nu)}{\eta}.$$

Here $\Upsilon_\nu[G(\nu)]$ is the left directional derivative of the function $\Gamma_N(\nu)$ in the direction of $G(\nu)$. Using the left directional derivative $\Upsilon_\nu[G(\nu)]$ in the sub-gradient descent algorithm for ν ensures that the gradient points in the descent direction along the boundary of ν whenever the ν-update hits its boundary.

Now consider the following equation,

$$\nu_{k+1} = \Gamma_{\mathcal{N}}\left[\nu_k - \zeta_1(k)\left(\omega + \sum_\xi \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu_k}-\omega+\beta\right) - \sum_\xi \mathbb{P}_\theta(\xi)\frac{J(\xi)}{\nu_k}\frac{\partial\phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu_k}-\omega+\beta} + \delta_{\nu_{k+1}}\right)\right],$$

where

$$\delta_{\nu_{k+1}} = \left(\omega + \sum_{j=1}^{N} \frac{1}{N}\phi^*\left(\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta\right) - \sum_{j=1}^{N} \frac{1}{N}\frac{J(\xi_{j,k})}{\nu_k}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega+\beta}\right)$$
$$- \left(\omega + \sum_{\xi} \mathbb{P}_\theta(\xi)\phi^*\left(\frac{J(\xi)}{\nu_k} - \omega + \beta\right) - \sum_{\xi} \mathbb{P}_\theta(\xi)\frac{J(\xi)}{\nu_k}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu_k}-\omega+\beta}\right).$$

In order to show that the update rule converges to the solution of the o.d.e, we need to verify several conditions. Before going through this process, we firstly make the following assumptions, which will be used to guarantee the convergence of our algorithm.

Assumption 3. The parameters ν and ω are bounded, i.e., $\nu \in [V_{\min}, V_{\max}]$ and $\omega \in [W_{\min}, W_{\max}]$.

Assumption 4. Let $U_{\min}$ and $U_{\max}$ denote the bound for $u = \frac{J(\xi)}{\nu} - \omega + \beta$. The function ϕ satisfies:
1. The first derivative of the conjugate function ϕ^* is bounded in $[U_{\min}, U_{\max}]$.
2. The second derivative of the conjugate function ϕ^* is bounded in $[U_{\min}, U_{\max}]$.

In Lemma 1 in Appendix C, we show that $\widehat{\nabla_\nu} L(\nu, \omega, \theta)$ is Lipschitz continuous in ν. Given that the step size ζ_1 satisfies Assumption 2, we have $\sum_k \zeta_1(k) = \infty$ and $\sum_k \zeta_1^2(k) < \infty$. Furthermore, in Lemma 2 in Appendix C, we show that the sequence $\{\delta_{\nu_{k+1}}\}$ forms a martingale difference sequence. In addition, under Assumption 3, we have $\sup_k ||\nu_k|| < \infty$. With these conditions, we can invoke Corollary 4 in Chapter 5 of [34] to show that the update rule in our algorithm converges almost surely to the set $[V_{\min}, V_{\max}]$.

To complete the proof of convergence for the ν-update, we must show that the sequence converges to a fixed point of the o.d.e. (19). To establish this, we apply a Lyapunov stability analysis.

For any given ω and θ, define the following Lyapunov function

$$\mathcal{L}_{\omega,\theta}(\nu) = L(\nu, \omega, \theta) - L(\nu^*, \omega, \theta),$$

where ν^* is a minimum point.

To utilize the Lyapunov theory for asymptotically stable differential inclusions (Theorem 3.10 and Corollary 3.11 in [41]), we need to verify that the Lyapunov function defined above satisfies both Hypothesis 3.1 and Hypothesis 3.9 from [41].

We begin by verifying that the Lyapunov function satisfies Hypothesis 3.9, which requires showing that $\frac{d}{dt}\mathcal{L}_{\omega,\theta}(\nu) \leq 0$ and $\nabla_t \mathcal{L}_{\omega,\theta}(\nu)$ is non-zero if $\left\|\Gamma_N[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)]\right\| \neq 0$. Considering the continuous-time dynamics for ν, we have

$$\frac{d}{dt} L(\nu, \omega, \theta) = \widehat{\nabla_\nu} L(\nu, \omega, \theta) \Upsilon_N \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)\right].$$

Therefore, we obtain

$$\begin{aligned}
\frac{d}{dt}\mathcal{L}_{\theta,\lambda}(\nu) &= \widehat{\nabla_\nu} L(\nu, \omega, \theta) \Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)\right] - \widehat{\nabla_\nu} L(\nu, \omega, \theta) \Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)\right] \big|_{\nu=\nu^*} \\
&= \widehat{\nabla_\nu} L(\nu, \omega, \theta) \Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)\right] \\
&= \frac{d}{dt} \widehat{\nabla_\nu} L(\nu, \omega, \theta).
\end{aligned}$$

We need to demonstrate that $\frac{d}{dt}\mathcal{L}_{\omega,\theta}(\nu) \leq 0$ and that this quantity is non-zero whenever

$$\left\|\Gamma_N[-\widehat{\nabla_\nu} L(\nu, \omega, \theta)]\right\| \neq 0.$$

Case 1: $\nu \in (V_{\min}, V_{\max})$.

