Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

DISTORTION RISK MEASURE-BASED DEEP REINFORCEMENT LEARNING

Jinyang Jiang[1,2], Bernd Heidergott[3], Jiaqiao Hu[4], and Yijie Peng[1,2]

[1]Wuhan Institute for Artificial Intelligence, Guanghua School of Management,
Peking University, Beijing, CHINA
[2]Xiangjiang Laboratory, Changsha, Hunan, CHINA
[3]Vrije Universiteit Amsterdam, Amsterdam, THE NETHERLANDS
[4]Dept. of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, USA

ABSTRACT

Mainstream reinforcement learning (RL) typically focuses on maximizing expected cumulative rewards. In this paper, we explore a risk-sensitive RL setting where the objective is to optimize the distortion risk measure (DRM), a criterion better reflecting human risk perception. We parameterize the action selection policy by neural networks and propose a novel policy gradient algorithm, DRM-based Policy Optimization (DPO), along with its accelerated variant, DRM-based Proximal Policy Optimization (DPPO), to address deep RL tasks with DRM objectives. DPO integrates three coupled recursions operating at different timescales to estimate gradient components and update parameters simultaneously. Our experiments provide numerical results across diverse scenarios, demonstrating that our proposed algorithms outperform the existing baselines under the DRM criterion.

1 INTRODUCTION

Reinforcement learning (RL) formulates sequential decision-making as Markov Decision Processes (MDPs) and keeps improving the control policy by leveraging historical data collected when interacting with the environment. Standard RL frameworks take the expectation of cumulative reward as their objective, which tends to produce risk-neutral policies. Given the profound achievements deep RL has made in games (Silver et al. 2016; Mnih et al. 2015), robotic control (Levine et al. 2016), and other fields, researchers are motivated to enhance the learning capacities of these algorithms. Despite these advances, the real-world deployment of RL has not scaled up commensurately, particularly in high-risk sectors, since overlooking the risk from uncertainty can lead to serious consequences, as even a minor shortfall in the last 1% service level may trigger widespread public events of a company.

Classical RL algorithms are trustworthy for reaching a well-behaved policy in tasks where the underlying mechanism is almost deterministic, which is usually not the case when humans participate in the interaction. To manage the impact of uncertainty, risk measures have been applied to emphasize various considerations during training, which is referred to as risk-sensitive RL in the literature. For instance, value-at-risk (VaR) and conditional value-at-risk (CVaR) are common choices to capture the tail behavior, while Cumulative Prospect Theory (CPT) (Tversky and Kahneman 1992) assigns different credits to gains and losses. These risk measures are integrated into mainstream RL as either objectives or constraints.

In this paper, we explore a risk-sensitive RL setting where the objective is to optimize the distortion risk measure (DRM). As a weighted integral of the cumulative distribution function (CDF), DRM provides a unified formulation for a wide range of risk measures by distorting the probability with different distortion functions as shown in Figure 1. The specific definitions are provided in later sections. The weight coefficients are associated with the derivatives of distortion functions. The derivative corresponding to CPT implies that two-sided tail behavior is given increased consideration, while other listed risk measures pay attention to one-sided tail. With careful configuration, DRM has a better capacity to reflect human risk perception.

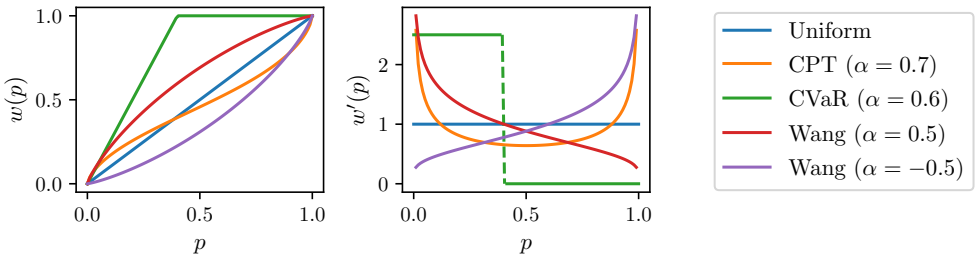


Figure 1: Visualization of example distortion functions and their derivatives.

We parameterize the action selection policy by deep neural networks which can consist of enormous policy parameters. Due to the high dimensionality, gradient-based optimization turns out to be the only reliable technical path. We transform the DRM into a weighted summation of quantiles, whose gradient can be expressed as a function of multiple distribution-related terms. Nonetheless, the gradient estimation for DRM is far beyond the likelihood ratio technique in classical RL algorithms, since the formulation involves complicated ratios and has variable coupling issues. We propose a novel policy gradient algorithm, DRM-based Policy Optimization (DPO), and its accelerated variant, DRM-based Proximal Policy Optimization (DPPO), to address deep RL tasks with DRM objectives. DPO integrates three coupled recursions operating at different timescales to estimate several components in the DRM gradient and update policy parameters simultaneously. We conduct experiments for our proposed algorithms across diverse scenarios, including inventory management examples. Our numerical results demonstrate that both algorithms outperform the existing baselines under the DRM criterion. To the best of our knowledge, we are the first to propose a three-timescale iterative algorithm under the DRM criterion in the deep RL context, addressing the large-scale optimization when deep neural networks are applied to represent policies.

The rest of the paper is organized as follows: In Section 2, we review the related work in the fields of deep RL and gradient estimation. In Section 3, we describe the MDP settings, the mean-based policy gradient theorem, and the DRM criterion for RL. Section 4 introduces our new algorithm, DPO, and its accelerated variant, DPPO. Theoretical results for DPO are listed in Section 5. Numerical results are presented in Section 6. We conclude the paper and discuss future directions in Section 7.

2 RELATED WORK

2.1 Mean-Based RL

Deep RL with expectation criterion has been the standard framework for years, which can be roughly divided into two categories. The first category parameterizes the control policy as certain end-to-end forms with current observation as input. As a pioneer, REINFORCE (Williams 1992) establishes the policy gradient theorem to adopt stochastic gradient ascent for policy optimization. Trust Region Policy Optimization (TRPO) (Schulman et al. 2015) leverages the importance sampling technique under complicated constraints to achieve better data utilization efficiency. Its simplified version, Proximal Policy Optimization (PPO) (Schulman et al. 2015) reduces the constraint to a clipped surrogate objective and has become one of the state-of-art baseline algorithms among mainstream RL algorithms. The second category estimates the action values given the state, with policies induced by selecting actions of the best value (Mnih et al. 2015; Hessel et al. 2018). Recently, it has been extended to estimate the value distribution (Bellemare et al. 2017). Others also make attempts to integrate both ideas (Lillicrap et al. 2015; Haarnoja et al. 2018).