There exists a sufficiently small $\eta_0 > 0$ such that $\nu - \eta_0 \widehat{\nabla_\nu} L(\nu, \omega, \theta) \in [V_{\min}, V_{\max}]$ and

$$\Gamma_N \left[\nu - \eta_0 \widehat{\nabla_\nu} L(\nu, \omega, \theta) \right] = -\eta_0 \widehat{\nabla_\nu} L(\nu, \omega, \theta).$$

Recalling the definition of $\Upsilon \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta) \right]$, we obtain

$$\frac{d}{dt} L(\nu, \theta, \lambda) = - \left\| \widehat{\nabla_\nu} L(\nu, \omega, \theta) \right\| \leq 0$$

and $\frac{d}{dt} L(\nu, \theta, \lambda) < 0$ if $\widehat{\nabla_\nu} L(\nu, \omega, \theta) \neq 0$.

Case 2: $\nu \in \{V_{\min}, V_{\max}\}$.

Notice that there are two cases, which depend on whether the set

$$F(\nu) := \left\{ \widehat{\nabla_\nu} L(\nu, \omega, \theta) \middle| \forall \eta_0 > 0, \exists \eta \in [0, \eta_0] \text{ such that } \nu - \eta \widehat{\nabla_\nu} L(\nu, \omega, \theta) \notin [V_{\min}, V_{\max}] \right\}$$

is empty or not.

Case 2-1: $F(\nu)$ is empty.

Since $\nu \in \{V_{\min}, V_{\max}\}$ and $\nu - \eta \widehat{\nabla_\nu} L(\nu, \omega, \theta) \in [V_{\min}, V_{\max}]$, we know

$$\Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta) \right] = -\widehat{\nabla_\nu} L(\nu, \omega, \theta),$$

which implies that

$$\frac{d}{dt} L(\nu, \theta, \lambda) = - \left\| \widehat{\nabla_\nu} L(\nu, \omega, \theta) \right\| \leq 0$$

and $\frac{d}{dt} L(\nu, \theta, \lambda) < 0$ if $\widehat{\nabla_\nu} L(\nu, \omega, \theta) \neq 0$.

Case 2-2: $F(\nu)$ is not empty.

For any $\eta > 0$, define $\nu_\eta := \nu - \eta \widehat{\nabla_\nu} L(\nu, \omega, \theta)$. When $0 < \eta \to 0$, $\Gamma_N[\nu_\eta]$ is the projection of ν_η to the tangent space of $[V_{\min}, V_{\max}]$. For any $\hat{\nu} \in [V_{\min}, V_{\max}]$, since the set $\{\nu \in [V_{\min}, V_{\max}] : ||\nu - \nu_\eta||_2 \leq ||\hat{\nu} - \nu_\eta||_2\}$ is compact, then the project of ν_η on $[V_{\min}, V_{\max}]$ exists. Furthermore, since $g(\nu) = \frac{1}{2}(\nu - \nu_\eta)^2$ is a strongly convex function and $\nabla_\nu g(\nu) = \nu - \nu_\eta$. By the first order optimal condition, we obtains $\forall \nu \in [V_{\min}, V_{\max}]$,

$$\nabla g(\nu_\eta^*)(\nu - \nu_\eta^*) = (\nu_\eta^* - \nu_\eta)(\nu - \nu_\eta^*) \geq 0,$$

where ν_η^* is the unique projection of ν_η. Due to the uniqueness, we know only if $\nu = \nu_\eta^*$, the above equality holds. Therefore, for any $\nu \in [V_{\min}, V_{\max}]$ and $\eta > 0$,

$$\begin{aligned}
\widehat{\nabla_\nu} L(\nu, \omega, \theta) \Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta) \right] &= \widehat{\nabla_\nu} L(\nu, \omega, \theta) \lim_{0 < \eta \to 0} \frac{\nu_\eta^* - \nu}{\eta} \\
&= \lim_{0 < \eta \to 0} \frac{\nu - \nu_\eta}{\eta} \lim_{0 < \eta \to 0} \frac{\nu_\eta^* - \nu}{\eta} \\
&= \lim_{0 < \eta \to 0} \frac{-||\nu_\eta^* - \nu||^2}{\eta^2} + \lim_{0 < \eta \to 0} (\nu^* - \nu_\eta) \frac{\nu^* - \nu}{\eta^2} \leq 0.
\end{aligned}$$

Note that for any $\widehat{\nabla_\nu} L(\nu, \omega, \theta) \cap F(\nu)^c$, $\nu - \eta \widehat{\nabla_\nu} L(\nu, \omega, \theta) \in [V_{\min}, V_{\max}]$ for any $\eta \in [0, \eta_0]$ and some $\eta_0 > 0$. Thus, this follows the statement in the empty case.

Combining all these arguments, we conclude that $\frac{d}{dt} L(\nu, \omega, \theta) \leq 0$, and this inequality holds strictly whenever $\Upsilon_\nu \left[-\widehat{\nabla_\nu} L(\nu, \omega, \theta) \right] \neq 0$. As a result, $\frac{d}{dt} \mathcal{L}_{\omega, \theta}(\nu) \leq 0$ and remains non-zero under the same condition.

Having shown that $\mathcal{L}_{\omega, \theta}(\nu)$ satisfies Hypotheses 3.1 and 3.9, we can now apply the results from [41]. This ensures that the ν-update converges almost surely to the solution of the o.d.e. (19), which, in turn, converges to $\nu^* \in [V_{\min}, V_{\max}]$.

B. Convergence of ω-update

After establishing the convergence of the ν-update, we proceed to demonstrate the convergence of the ω-update. Given that ν converges on a faster timescale than ω, and θ operates on the slowest timescale, the ω-update can be expressed using the converged value $\nu^*(\omega)$ while treating θ as an invariant quantity, i.e.,

$$\omega_{k+1} = \Gamma_{\mathcal{R}}\left[\omega_k - \zeta_2(k)\left(\nu^*(\omega_k) - \nu^*(\omega_k)\sum_{j=1}^{N}\frac{1}{N}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu^*(\omega_k)}-\omega_k+\beta}\right)\right].$$

Considering the continuous time dynamic of ω,

$$\dot{\omega} \in \Upsilon_{\omega}\left[-\widehat{\nabla_{\omega}}L(\nu,\omega,\theta)\right], \tag{20}$$

where

$$\Upsilon_{\omega}[G(\omega)] := \lim_{0<\eta\to 0}\frac{\Gamma_{\mathcal{R}}(\omega+\eta G(\omega)) - \Gamma_{\mathcal{R}}(\omega)}{\eta}.$$

The ω-update can be rewritten as a stochastic approximation, i.e.,

$$\omega_{k+1} = \Gamma_{\mathcal{R}}\left[\omega_k - \zeta_2(k)\left(\nu^*(\omega_k) - \nu^*(\omega_k)\cdot\sum_{\xi}\mathbb{P}_{\theta}(\xi)\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu^*(\omega_k)}-\omega_k+\beta}\right) + \delta_{\omega_{k+1}}\right],$$

where

$$\begin{aligned}\delta_{\omega_{k+1}} = &-\left(\nu^*(\omega_k) - \nu^*(\omega_k)\sum_{\xi}\mathbb{P}_{\theta}(\xi)\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu^*(\omega_k)}-\omega_k+\beta}\right)\\ &+\left(\nu^*(\omega_k) - \nu^*(\omega_k)\sum_{j=1}^{N}\frac{1}{N}\frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu^*(\omega_k)}-\omega_k+\beta}\right).\end{aligned}$$

To demonstrate that the update rule converges to the solution of the o.d.e., we need to verify conditions similar to those established previously. In particular, in Lemma 3 of Appendix C, we show that $\widehat{\nabla_{\omega}}L(\nu,\omega,\theta)$ is Lipschitz continuous in ω. The step size $\zeta_2(k)$ satisfies $\sum_k \zeta_2(k) = \infty$ and $\sum_k \zeta_2^2(k) < \infty$, as stated in Assumption 2. Moreover, Assumption 3 ensures that $\sup_k ||\omega_k|| < \infty$.

Next, we focus on the Lyapunov analysis for the ω-update. For any fixed θ, we define the Lyapunov function as:

$$\mathcal{L}_{\theta}(\omega) = L(\nu^*(\omega),\omega,\theta) - L(\nu^*(\omega),\omega^*,\theta),$$

where ω^* is a local minimum point. Analogous to the approach used for the ν-update, we can express:

$$\frac{d}{dt}\mathcal{L}_{\theta}(\omega) = \frac{d}{dt}\widehat{\nabla_{\omega}}L(\nu^*(\omega),\omega,\theta).$$

Following a method similar to the Lyapunov analysis for the ν-update, we can show that $\frac{d}{dt}\mathcal{L}_{\theta}(\omega) \leq 0$ and that this quantity is strictly non-zero whenever $\left\|\Gamma_{\mathcal{R}}[-\widehat{\nabla_{\omega}}L(\nu,\omega,\theta|_{\nu=\nu^*(\omega)})]\right\| \neq 0$. Consequently, we demonstrate that the ω-update converges almost surely to the solution of the o.d.e. (19), which, in turn, converges to $\omega^* \in [W_{\min}, W_{\max}]$.

C. Local minimum

In this section, we aim to establish the convergence of the sequence $\{\nu_k, \omega_k\}$ towards a local minimum of the objective function $L(\nu, \omega, \theta)$, while keeping θ fixed. Building upon the arguments presented in the previous sections, we show that, for any given initial states $\nu(0)$ and $\omega(0)$, the sequences $\nu(t)$ and $\omega(t)$ converge to their respective optimal stationary points, ν^* and ω^*. This further implies

$$\begin{aligned} L(\nu^*, \omega^*, \theta) &\leq L(\nu(\omega^*(t)), \omega(t), \theta) \\ &\leq L(\nu(\omega(0)), \omega(0), \theta) \\ &\leq L(\nu(t), \omega(0), \theta) \\ &\leq L(\nu(0), \omega(0), \theta). \end{aligned}$$

We demonstrate the existence of a local minimum through contraction.

Suppose that (ν^*, ω^*) is not a local minimum, then there exits a point $(\bar{\nu}, \bar{\omega}) \in [V_{\min}, V_{\max}] \times [W_{\min}, W_{\max}] \cap \mathcal{B}_{(\nu^*, \omega^*)}(r)$ such that

$$L(\bar{\nu}, \bar{\omega}, \theta) = \min_{(\nu,\omega) \in [V_{\min}, V_{\max}] \times [W_{\min}, W_{\max}] \cap \mathcal{B}_{(\nu^*, \omega^*)}(r)} L(\nu, \omega, \theta).$$

The minimum is attained by the Weierstrass extreme value theorem. By setting $\omega(0) = \bar{\omega}$, we have

$$\begin{aligned} L(\bar{\nu}, \bar{\omega}, \theta) &= \min_{(\nu,\omega) \in [V_{\min}, V_{\max}] \times [W_{\min}, W_{\max}] \cap \mathcal{B}_{(\nu^*, \omega^*)}(r)} L(\nu, \omega, \theta) \\ &\leq L(\nu^*, \omega^*, \theta) \\ &\leq L(\bar{\nu}, \bar{\omega}, \theta), \end{aligned}$$

which is a contraction.

Therefore, (ν^*, ω^*) is a local minimum for $L(\nu, \omega, \theta)$ for any fixed θ.