2.2 Gradient Estimation of Risk Measures

Unlike mean-based objectives, even basic risk measures may impose huge difficulties in estimating their gradients, which motivates researchers to develop various techniques, such as kernel estimation (Liu and Hong 2009; Hong and Liu 2009), infinite perturbation analysis (Hong 2009; Jiang and Fu 2015), and

measure-valued differentiation (Heidergott and Volk-Makarewicz 2016). Glynn et al. (2021) further utilize the generalized likelihood ratio method to derive DRM estimators, which is the most relevant to our work. Unfortunately, previous methods only work when having perfect knowledge of underlying dynamics in the MDP, which violates the general settings in RL. Black-box gradient estimation approaches like simultaneous perturbation stochastic approximation (SPSA) (Spall 1992) and evolution strategy (Salimans et al. 2017) also used to be considered as alternatives but suffer from a significant performance decline with high-dimensional parameter spaces when neural networks are deployed.

2.3 Risk-Sensitive RL

Risk measures are introduced into the RL framework for producing safe and robust decision-making. Early work includes studies on the expected exponential utility (Borkar 2001) and CVaR (Petrik and Subramanian 2012; Tamar et al. 2014) as objectives. Prashanth and Ghavamzadeh (2013) and Prashanth et al. (2016) utilize SPSA to optimize variance-related risk measures and CPT. Jiang et al. (2022) develop a two-timescale iterative algorithm to optimize the VaR objective. Bertsekas (1997) employs a Lagrangian approach to solve risk-constrained RL. VaR and CVaR are also considered as risk constraints in Borkar and Jain (2014), Chow et al. (2018). Recently, Prashanth and Fu (2022) investigate the challenges involved in risk-sensitive RL using policy gradient methods. Moreover, there are value-based algorithms for managing the uncertainty in deep RL, where value distribution approximation and risk-sensitive greedy selection are combined to increase the decision robustness (Bellemare et al. 2017; Dabney et al. 2018).

3 PROBLEM FORMULATION & PRELIMINARIES

An MDP consists of the state space $\mathscr{S}$ and action space $\mathscr{A}$, as well as the transition kernel $p(s'|s,a)$, reward function $r(s,a,s')$, and reward discount factor $\eta \in (0,1)$. We consider the stochastic decision policy parameterized by a vector θ, which is a conditional density $\pi(\cdot|s;\theta)$ over $\mathscr{A}$ for every $s \in \mathscr{S}$. A trajectory $\tau = \{s_0, a_0, s_1, \cdots, s_T\}$ is generated from the interaction between the MDP dynamics and parametric policy, where $s_{t+1} \sim p(\cdot|s_t,a_t)$, $a_t \sim \pi(\cdot|s_t;\theta)$, s_0 follows some initializing density $p(\cdot)$, and T is the time horizon. The discounted cumulative reward is written as $R(\tau) = \sum_{t=0}^{T-1} \eta^t r(s_t,a_t,s_{t+1})$, which can be viewed as a function of τ and follows a distribution denoted by $F(\cdot;\theta)$.

In the context of classical deep RL, the objective is to solve the optimal policy that maximizes the expected cumulative reward, i.e., $\max_{\theta\in\Theta} \mathbb{E}_{R(\tau)\sim F(\cdot;\theta)}[R(\tau)]$, where Θ is the parameter space. The corresponding deep RL algorithms rely on gradient descent recursions, where the likelihood ratio method is frequently used to derive unbiased gradient estimators without any knowledge of the underlying dynamics. Let $\Pi(\tau;\theta) = p(s_0)\prod_{t=0}^{T-1} \pi(a_t|s_t;\theta)p(s_{t+1}|s_t,a_t)$. With mild integrability conditions (L'Ecuyer et al. 1992), the gradient of the expectation objective can be rewritten as

$$\nabla_\theta \mathbb{E}[R(\tau)] = \int_{\Omega_\tau} R(\tau)\nabla_\theta \log \Pi(\tau;\theta)\Pi(\tau;\theta)d\tau = \mathbb{E}\left[R(\tau)\sum_{t=0}^{T-1}\nabla_\theta \log \pi(a_t|s_t;\theta)\right]. \tag{1}$$

Advanced variants (Schulman et al. 2015; Schulman et al. 2017) may leverage the structural characteristics of MDPs and simulation-based techniques to obtain further acceleration.

We propose incorporating human risk perception into the learned policy by optimizing the DRM of cumulative rewards instead of maximizing expectations, which can be supported by CPT to serve as an alternative to basing decisions on expected utility theory (Tversky and Kahneman 1992). For a given distortion risk function $w(\cdot)$ with $w(0) = 0$ and $w(1) = 1$, the vanilla representation of DRM (Denuit et al. 2006) is written as

$$J(\theta) = \int_0^{+\infty} w(1-F(r;\theta))dr + \int_{-\infty}^{0} (w(1-F(r;\theta))-1)dr, \tag{2}$$

where the massively studied measures, e.g., VaR and CVaR, admit the formula when $w(y) = \mathbf{1}\{y > 1-\alpha\}$ and $w(y) = \min\{\frac{y}{1-\alpha}, 1\}$, respectively. Our generalized risk-sensitive deep RL aims at maximizing the

DRM on a compact convex set Θ, i.e.,

$$\max_{\theta\in\Theta} J(\theta). \tag{3}$$

The convexity of $w(\cdot)$ can lead to a risk-averse preference and vice versa. The direct gradient of the objective (2) can be given by $\nabla_\theta J(\theta) = \int_{-\infty}^{+\infty} -w'(1-F(x;\theta))\nabla_\theta F(x;\theta)dx$. Due to the numerical integration, algorithms can not avoid compromising to truncate the integral region, which requires problem-specific knowledge of rewards and results in huge computation wastes.

Glynn et al. (2021) provide an alternative gradient formulation with a finite integration region. If $w(\cdot)$ is non-decreasing and left-continuous, then the equality (2) can be transformed into a Lebesgue-Stieltjes integral over a bounded interval (Dhaene et al. 2012), i.e.,

$$J(\theta) = \int_0^1 F^{-1}(1-y;\theta)dw(y). \tag{4}$$

Its gradient is given by the weighted integral of VaR gradients, also known as quantile gradients, which can be obtained by applying the implicit function theorem (Fu et al. 2009), i.e.,

$$\nabla_\theta J(\theta) = -\int_0^1 \nabla_\theta F^{-1}(y;\theta)d\tilde{w}(y) = -\int_0^1 -\frac{\nabla_\theta F(q;\theta)}{f(q;\theta)}\bigg|_{q=F^{-1}(y;\theta)} d\tilde{w}(y), \tag{5}$$

where $\tilde{w}(y) = w(1-y)$ and $f(\cdot;\theta)$ is the density of the cumulative reward. There are three major difficulties in deriving an estimator for the gradient (5):

- The DRM gradient (5) is in the integral form that is hard to calculate analytically.
- The density function $f(\cdot;\theta)$ usually does not have an explicit form in the deep RL context.
- While the quantile function $F^{-1}(\cdot;\theta)$ is involved in the right-hand-side of the DRM gradient (5), its formula remains unknown and the value changes as the underlying θ varies.

4 DRM-BASED POLICY OPTIMIZATION

4.1 On-Policy Deep RL for Optimizing DRM

We first partition the integration interval $[0,1]$ into N uniform segments $[\frac{i-1}{N}, \frac{i}{N}]$ for $i = 1,\cdots,N$, and consider the algorithm framework that the policy optimization follows the numerical integration for computing the DRM gradient (5), which is given by

$$\hat{\nabla}_\theta J(\theta) = -\sum_{i=1}^{N} d_i(\theta)(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N})). \tag{6}$$

where $d_i(\theta) = -\frac{\nabla_\theta F(q;\theta)}{f(q;\theta)}|_{q=F^{-1}(\frac{2i-1}{2N};\theta)}$ is the true gradient of the $\frac{2i-1}{2N}$-quantile. This essentially leads the algorithm to solve the approximated problem, i.e.,

$$\max_{\theta\in\Theta} \hat{J}(\theta) = \max_{\theta\in\Theta} -\sum_{i=1}^{N} F^{-1}(\frac{2i-1}{2N};\theta)(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N})), \tag{7}$$

and $\hat{\nabla}_\theta J(\theta) = \nabla_\theta \hat{J}(\theta)$. The difference between $J(\theta)$ and $\hat{J}(\theta)$ diminishes with N increasing. The numerical integration substitute enables us to focus on tracing a series of quantiles and their gradients. Though Hu et al. (2022) propose coupled double recursions to track a single quantile and its gradient, this approach can be difficult to apply in deep RL since the density $f(\cdot;\theta)$ relies on the explicit forms of the transition probability and reward function that are typically unavailable.

Instead of using a large batch of samples for plug-in estimation each time, we propose a method to estimate different components of the DRM gradient and update the policy parameters simultaneously. This approach involves three recursions running on multiple timescales. The first two recursions trace a set of quantiles that occurred in the equality (7) and their gradients, while the last one updates the policy parameter θ. Given the current policy parameter θ, the quantile estimates can be found using the following recursive procedure:

$$q_{k+1,i} = q_{k,i} + \gamma_k^q \left(\frac{2i-1}{2N} - \mathbf{1}\{R(\tau_k) \leq q_{k,i}\}\right), \quad \text{for } i = 1, \cdots, N, \tag{8}$$

where γ_k^q is the step-size, τ_k is the trajectory simulated by following the policy $\pi(\cdot|\cdot;\theta)$, and $q_{k,i}$ is an estimate of $F^{-1}(\frac{2i-1}{2N};\theta)$ at the k-th iteration. Note that $F(r;\theta) = \mathbb{E}[\mathbf{1}\{R(\tau) \leq r\}]$, the recursion (8) is a stochastic approximation (SA) iteration for solving $F(q;\theta) = \frac{2i-1}{2N}$.

Since the quantile gradient $d_i(\theta)$ takes a complicated ratio form, it has no analytical formula to estimate as a whole and faces the ratio bias issue when the numerator and denominator are estimated separately with finite samples. Note that the definition of $d_i(\theta)$ can be transformed into $-\nabla_\theta F(q;\theta)|_{q=F^{-1}(\frac{2i-1}{2N};\theta)} - d_i(\theta) f(F^{-1}(\frac{2i-1}{2N};\theta);\theta) = 0$. We construct the following recursions to track the quantile gradients when θ and $\{q_i\}_{i=1}^N$ are fixed, i.e.,

$$D_{k+1,i} = D_{k,i} + \gamma_k^D \left(G_1(\tau_k;\theta,q_i) - G_2(\tau_k;\theta,q_i) D_{k,i}\right), \quad \text{for } i = 1, \cdots, N, \tag{9}$$

where γ_k^D is the step-size, $G_1(\tau;\theta,q_i)$ and $G_2(\tau;\theta,q_i)$ are estimators of $-\nabla_\theta F(q;\theta)|_{q=q_i}$ and $f(q_i;\theta)$, respectively, and $D_{k,i}$ is an estimate of $d_i(\theta)$ at the k-th iteration. With the likelihood ratio technique similar to that in the equality (1), we can rewrite the CDF gradient as follows:

$$\begin{aligned}
\nabla_\theta F(q;\theta) &= \nabla_\theta \mathbb{E}[\mathbf{1}\{R(\tau) \leq q\}] = \nabla_\theta \int_{\Omega_\tau} \mathbf{1}\{R(\tau) \leq q\} \Pi(\tau;\theta) d\tau \\
&= \mathbb{E}[\mathbf{1}\{R(\tau) \leq q\} \nabla_\theta \log \Pi(\tau;\theta)] = \mathbb{E}\left[\mathbf{1}\{R(\tau) \leq q\} \sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|s_t;\theta)\right],
\end{aligned}$$

which yields $G_1(\tau;\theta,q) = -\mathbf{1}\{R(\tau) \leq q\} \sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|s_t;\theta)$. However, only if we have an explicit and differentiable density function $f(\cdot;\theta)$, can we employ the generalized likelihood ratio method (Lei et al. 2018) to derive an unbiased estimator. Because this is usually not the case for the deep RL settings, we use the kernel-based method (Wand and Jones 1994) to estimate the density value, i.e., $G_2(\tau;\theta,q,h) = \frac{1}{h} K(\frac{R(\tau)-q}{h})$, where the additional parameter h is the bandwidth, $K(\cdot)$ denotes a kernel function satisfying $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$ and $\int_{\mathbb{R}} x^2 K(x)dx < \infty$, e.g., the density function of a standard normal distribution $K(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}$. To control the estimation variance, we also require the kernel function to be square integrable, i.e., $\int_{\mathbb{R}} K^2(x)dx < \infty$, which is a common setting in kernel-based methods. Although $G_2(\tau;\theta,q,h)$ is not an unbiased estimator of $f(q;\theta)$, $\mathbb{E}[G_2(\tau;\theta,q,h)] - f(q;\theta)$ can be controlled by varying h.