D. Convergence of θ-update

Given that θ converges on the slowest timescale, we can express the θ-update as:

$$\theta_{k+1} = \Gamma_\Theta \left[\theta_k - \zeta_3(k) \left(\nu^*(\theta) \sum_{j=1}^{N} \frac{1}{N} \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,k}) \cdot \phi^* \left(\frac{J(\xi_{j,k})}{\nu^*(\theta)} - \omega^*(\theta) + \beta \right) \right) \right].$$

We now consider the following o.d.e. for θ:

$$\dot{\theta} \in \Upsilon_\theta \left[-\widehat{\nabla_\theta} L(\nu, \omega, \theta) \right], \tag{21}$$

where

$$\Upsilon_\theta[G(\theta)] := \lim_{0 < \eta \to 0} \frac{\Gamma_\Theta(\theta + \eta G(\theta)) - \Gamma_\Theta(\theta)}{\eta}.$$

The θ-update can be rewritten as a stochastic approximation, i.e.,

$$\theta_{k+1} = \Gamma_\Theta \left[\theta_k - \zeta_3(k) \cdot \left(\widehat{\nabla_\theta} L(\nu, \omega, \theta) \big|_{\nu = \nu^*(\theta_i), \omega = \omega_k, \theta = \theta_k} + \delta_{\theta_{k+1}} \right) \right],$$

where

$$\begin{aligned} \delta_{\theta_{k+1}} = &-\nu^*(\theta) \sum_{\xi} \mathbb{P}_\theta(\xi) \nabla_\theta \log \mathbb{P}_\theta(\xi) \cdot \phi^* \left(\frac{J(\xi)}{\nu^*(\theta)} - \omega^*(\theta_k) + \beta \right) \\ &+ \nu^*(\theta) \sum_{\xi} \frac{1}{N} \nabla_\theta \log \mathbb{P}_\theta(\xi) \cdot \phi^* \left(\frac{J(\xi)}{\nu^*(\theta)} - \omega^*(\theta_k) + \beta \right). \end{aligned}$$

To demonstrate that the update rule converges to the solution of the o.d.e., we need to verify several conditions. First, in Lemma 5 in Appendix C, we show that $\widehat{\nabla}_\theta L(\nu,\omega,\theta)$ is Lipschitz continuous in θ. Second, the step size $\zeta_3(k)$ satisfies $\sum_k \zeta_3(k)=\infty$ and $\sum_k \zeta_3^2(k)<\infty$, which follows from Assumption 2. Additionally, in Lemma 6 of Appendix C, we show that $\{\delta_{\omega_{k+1}}\}$ forms a martingale difference sequence. Finally, θ is in a compact and closed set Θ, which ensures that $\sup_k ||\theta_k|| < \infty$.

It remains to check the Lyapunov analysis for θ-update. The general idea here is same with the Lyapunov analysis above, but the difference here is that θ is vector other than a scalar. We first define the Lyapunov function

$$\mathcal{L}(\theta) = L(\nu^*(\theta),\omega^*(\theta),\theta) - L(\nu^*(\theta^*),\omega^*(\theta^*),\theta^*),$$

where θ^* is a local minimum point. Consider the continuous time dynamics for θ, we have

$$\frac{d}{dt}\mathcal{L}(\theta) = \frac{d}{dt}\widehat{\nabla}_\theta L(\nu^*(\theta),\omega^*(\theta),\theta).$$

It remains to show that $\frac{d}{dt}\widehat{\nabla}_\theta L(\nu^*(\theta),\omega^*(\theta),\theta)\le 0$ and the equality holds if and only if

$$\Upsilon_\theta\left[-\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right]=0.$$

There are three cases we have to consider.

Case 1: θ is in the interior of Θ (not on the boundary).

Since Θ is a compact closed set, there exists a sufficient small $\eta_0>0$ such that

$$\theta-\eta_0\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\in\Theta$$

and

$$\Gamma_\Theta\left(\theta-\eta_0\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right)-\theta=-\eta_0\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}.$$

Recall the definition of Υ_θ, we have

$$\frac{d}{dt}L(\nu,\omega,\theta)\bigg|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}=-\left\|\widehat{\nabla}_\theta L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right\|^2\le 0.$$

Furthermore, the equality only holds when $\frac{d}{dt}L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}=0$.

Case 2: θ is on the boundary of Θ and $\theta-\eta\widehat{\nabla}_\theta L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\in\Theta$ for any $\eta\in(0,\eta_0]$ and some $\eta_0>0$.

In this case, we have

$$\Upsilon_\theta\left[-\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right]=-\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}.$$

Therefore, we obtain

$$\frac{d}{dt}L(\nu,\omega,\theta)\bigg|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}=-\left\|\widehat{\nabla}_\theta L(\nu,\omega,\theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right\|^2\le 0.$$

Moreover, the equality only holds when $\widehat{\nabla}_\theta L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}=0$.

Case 3: θ is on the boundary of Θ but $\theta-\eta\widehat{\nabla}_\theta L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\notin\Theta$ for some $\eta\in(0,\eta_0]$ and any $\eta_0>0$.