In the last recursion, we use estimated quantile gradients instead of $d_i(\theta_k)$ to compute the numerical integral (6) and perform the gradient ascent method for solving the problem (7). This leads to our proposed DRM-based Policy Optimization (DPO) algorithm as follows:

$$D_{k+1,i} = D_{k,i} + \gamma_k^D \left(G_1(\tau_k;\theta_k,q_{k,i}) - G_2(\tau_k;\theta_k,q_{k,i},h_k) D_{k,i}\right), \qquad \text{for } i = 1, \cdots, N, \tag{10}$$

$$q_{k+1,i} = q_{k,i} + \gamma_k^q \left(\frac{2i-1}{2N} - \mathbf{1}\{R(\tau_k) \leq q_{k,i}\}\right), \qquad \text{for } i = 1, \cdots, N, \tag{11}$$

$$\theta_{k+1} = \varphi_\Theta\Big(\theta_k + \gamma_k^\theta \sum_{i=1}^{N} -D_{k,i} \big(\tilde{w}\big(\frac{i}{N}\big) - \tilde{w}\big(\frac{i-1}{N}\big)\big)\Big), \tag{12}$$

where γ_k^θ is the gradient search step-size, $\varphi_\Theta(\cdot)$ represents a projection operation that brings an iterate θ_{k+1} back to the parameter space Θ whenever it becomes infeasible, and h_k is a decreasing bandwidth of kernel smoothing. The recursion (12) is equivalent to

$$\theta_{k+1} = \theta_k + \gamma_k^\theta \sum_{i=1}^{N} -D_{k,i}(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N})) + \gamma_k^\theta P_k, \tag{13}$$

where $\gamma_k^\theta P_k := \theta_{k+1} - \theta_k + \gamma_k^\theta \sum_{i=1}^{N} D_{k,i}(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N}))$ is the real vector with the smallest L^2-norm needed to keep $\{\theta_k\}$ in Θ, i.e., $P_k \in -C(\theta_{k+1})$, where $C(\theta)$ is the normal cone to Θ at θ. The iterative update only requires at least one sample to perform the estimation and enables us to utilize the historical samples during the past steps. The pseudo-code of DPO is presented in Algorithm 1.

Algorithm 1 DRM-based Policy Optimization (DPO)

1: **Input:** Policy network $\pi(\cdot|\cdot;\theta)$, distortion funcion $w(\cdot)$.
2: **Initialize:** Policy parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$, quantile estimates $\{q_{0,i}\}_{i=1}^N \in \mathbb{R}$ and quantile gradient estimates $\{D_{0,i}\}_{i=1}^N \in \mathbb{R}^d$.
3: **for** $k = 0, \cdots, K-1$ **do**
4: Generate one episode $\tau_k = \{s_0^k, a_0^k, s_1^k, \cdots, a_{T-1}^k, s_T^k\}$ following policy $\pi(\cdot|\cdot;\theta_k)$;
5: **for** $i = 1, \cdots, N$ **do** $D_{k+1,i} \leftarrow D_{k,i} + \gamma_k^D(G_1(\tau_k;\theta_k,q_{k,i}) - G_2(\tau_k;\theta_k,q_{k,i},h_k)D_{k,i})$ **end for**;
6: **for** $i = 1, \cdots, N$ **do** $q_{k+1,i} \leftarrow q_{k,i} + \gamma_k^q(\frac{2i-1}{2N} - \mathbf{1}\{R(\tau_k) \le q_{k,i}\})$ **end for**;
7: $\theta_{k+1} \leftarrow \varphi_\Theta\big(\theta_k + \gamma_k^\theta \sum_{i=1}^N -D_{k,i}(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N}))\big)$.
8: **end for**
9: **Output:** Trained policy network $\pi(\cdot|\cdot;\theta_K)$.

4.2 Off-Policy Deep RL for Optimizing DRM

For the state-of-art mean-based deep RL algorithms such as the one in Schulman et al. (2017), the policy parameters can be updated multiple times before generating new simulated data. In contrast, DPO updates parameters only once per episode and drops the data when parameters change, potentially leading to inefficiency in leveraging simulation samples. Define $[k] = K\lfloor \frac{k}{K} \rfloor$, where $K > 0$ is an integer. By applying the importance sampling technique, we can obtain an accelerated variant of DPO, namely DRM-based Proximal Policy Optimization (DPPO), which uses the trajectory generated at the $[k]$-th iteration to update the parameters for the next K iterations. The recursions (10) and (11) are transformed into

$$D_{k+1,i} = D_{k,i} + \gamma_k^D \rho_k(G_1(\tau_{[k]};\theta_k,q_{k,i}) - G_2(\tau_{[k]};\theta_k,q_{k,i},h_k)D_{k,i}), \qquad \text{for } i = 1, \cdots, N, \tag{14}$$

$$q_{k+1,i} = q_{k,i} + \gamma_k^q(\frac{2i-1}{2N} - \rho_k \mathbf{1}\{R(\tau_{[k]}) \le q_{k,i}\}), \qquad \text{for } i = 1, \cdots, N, \tag{15}$$

where $\rho_k = \frac{\Pi(\tau_{[k]};\theta_k)}{\Pi(\tau_{[k]};\theta_{[k]})} = \prod_{t=0}^{T-1} \frac{\pi(a_{[k],t}|s_{[k],t};\theta_k)}{\pi(a_{[k],t}|s_{[k],t};\theta_{[k]})}$ is the importance sampling ratio. Recursions (10) and (11) are equivalent to recursions (14) and (15) in a probability sense (Glynn and Iglehart 1989).

However, the importance sampling technique can cause increasing estimation variance when θ_k differs from $\theta_{[k]}$ too much, which may lead to an explosion. Mean-based TRPO (Schulman et al. 2015) manages to control the difference by imposing an additional constraint $D_{\mathrm{KL}}(\Pi(\cdot,\theta_k)\|\Pi(\cdot,\theta_{[k]})) \le \delta$, where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler (KL) divergence, but PPO reduces it to an unconstrained optimization of the surrogate objective with clipping and achieves reliable outcome without heavy computation consumption (Schulman et al. 2017). With the inspiration of clipping the importance sampling ratio in PPO, a similar option can be provided to DPPO to exclude the samples where the ratio deviates significantly from 1. It can be implemented by using $\tilde{\rho}_k = \rho_k \mathbf{1}\{1-\varepsilon \le \rho \le 1+\varepsilon\}$ instead of the original ρ_k in recursions (14) and (15), where ε is the tolerance.

5 CONVERGENCE ANALYSIS

Let $(\Omega, \mathscr{F}, P)$ be a probability space. We simplify the notations $\{q_{k,i}\}_{i=1}^N$ and $\{D_{k,i}\}_{i=1}^N$ as q_k and D_k, respectively, and define $\mathscr{F}_k = \sigma\{\theta_0, q_0, D_0, \cdots, \theta_k, q_k, D_k\}$ as the filtration generated by our algorithm for $k = 0, 1, \cdots$. Though only one simulation sample per iteration is acceptable in practice with careful choices of hyperparameters, we consider the following recursions instead of recursions (10) and (11) in convergence analysis for a more flexible trade-off between the bias and variance of estimation in kernel smoothing:

$$D_{k+1,i} = D_{k,i} + \gamma_k^D \Big(\frac{1}{B_k}\sum_{n=1}^{B_k}(G_1(\tau_{k,n};\theta_k,q_{k,i}) - G_2(\tau_{k,n};\theta_k,q_{k,i},h_k)D_{k,i})\Big), \qquad \text{for } i = 1,\cdots,N, \tag{16}$$

$$q_{k+1,i} = q_{k,i} + \gamma_k^q \Big(\frac{2i-1}{2N} - \frac{1}{B_k}\sum_{n=1}^{B_k}\mathbf{1}\{R(\tau_{k,n}) \le q_{k,i}\}\Big), \qquad \text{for } i = 1,\cdots,N, \tag{17}$$

where B_k and $\tau_{k,n}$ denote the batch size and the n-th trajectory at the k-th iteration, respectively. Here, we introduce some assumptions before the analysis.

Assumption 1 For $i = 1,\cdots,N$, $F^{-1}(\frac{2i-1}{2N};\theta) \in C^1(\Theta)$.

Assumption 2 The step-size sequences $\{\gamma_k^D\}$, $\{\gamma_k^q\}$, $\{\gamma_k^\theta\}$, batch size sequence $\{B_k\}$, and bandwidth sequence $\{h_k\}$ satisfy

(a) $\gamma_k^D > 0$, $\sum_{k=0}^\infty \gamma_k^D = \infty$, $\sum_{k=0}^\infty (\gamma_k^D)^2 < \infty$; (b) $\gamma_k^q > 0$, $\sum_{k=0}^\infty \gamma_k^q = \infty$, $\sum_{k=0}^\infty (\gamma_k^q)^2 < \infty$;
(c) $\gamma_k^\theta > 0$, $\sum_{k=0}^\infty \gamma_k^\theta = \infty$, $\sum_{k=0}^\infty (\gamma_k^\theta)^2 < \infty$; (d) $\gamma_k^\theta = o(\gamma_k^q)$, $\gamma_k^q = o(\gamma_k^D)$;
(e) $B_k > 0$, $h_k > 0$, $h_k^2 + B_k^{-1}h_k^{-1} = o(1)$.

Assumption 3 The log gradient of neural network outputs w.r.t. θ is bounded, i.e., there exists a constant $C_\pi > 0$, for any state-action pair $(s,a) \in \mathscr{S} \times \mathscr{A}$ and parameter $\theta \in \Theta$, $\sup_{a,s,\theta} \|\nabla_\theta \log \pi(a|s;\theta)\| < C_\pi$.

Assumption 4 There exist constants $\varepsilon_f \in (0,1)$ and $C_f > 0$ such that for $i = 1,\cdots,N$, w.p.1,

(a) $\inf_k f(q_{k,i};\theta_k) > \varepsilon_f$ and $\sup_k f(q_{k,i};\theta_k) < C_f$; (b) $\sup_k |f'(q_{k,i};\theta_k)| < \infty$.

Assumption 1 is a common setting in continuous optimization and implies the smoothness of DRM objectives. Assumption 2 is standard in multi-timescale SA analysis. Assumption 3 is naturally made to avoid computational overflow. Assumption 4 includes necessary conditions to control the bias and variance in kernel estimation and to ensure that the quantile gradients are well-defined. We expect the recursions (16), (17), and (12) or (13) to track three coupled ordinary differential equations (ODEs):

$$\begin{aligned}
\dot{D}_i(t) &= -\nabla_\theta F(q;\theta(t))\big|_{q=q_i(t)} - f(q_i(t);\theta(t))D_i(t), \quad \text{for } i = 1,\cdots,N,\\
\dot{q}_i(t) &= \frac{2i-1}{2N} - F(q_i(t);\theta(t)), \qquad\qquad \text{for } i = 1,\cdots,N,\\
\dot{\theta}(t) &= -\sum_{i=1}^N D_i(t)(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N})) + p(t),
\end{aligned} \tag{18}$$

where $p(t) \in -C(\theta(t))$ is the vector with the smallest norm needed to keep $\theta(t)$ in Θ. Due to the relationships of three step-sizes in Assumption 2, the later ODE(s) can be viewed as static for analyzing the dynamic of the former one(s). Recursion (12) can be regarded as tracking the ODE

$$\dot{\theta}(t) = -\sum_{i=1}^N d_i(\theta(t))(\tilde{w}(\frac{i}{N}) - \tilde{w}(\frac{i-1}{N})) + p(t), \tag{19}$$

whereas other variables are always close to converging. If $\theta^* = \arg\max_{\theta\in\Theta}\hat{J}(\theta)$ is the unique global asymptotically stable equilibrium of (19), then strong convergence of DPO to the optimum can be established. Therefore, we first prepare the following analysis on asymptotic behaviors of decoupled ODEs as shown in Lemmas 1-3.

Lemma 1 For all $\bar{\theta} \in \Theta$ and $i = 1, \cdots, N$, $F^{-1}(\frac{2i-1}{2N}; \bar{\theta})$ is the unique global asymptotically stable equilibrium of the ODE $\dot{q}_i(t) = \frac{2i-1}{2N} - F(q_i(t); \bar{\theta})$.