For any $\eta>0$, we define

$$\theta_\eta=\theta-\eta\widehat{\nabla}_\theta L(\nu,\omega,\theta)|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}.$$

In this case, when $0 < \eta \to 0$, $\Gamma_\Theta[\theta_\eta]$ is the projection of θ_η to the tangent space of Θ. For any $\hat{\theta} \in \Theta$, since $\{\theta \in \Theta : ||\theta - \theta_\eta||_2 \leq \left\|\hat{\theta} - \theta_\eta\right\|_2\}$ is a compact set, the project of θ_η exists. Define $g(\theta) = \frac{1}{2}||\theta - \theta_\eta||_2^2$, since $g(\theta)$ is a strong convex function and $\nabla_\theta g(\theta) = \theta - \theta_\eta$, we obtain

$$\nabla g(\theta_\eta^*)^\top(\theta - \theta_\eta^*) = (\theta_\eta^* - \theta_\eta)^\top(\theta - \theta_\eta^*) \geq 0,$$

for any $\theta \in \Theta$, where θ_η^* is the projection of θ_η. Due to the uniqueness of this projection, the equality holds if and only if $\theta = \theta_\eta^*$. Therefore, for any $\theta \in \Theta$ and $\eta > 0$,

$$\begin{aligned}
&\left(\widehat{\nabla}_\theta L(\nu, \omega, \theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right)^\top \cdot \Upsilon_\nu \left[-\widehat{\nabla}_\nu L(\nu, \omega, \theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right] \\
&= \left(\widehat{\nabla}_\theta L(\nu, \omega, \theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right)^\top \lim_{0<\eta\to 0} \frac{\theta_\eta^* - \theta}{\eta} \\
&= \left(\lim_{0<\eta\to 0} \frac{\theta - \theta_\eta}{\eta}\right)^\top \lim_{0<\eta\to 0} \frac{\theta_\eta^* - \theta}{\eta} \\
&= \lim_{0<\eta\to 0} \frac{-||\theta_\eta^* - \theta||^2}{\eta^2} + \lim_{0<\eta\to 0} (\theta^* - \theta_\eta)^\top \frac{\theta^* - \theta}{\eta^2} \leq 0.
\end{aligned}$$

Combining all these arguments, we have $\frac{d}{dt}L(\nu^*(\theta), \omega^*(\theta), \theta) \leq 0$ and it is non-zero whenever

$$\Upsilon_\theta \left[-\widehat{\nabla}_\theta L(\nu, \omega, \theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right] \neq 0.$$

Therefore, we know that $\frac{d}{dt}\mathcal{L}(\theta) \leq 0$ and it is non-zero whenever $\Upsilon_\theta \left[-\widehat{\nabla}_\theta L(\nu, \omega, \theta)\big|_{\nu=\nu^*(\theta),\omega=\omega^*(\theta)}\right] \neq 0$. Now, we can establish the almost sure convergence of the θ-update to the solution of the o.d.e given by equation (21), which in turn converges to $\theta^* \in \Theta$.

Combining with the fact that (ν^*, ω^*) are local minimum for $L(\nu, \omega, \theta)$, we further conclude that θ^* is a local optimal policy for the ϕ-divergence optimization problem.

APPENDIX C
TECHNICAL LEMMAS

In this section, we present the technical lemmas that are used in the convergence analysis in the proof of Theorem 2. We begin by introducing the following propositions, derived from the definition of $\mathbb{P}_\theta$, which are crucial for demonstrating that the gradient estimates in Algorithm 1 are Lipschitz continuous. These results further aid in establishing the technical lemmas that will be discussed later.

Proposition 1. *By the definition of $\mathbb{P}_\theta(\xi)$ and $\nabla_\theta \log \mathbb{P}_\theta(\xi)$, we have*

$$\begin{aligned}
&\mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi) \\
&= P_0(x_0) \prod_{k=0}^{T-1} \pi(a_k|x_k, \theta) P(x_{k+1}|x_k, a_k) \sum_{k=0}^{T-1} \frac{\nabla_\theta \pi(a_k|x_k, \theta)}{\pi(a_k|x_k, \theta)} \\
&= P_0(x_0) \sum_{k=0}^{T-1} \prod_{i\neq k}^{T-1} \nabla_\theta \pi(a_k|x_k, \theta) \pi(a_k|x_k, \theta) P(x_{k+1}|x_k, a_k).
\end{aligned}$$

Combining Assumption 1 and the fact that the sum of products of Lipschitz function is Lipschitz, $\mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)$ and $\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)$ are Lipschitz in θ. Furthermore, since the gradient of Lipschitz function is bounded, we have

$$\left|\nabla_\theta \left(\sum_\xi \mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)\right)\right| \leq K_1(\xi).$$

Also,

$$\mathbb{E}[\nabla_\theta \log \mathbb{P}_\theta(\xi)] = \sum_\xi \mathbb{P}_\theta(\xi) \nabla_\theta \log \mathbb{P}_\theta(\xi) = 0.$$

Proposition 2. *By Assumption 1,* $\nabla_\theta \log \mathbb{P}_\theta(\xi)$ *is bounded, i.e.,* $|\nabla_\theta \log \mathbb{P}_\theta(\xi)| \leq K_2(\xi)$.

Lemma 1. $\widehat{\nabla_\nu} L(\nu, \omega, \theta)$ *is Lipschitz in* ν.

Proof. Recall that

$$\widehat{\nabla_\nu} L(\nu, \omega, \theta) = \omega + \sum_\xi \mathbb{P}_\theta(\xi) \phi^* \left(\frac{J(\xi)}{\nu} - \omega + \beta \right) - \sum_\xi \mathbb{P}_\theta(\xi) \frac{J(\xi)}{\nu} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.$$