Lemma 2 For all $\bar{\theta} \in \Theta$, $i = 1, \cdots, N$, and $\bar{q}_i \in \mathbb{R}$, $\nabla_\theta F^{-1}(\frac{2i-1}{2N}; \bar{\theta})$ is the unique global asymptotically stable equilibrium of the ODE $\dot{D}_i(t) = -\nabla_\theta F(q; \bar{\theta})\big|_{q=F^{-1}(\frac{2i-1}{2N}; \bar{\theta})} - f(F^{-1}(\frac{2i-1}{2N}; \bar{\theta}); \bar{\theta}) D_i(t)$.

Lemma 3 If $\hat{J}(\theta)$ is strictly concave on Θ, then θ^* is the unique global asymptotically stable equilibrium of the ODE (19).

To prove that recursions (16), (17) and (12) track coupled ODEs (18), we can iteratively apply the following convergence theorem for single-timescale SA with bias term, which is slightly modified from Theorem 2.3 in Chapter 5 of Kushner and Yin (1997).

Theorem 1 Consider a recursion $x_{k+1} = x_k + \gamma_k(h(x_k) + m_k + b_k)$, where $h(\cdot)$ is bounded; $\{\gamma_k\}$ satisfies $\gamma_k > 0$, $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k^2 < \infty$; $\{m_k\}$ is a square-integrable martingale difference sequence, i.e., there exists a constant $K > 0$ such that $\mathbb{E}[m_{k+1}|\mathscr{F}_k] = 0$ and $\mathbb{E}[\|m_{k+1}\|^2|\mathscr{F}_k] \leq K$ a.s., for $k \geq 0$, with sigma fields denoted as $\mathscr{F}_k = \sigma\{x_0, \cdots, x_k\}$; and $\{b_k\}$ is $o(1)$. If the sequence $\{x_k\}$ is bounded and the ODE $\dot{x}(t) = h(x(t))$ is well-posed, then $\{x_k\}$ converges a.s. to some limit set of the ODE or the unique global asymptotically stable equilibrium, if one exists; otherwise, if a projection onto some compact set is applied to $\{x_k\}$, then it converges a.s. to the counterpart of the corresponding projected ODE.

The main difficulty of applying the result of Theorem 1 lies in showing the boundedness of the sequences generated by DPO. This is established in Lemma 4 and Theorem 2.

Lemma 4 If Assumptions 1 and 2(b) hold, then the sequence $\{q_k\}$ generated by recursion (11) is bounded w.p.1, i.e., for $i = 1, \cdots, N$, $\sup_k |q_{k,i}| < \infty$ w.p.1.

Theorem 2 If Assumptions 2(a), (e), 3, and 4 hold, then the sequence $\{D_k\}$ generated by recursion (10) is bounded w.p.1, i.e., for $i = 1, \cdots, N$, $\sup_k \|D_{k,i}\| < \infty$ w.p.1.

Now we have the following main convergence result for DPO.

Theorem 3 If Assumptions 1-4 hold and $\hat{J}(\theta)$ is strictly concave on Θ, then the sequences $\{D_k\}$ $\{q_k\}$ and $\{\theta_k\}$ generated by recursions (16), (17) and (12) converge to the unique optimal solution $\{\{d_i(\theta^*)\}, \{F^{-1}(\frac{2i-1}{2N}; \theta^*)\}, \theta^*\}$ of problem (7) w.p.1.

The concavity assumption is to guarantee convergence to the global optimum; otherwise, it can be relaxed, and the algorithm will then converge to a stationary point of the ODE (19). The detailed proofs of the lemmas and theorems mentioned above will be presented in the full journal version.

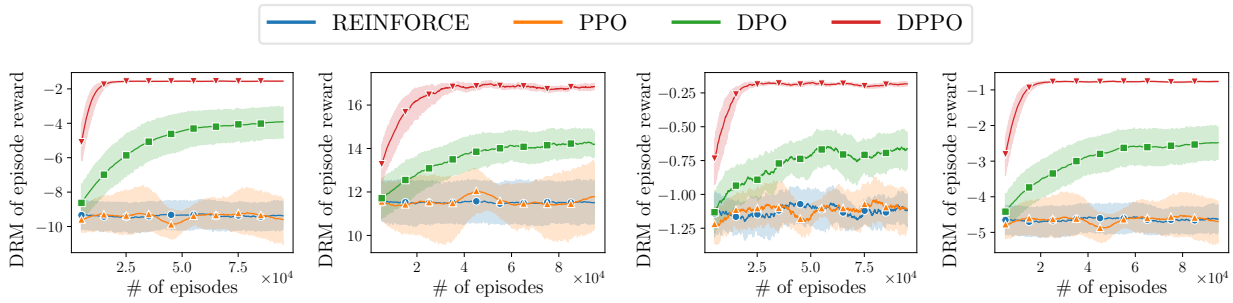
6 EXPERIMENTS

In this section, we conduct simulation experiments on different RL tasks to evaluate our DPO and DPPO against baseline algorithms within the same framework, specifically REINFORCE and PPO. Since black-box algorithms exhibit a rather low efficiency in optimizing policies parameterized by deep neural networks and the issue is exacerbated by the costly DRM evaluation, they are not included in the formal comparison.

6.1 Fair Lottery

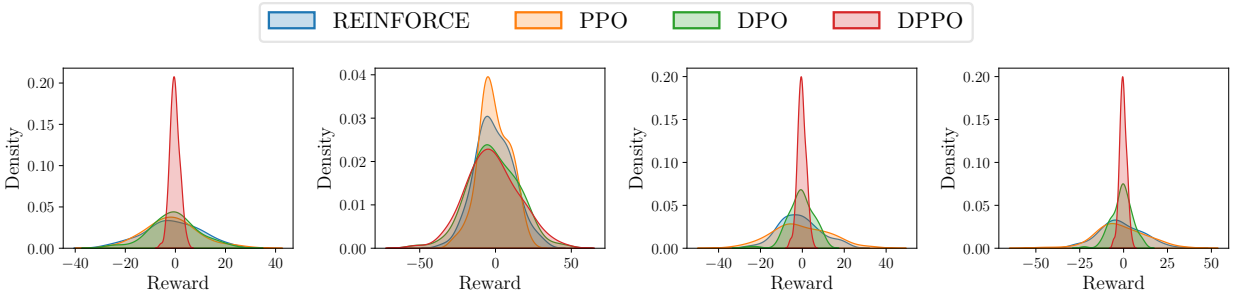
We conduct experiments for our algorithms on an actuarially fair lottery game. The state $s_t = (s_t^1, \cdots, s_t^N) \in \mathbb{R}^N$ is a random vector generated by shuffling the components of s_0 and represents the current risk levels of each lottery. The agent makes purchasing decisions among all alternatives at each time step, i.e., $a_t \in \{1, \cdots, N\}$. The step reward is then sampled from Uniform$(-s_t^{a_t}, s_t^{a_t})$, where the support of reward distribution is determined by the selected state entry. Since the expectation of cumulative rewards is fixed at zero, the risk-neutral algorithms are expected to fail in this environment, while others can still learn some strategies according to their risk preferences.