Let $f(\nu)$ denote $\widehat{\nabla_\nu} L(\nu, \omega, \theta)$, we have

$$\begin{aligned}
f'(\nu) &= \sum_\xi \mathbb{P}_\theta(\xi) \nabla_\nu \left(\phi^* \left(\frac{J(\xi)}{\nu} - \omega + \beta \right) \right) - \sum_\xi \mathbb{P}_\theta(\xi) \nabla_\nu \left(\frac{J(\xi)}{\nu} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} \right) \\
&= -\sum_\xi \mathbb{P}_\theta(\xi) \frac{J(\xi)}{\nu^2} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} - \sum_\xi \mathbb{P}_\theta(\xi) \left(-\frac{J(\xi)}{\nu^2} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} - \frac{J^2(\xi)}{\nu^3} \frac{\partial \phi^*}{\partial u^2}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} \right) \\
&= -\sum_\xi \mathbb{P}_\theta(\xi) \frac{J(\xi)}{\nu^2} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} + \sum_\xi \mathbb{P}_\theta(\xi) \left(\frac{J(\xi)}{\nu^2} + \frac{J^2(\xi)}{\nu^3} \right) \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta} \\
&= \sum_\xi \mathbb{P}_\theta(\xi) \frac{J^2(\xi)}{\nu^3} \frac{\partial \phi^*}{\partial u^2}\bigg|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.
\end{aligned}$$

Notice that $J(\xi)$ is bounded by $[-\frac{C_{\max}}{1-\gamma}, \frac{C_{\max}}{1-\gamma}]$, ν is bounded by $[V_{\min}, V_{\max}]$ and ω is bounded by $[W_{\min}, W_{\max}]$. By Assumption 4, we know that $f'(\nu)$ is bounded. Thus, $\widehat{\nabla_\nu} L(\nu, \omega, \theta)$ is Lipschitz in ν. □

Lemma 2. $\{\delta_{\nu_{k+1}}\}$ *is a martingale difference sequence.*

Proof. Due to the fact that the trajectories are generated based on the sampling p.m.f and all these trajectories are independent, we have $\mathbb{E}[\delta_{\nu_{k+1}} | \mathcal{F}_{\nu,k}] = 0$ where $\mathcal{F}_{\nu,k} = \sigma(\nu_m, \delta_{\nu_m}, m \leq k)$ is the filtration of ν_k generated by different independent trajectories.

We need to prove that $\mathbb{E}[||\delta_{\nu_{k+1}}||^2 | \mathcal{F}_{\nu,k}]$ is bounded. Consider

$$\begin{aligned}
\delta_{\nu_{k+1}} &= \left(\omega + \sum_{j=1}^N \frac{1}{N} \phi^* \left(\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta \right) - \sum_{j=1}^N \frac{1}{N} \frac{J(\xi_{j,k})}{\nu_k} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega+\beta} \right) \\
&\quad - \left(\omega + \sum_\xi \mathbb{P}_\theta(\xi) \phi^* \left(\frac{J(\xi)}{\nu_k} - \omega + \beta \right) - \sum_\xi \mathbb{P}_\theta(\xi) \frac{J(\xi)}{\nu_k} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu_k}-\omega+\beta} \right) \\
&= -\sum_\xi \mathbb{P}_\theta(\xi) \phi^* \left(\frac{J(\xi)}{\nu_k} - \omega + \beta \right) + \sum_{j=1}^N \frac{1}{N} \phi^* \left(\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta \right) \\
&\quad - \sum_{j=1}^N \frac{1}{N} \frac{J(\xi_{j,k})}{\nu_k} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi_{j,k})}{\nu_k}-\omega+\beta} + \sum_\xi \mathbb{P}_\theta(\xi) \frac{J(\xi)}{\nu_k} \frac{\partial \phi^*}{\partial u}\bigg|_{u=\frac{J(\xi)}{\nu_k}-\omega+\beta}.
\end{aligned}$$

Notice that ϕ^* is a convex function and $\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta$ is bounded. Then, $\phi^*(\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta)$ is bounded. For convenience, we denote it as $\phi^*(\frac{J(\xi_{j,k})}{\nu_k} - \omega + \beta) \in [c_1, c_2]$, where $c_1, c_2 \in \mathbb{R}$. By Assumption 4, we have $\left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi)}{\nu_k}-\omega+\beta} \in [c_3, c_4]$, where $c_3, c_4 \in \mathbb{R}$. Then, we have

$$\delta_{\nu_{k+1}} \leq c_2 - c_1 + \frac{C_{\max}}{(1-\gamma)V_{\min}}(c_4 - c_3).$$

Let $c_5 \in \mathbb{R}$ denote the real value on the right side, we further have, $||\delta_{\nu_{k+1}}||^2 \leq (c_5)^2$, which implies $\{\delta_{\nu_{k+1}}\}$ is a martingale difference sequence. □

Lemma 3. *$\widehat{\nabla_\omega} L(\nu, \omega, \theta)$ is Lipschitz in ω.*

Proof. Recall that

$$\widehat{\nabla_\omega} L(\nu, \omega, \theta) = \nu - \nu \sum_\xi \mathbb{P}_\theta(\xi) \left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.$$

For convenience, denote $\widehat{\nabla_\omega} L(\nu, \omega, \theta)$ as $f(\omega)$. We have

$$f'(\omega) = \nu \sum_\xi \mathbb{P}_\theta(\xi) \left.\frac{\partial \phi^*}{\partial u^2}\right|_{u=\frac{J(\xi)}{\nu}-\omega+\beta}.$$

Recall that the second derivative of ϕ^* is bounded in a closed set and ν is also bounded. We know $f'(\omega)$ is bounded, thus, $\widehat{\nabla_\omega} L(\nu, \omega, \theta)$ is Lipschitz in ω. □

Lemma 4. *$\{\delta_{\omega_{k+1}}\}$ is a martingale difference sequence.*

Proof. Note that the trajectories are generated based on the sampling p.m.f and all these trajectories are independent, we have $\mathbb{E}[\delta_{\omega_{k+1}} | \mathcal{F}_{\nu,k}] = 0$ where $\mathcal{F}_{\omega,k} = \sigma(\omega_m, \delta_{\omega_m}, m \leq k)$ is the filtration of ω_k generated by different independent trajectories.

We now demonstrate that $\mathbb{E}[||\delta_{\omega_{k+1}}||^2 | \mathcal{F}_{\nu,k}]$ is bounded. Consider

$$\begin{aligned}
&\delta_{\omega_{k+1}} \\
&= -\left(\nu^*(\omega_k) - \nu^*(\omega_k) \sum_\xi \mathbb{P}_\theta(\xi) \left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi)}{\nu^*(\omega_k)}-\omega_k+\beta}\right) + \left(\nu^*(\omega_k) - \nu^*(\omega_k) \sum_{j=1}^N \frac{1}{N} \left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi_{j,k})}{\nu^*(\omega_k)}-\omega_k+\beta}\right) \\
&= \nu^*(\omega_k) \sum_\xi \mathbb{P}_\theta(\xi) \left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi)}{\nu^*(\omega_k)}-\omega_k+\beta} - \nu^*(\omega_k) \sum_{j=1}^N \frac{1}{N} \left.\frac{\partial \phi^*}{\partial u}\right|_{u=\frac{J(\xi_{j,k})}{\nu^*(\omega_k)}-\omega_k+\beta}.
\end{aligned}$$

Since the first derivative of ϕ^* is bounded in a closed set, for convenience, denote its bound as $[c_6, c_7]$, where $c_6, c_7 \in \mathbb{R}$, we have

$$\delta_{\omega_{k+1}} \leq V_{\max} |c_7 - c_6|.$$

Thus, $||\delta_{\omega_{k+1}}||^2$ is bounded, which further implies $\{\delta_{\omega_{k+1}}\}$ is a martingale difference sequence. □

Lemma 5. *$\widehat{\nabla_\theta} L(\nu, \omega, \theta)$ is Lipschitz in θ.*

Proof. Recall that

$$\widehat{\nabla_\theta} L(\nu, \omega, \theta) = \sum_\xi \mathbb{P}_\theta(\xi) \nabla_\theta \log \mathbb{P}_\theta(\xi) \phi^* \left(\frac{J(\xi)}{\nu} - \omega + \beta\right).$$

By Assumption 1 and 4, we know that $\nabla_\theta \mathbb{P}_\theta(\xi)$ and $\nabla_\theta \log \mathbb{P}_\theta(\xi)$ are Lipschitz in θ. By the fact that the sum of Lipschitz functions is Lipschitz, we know that $\widehat{\nabla_\theta} L(\nu, \omega, \theta)$ is Lipschitz in θ. □

Lemma 6. *$\{\delta_{\omega_{k+1}}\}$ is a martingale difference sequence.*

Proof. Since the trajectories are generated based on the sampling p.m.f and all these trajectories are independent, we have $\mathbb{E}[\delta_{\theta_{k+1}}|\mathcal{F}_{\nu,k}] = 0$ where $\mathcal{F}_{\theta,k} = \sigma(\theta_m, \delta_{\theta_m}, m \leq k)$ is the filtration of θ_k generated by different independent trajectories.

It remains to show $\mathbb{E}[||\delta_{\theta_{k+1}}||^2|\mathcal{F}_{\nu,k}]$ is bounded. Consider

$$\begin{aligned}\delta_{\theta_{k+1}} &= -\nu^*(\theta)\sum_{\xi}\mathbb{P}_\theta(\xi)\nabla_\theta \log \mathbb{P}_\theta(\xi)\cdot\phi^*\left(\frac{J(\xi)}{\nu^*(\theta)} - \omega^*(\theta_k) + \beta\right)\\ &\quad + \nu^*(\theta)\sum_{\xi}\frac{1}{N}\nabla_\theta \log \mathbb{P}_\theta(\xi)\cdot\phi^*\left(\frac{J(\xi)}{\nu^*(\theta)} - \omega^*(\theta_k) + \beta\right)\\ &\leq V_{\max}(K_1(\xi)+K_2(\xi))\max\left\{|U_{\min}|, |U_{\max}|\right\}.\end{aligned}$$

Thus, $||\delta_{\theta_{k+1}}||^2 \leq (c_8)^2$, where $c_8 = (V_{\max}K_1(\xi) + K_2(\xi))\max\{|U_{\min}|, |U_{\max}|\} \in \mathbb{R}$, which further implies that $\{\delta_{\theta_{k+1}}\}$ is a martingale difference sequence. □

REFERENCES

[1] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: John Wiley & Sons, 2014.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press, 2018.

[3] E. Delage and S. Mannor, "Percentile Optimization for Markov Decision Processes with Parameter Uncertainty," *Operations Research*, vol. 58, no. 1, pp. 203–213, 2010.

[4] N. Bäuerle and J. Ott, "Markov Decision Processes with Average-Value-at-Risk Criteria," *Mathematical Methods of Operations Research*, vol. 74, pp. 361–379, 2011.

[5] P. La and M. Ghavamzadeh, "Actor-Critic Algorithms for Risk-Sensitive MDPs," *Advances in Neural Information Processing Systems*, vol. 26, Dec, 2013.

[6] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-Sensitive Reinforcement Learning," *Neural Computation*, vol. 26, no. 7, pp. 1298–1328, 2014.

[7] Y. Fei, Z. Yang, Y. Chen, Z. Wang, and Q. Xie, "Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 384–22 395, Dec, 2020.

[8] L. Prashanth, M. C. Fu *et al.*, "Risk-Sensitive Reinforcement Learning via Policy Gradient Search," *Foundations and Trends® in Machine Learning*, vol. 15, no. 5, pp. 537–693, 2022.