We test REINFORCE, PPO, DPO, and DPPO in a unified environment setting, where the agent is offered two options each time and makes decisions in 10 time steps. The purchasing policy is parameterized as a neural network consisting of two hidden layers with each containing 16 neurons. We configure the DRM with different distortion functions to explore various risk preferences. In addition to VaR and CVaR, we implement the CPT and DRM proposed in Wang (2000) with distortion functions $w(p) = \frac{p^\alpha}{(p^\alpha+(1-p)^\alpha)^{1/\alpha}}$ and $w(p) = \Phi(\Phi^{-1}(p)+\alpha)$, respectively. The learning curves based on 10 independent experiments are presented in Figure 2. The DRM value is estimated by rewards in the past 100 episodes and smoothed via averaging with window sizes of 100 for better visibility. The shaded areas represent the 95% confidence bands. It can be observed that both DPO and DPPO significantly improve the agent's DRM performance across all risk configurations, while mean-based REINFORCE and PPO continue to oscillate around the initial level. Moreover, DPPO demonstrates a remarkable advantage in data utilization efficiency and converges to a high level at an early stage. We test trained agents with 100 replications and present the KDE plots of rewards as shown in Figure 3. The concentration of rewards reflects the risk preferences of agents, and all outcomes of DPO and DPPO are consistent with the properties of corresponding risk measures, where distributions are concentrated under risk-averse criteria and vice versa.



(a) VaR $(\alpha = 0.25)$ (b) CVaR $(\alpha = 0.7)$ (c) CPT $(\alpha = 0.7)$ (d) Wang $(\alpha = -0.5)$

Figure 2: DRM learning curves of REINFORCE, PPO, DPO, and DPPO under different distortion function configurations in Fair Lottery examples.

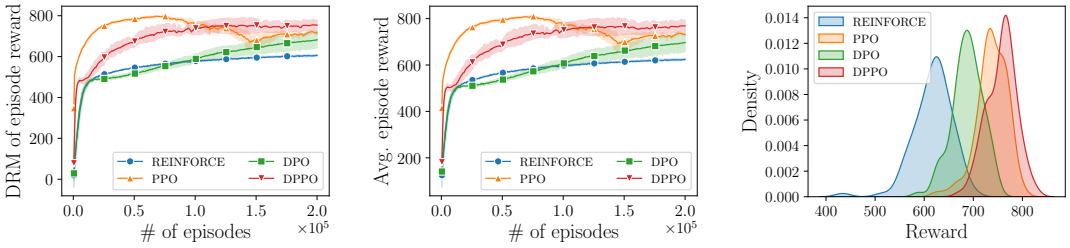


(a) VaR $(\alpha = 0.25)$ (b) CVaR $(\alpha = 0.7)$ (c) CPT $(\alpha = 0.7)$ (d) Wang $(\alpha = -0.5)$

Figure 3: KDE plots for rewards of REINFORCE, PPO, DPO, and DPPO under different distortion function configurations in Fair Lottery examples.

6.2 Inventory Management

We consider the inventory management problem within an N-echelon supply chain during T periods, where the two ends represent the customer and manufacturer. We use S_t^i, U_t^i and I_t^i to denote the products shipped by echelon i, the lost sales and on-hand inventory of echelon i at the end of period t. The reordering



(a) DRM learning curves (b) Average learning curves (c) KDE plots

Figure 4: Comparison of REINFORCE, PPO, DPO, and DPPO in the multi-echelon Inventory Management example.

quantity of echelon i is denoted as q_t^i, with q_t^0 as the randomly generated exogenous demand. In each period t, the agent has to select the best replenishment quantities $q_t = \{q_t^i\}_{i=1}^N$ for intermediate echelons. The shipped quantities are given by $S_t^i = q_t^{i-1} - [q_t^{i-1} - I_{t-1}^i - S_{t-L^i}^{i+1}]^+$, for $i = 1, \cdots, N$, where $[x]^+ = \max\{0, x\}$, L_i denotes the transit duration of echelon i, and the manufacturer is assumed to satisfy all demands from the last echelon, i.e., $S_t^{N+1} = q_t^N$. Then the lost sales and on-hand inventory are computed by $U_t^i = [q_t^{i-1} - S_t^i]^+$ and $I_t^i = [I_{t-1}^i + S_{t-L^i}^{i+1} - S_t^i]^+$. The profit of each echelon is calculated by $P_t^i = p^i S_t^i - p^{i+1} S_t^{i+1} - h^i I_t^i - l^i U_t^i$, where p^i, h^i and l^i are the unit price, unit holding cost and unit penalty for lost sales, respectively. The RL states cover the inventories, lost sales, shipped and ordering quantities in past $L = \max_i L_i$ periods, and the step rewards are determined by current overall profits, i.e., $\sum_{i=1}^N P_t^i$.

We test REINFORCE, PPO, DPO, and DPPO with the multi-echelon inventory management problem in which customer demands are generated by using a periodical stochastic model with a saw-wave trend. We choose DRM in Wang (2000) with $\alpha = -0.5$ as the DRM criterion. The agent policy is represented by a neural network consisting of two temporal convolutional layers and three parallel fully connected blocks to control each echelon. The agent makes replenishment decisions in 100 time steps. The learning curves obtained through 5 independent experiments and testing KDE plotted based on 100 replications are shown in Figure 4, where the shaded areas in the first two subfigures represent the 95% confidence bands. DRM-based algorithms outperform the corresponding baselines with the same architectures under both criteria, which suggests that it may be more desirable to address multi-echelon inventory management problems from a risk-sensitive perspective. The KDEs of rewards obtained by DPO and DPPO are more concentrated than those obtained by REINFORCE and PPO, indicating that the policies trained by DRM-based algorithms are more risk-averse. Furthermore, both off-policy algorithms, PPO and DPPO, have significant improvements in training efficiency compared to the other two algorithms.