[9] X. Ni and L. Lai, "Risk-Sensitive Reinforcement Learning via Entropic-VaR Optimization," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct, 2022, pp. 953–959.

[10] I. Greenberg, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Efficient Risk-Averse Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 639–32 652, Dec, 2022.

[11] K. Wang, N. Kallus, and W. Sun, "Near-Minimax-Optimal Risk-Sensitive Reinforcement Learning with CVaR," in *International Conference on Machine Learning*, Honolulu, HI, Jul, 2023, pp. 35 864–35 907.

[12] Q. Zhang, S. Leng, X. Ma, Q. Liu, X. Wang, B. Liang, Y. Liu, and J. Yang, "CVaR-Constrained Policy Optimization for Safe Reinforcement Learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[13] C. Cousins, E. Lobo, M. Petrik, and Y. Zick, "Percentile Criterion Optimization in Offline Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 36, Dec, 2023.

[14] X. Ni and L. Lai, "Robust Risk-Sensitive Reinforcement Learning with Conditional Value-at-Risk," *arXiv preprint arXiv:2405.01718*, 2024.

[15] Y. Chow and M. Ghavamzadeh, "Algorithms for CVaR Optimization in MDPs," *Advances in Neural Information Processing Systems*, vol. 27, Dec, 2014.

[16] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-Sensitive and Robust Decision-Making: A CVaR Optimization Approach," *Advances in Neural Information Processing Systems*, vol. 28, Dec, 2015.

[17] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the CVaR via Sampling," in *Proc. The AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Austin, TX, Feb, 2015.

[18] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy Gradient for Coherent Risk Measures," *Advances in Neural Information Processing Systems*, vol. 28, Dec, 2015.

[19] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-Constrained Reinforcement Learning with Percentile Risk Criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[20] R. Keramati, C. Dann, A. Tamkin, and E. Brunskill, "Being Optimistic to Be Conservative: Quickly Learning a CVaR Policy," in *Proc. The AAAI conference on artificial intelligence*, vol. 34, no. 04, New York, NY, Feb, 2020, pp. 4436–4443.

[21] Y. Fei, Z. Yang, and Z. Wang, "Entropic Risk-Sensitive reinforcement learning: A meta regret framework with function approximation," 2020.

[22] X. Ni and L. Lai, "Policy Gradient Based Entropic-VaR Optimization in Risk-Sensitive Reinforcement Learning," in *Proc. Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, Oct, 2022, pp. 1–6.

[23] J. L. Hau, M. Petrik, M. Ghavamzadeh, and R. Russel, "RASR: Risk-Averse Soft-Robust MDPs with EVaR and Entropic Risk," *arXiv preprint arXiv:2209.04067*, 2022.

[24] J. L. Hau, M. Petrik, and M. Ghavamzadeh, "Entropic Risk Optimization in Discounted MDPs," in *International Conference on Artificial Intelligence and Statistics*. Tamil Nadu, India: PMLR, Feb, 2023, pp. 47–76.

[25] X. Ni, G. Liu, and L. Lai, "Risk-Sensitive Reward-Free Reinforcement Learning with CVaR," in *Proc. International Conference on Machine Learning*, Vienna, Austria, Jul, 2024.

[26] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent Measures of Risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[27] A. Ahmadi-Javid, "Entropic Value-at-Risk: A New Coherent Risk Measure," *Journal of Optimization Theory and Applications*, vol. 155, pp. 1105–1123, 2012.

[28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust Region Policy Optimization," in *Proc. International Conference on Machine Learning*. Lille, France: PMLR, July, 2015, pp. 1889–1897.

[29] B. Belousov and J. Peters, "f-Divergence Constrained policy Improvement," *arXiv preprint arXiv:1801.00056*, 2017.

[30] A. Jain and A. Orlitsky, "A General Method for Robust Learning from Batches," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 775–21 785, Dec, 2020.

[31] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa, "Lmitation Learning as f-Divergence Minimization," in *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*. Springer, 2021, pp. 313–329.

[32] C. Gong, Q. He, Y. Bai, Z. Yang, X. Chen, X. Hou, X. Zhang, Y. Liu, and G. Fan, "The f-Divergence Reinforcement Learning Framework," *arXiv preprint arXiv:2109.11867*, 2021.

[33] C. P. Ho, M. Petrik, and W. Wiesemann, "Robust Phi-Divergence MDPs," *arXiv preprint arXiv:2205.14202*, 2022.

[34] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2009, vol. 48.

[35] A. Tamar, Y. Glassner, and S. Mannor, "Policy Gradients Beyond Expectations: Conditional Value-at-Risk." Citeseer, 2015.

[36] L. Prashanth, "Policy Gradients for CVaR-Constrained MDPs," in *Proc. International Conference on Algorithmic Learning Theory*, Bled, Slovenia, 2014, pp. 155–169.

[37] C. Ying, X. Zhou, H. Su, D. Yan, N. Chen, and J. Zhu, "Towards Safe Reinforcement Learning via Constraining Conditional Value-at-Risk," *arXiv preprint arXiv:2206.04436*, 2022.

[38] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of Conditional Value-at-Risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000.

[39] W. Sun, S. Rachev, F. Fabozzi, and P. Kalev, "Long-Range Dependence and Heavy Tailedness in Modelling Trade Duration," *Working Paper, University of Karlsruhe*, 2005.

[40] G. Brockman, "OpenAI Gym," *arXiv preprint arXiv:1606.01540*, 2016.

[41] M. Benaïm, "Dynamics of Stochastic Approximation Algorithms," in *Seminaire de Probabilites XXXIII*. Springer, 2006, pp. 1–68.