7 CONCLUSION

In this paper, we propose a DPO algorithm and its accelerated variant DPPO to address deep RL tasks under DRM criteria from the perspective of policy optimization. Both algorithms utilize the multi-timescale SA technique, enabling simultaneous updates of partial estimates within the gradient function and policy parameters with a few simulation runs. Numerical experiments demonstrate that our algorithms effectively optimize the DRM objectives across various distortion function configurations. In future work, the algorithms have the potential for further enhancement by leveraging more structural features of MDPs.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 72325007, 72250065, and 72022001, as well as the Special Funds for Guiding Local Scientific and Technological Development by the Central Government under Grant 2023EGA035.

REFERENCES

Bellemare, M. G., W. Dabney, and R. Munos. 2017. "A Distributional Perspective on Reinforcement Learning". In *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh, 449–458: Proceedings of Machine Learning Research.

Bertsekas, D. P. 1997. "Nonlinear Programming". *Journal of the Operational Research Society* 48(3):334–334.

Borkar, V. and R. Jain. 2014. "Risk-Constrained Markov Decision Processes". *IEEE Transactions on Automatic Control* 59(9):2574–2579.

Borkar, V. S. 2001. "A Sensitivity Formula for Risk-Sensitive Cost and the Actor-Critic Algorithm". *Systems & Control Letters* 44(5):339–346.

Chow, Y., M. Ghavamzadeh, L. Janson, and M. Pavone. 2018. "Risk-Constrained Reinforcement Learning with Percentile Risk Criteria". *Journal of Machine Learning Research* 18(167):1–51.

Dabney, W., M. Rowland, M. Bellemare, and R. Munos. 2018. "Distributional Reinforcement Learning with Quantile Regression". In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32. Palo Alto, California, USA: Association for the Advancement of Artificial Intelligence Press.

Denuit, M., J. Dhaene, M. Goovaerts, and R. Kaas. 2006. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. John Wiley & Sons.

Dhaene, J., A. Kukush, D. Linders, and Q. Tang. 2012. "Remarks on Quantiles and Distortion Risk Measures". *European Actuarial Journal* 2:319–328.

Fu, M. C., L. J. Hong, and J.-Q. Hu. 2009. "Conditional Monte Carlo Estimation of Quantile Sensitivities". *Management Science* 55(12):2019–2027.

Glynn, P. W. and D. L. Iglehart. 1989. "Importance Sampling for Stochastic Simulations". *Management Science* 35(11):1367–1392.

Glynn, P. W., Y. Peng, M. C. Fu, and J.-Q. Hu. 2021. "Computing Sensitivities for Distortion Risk Measures". *INFORMS Journal on Computing* 33(4):1520–1532.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, 1861–1870: Proceedings of Machine Learning Research.

Heidergott, B. and W. Volk-Makarewicz. 2016. "A Measure-Valued Differentiation Approach to Sensitivities of Quantiles". *Mathematics of Operations Research* 41(1):293–317.

Hessel, M., J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney *et al*. 2018. "Rainbow: Combining Improvements in Deep Reinforcement Learning". In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32. Palo Alto, California, USA: Association for the Advancement of Artificial Intelligence Press.

Hong, L. J. 2009. "Estimating Quantile Sensitivities". *Operations Research* 57(1):118–130.

Hong, L. J. and G. Liu. 2009. "Simulating Sensitivities of Conditional Value at Risk". *Management Science* 55(2):281–293.

Hu, J., Y. Peng, G. Zhang, and Q. Zhang. 2022. "A Stochastic Approximation Method for Simulation-Based Quantile Optimization". *INFORMS Journal on Computing* 34(6):2889–2907.

Jiang, G. and M. C. Fu. 2015. "On Estimating Quantile Sensitivities via Infinitesimal Perturbation Analysis". *Operations Research* 63(2):435–441.

Jiang, J., Y. Peng, and J. Hu. 2022. "Quantile-Based Policy Optimization for Reinforcement Learning". In *2022 Winter Simulation Conference (WSC)*, 2712–2723 https://doi.org/10.1109/WSC57314.2022.10015456.

Kushner, H. J. and G. G. Yin. 1997. *Stochastic Approximation Algorithms and Applications*. Springer.

Lei, L., Y. Peng, M. C. Fu, and J.-Q. Hu. 2018. "Applications of Generalized Likelihood Ratio Method to Distribution Sensitivities and Steady-State Simulation". *Discrete Event Dynamic Systems* 28:109–125.

Levine, S., C. Finn, T. Darrell, and P. Abbeel. 2016. "End-to-End Training of Deep Visuomotor Policies". *Journal of Machine Learning Research* 17(1):1334–1373.

Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa *et al*. 2015. "Continuous Control with Deep Reinforcement Learning". *arXiv preprint arXiv:1509.02971*.

Liu, G. and L. J. Hong. 2009. "Kernel Estimation of Quantile Sensitivities". *Naval Research Logistics (NRL)* 56(6):511–525.

L'Ecuyer, P., N. Giroux, and P. W. Glynn. 1992. "Experimental Results for Gradient Estimation and Optimization of a Markov Chain in Steady-State". In *Simulation and Optimization*, 14–23. Springer.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al*. 2015. "Human-Level Control through Deep Reinforcement Learning". *Nature* 518(7540):529–533.

Petrik, M. and D. Subramanian. 2012. "An Approximate Solution Method for Large Risk-Averse Markov Decision Processes". In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 805–814. Arlington, Virginia, USA: Association for Uncertainty in Artificial Intelligence Press.

Prashanth, L. and M. C. Fu. 2022. "Risk-Sensitive Reinforcement Learning via Policy Gradient Search". *Foundations and Trends® in Machine Learning* 15(5):537–693.

Prashanth, L. and M. Ghavamzadeh. 2013. "Actor-Critic Algorithms for Risk-Sensitive MDPs". In *Advances in Neural Information Processing Systems*, edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Volume 26: Curran Associates, Inc.

Prashanth, L., C. Jie, M. Fu, S. Marcus and C. Szepesvári. 2016. "Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control". In *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger, Volume 48, 1406–1415. New York, New York, USA: Proceedings of Machine Learning Research.

Salimans, T., J. Ho, X. Chen, S. Sidor and I. Sutskever. 2017. "Evolution Strategies as a Scalable Alternative to Reinforcement Learning". *arXiv preprint arXiv:1703.03864*.

Schulman, J., S. Levine, P. Abbeel, M. Jordan and P. Moritz. 2015. "Trust Region Policy Optimization". In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, 1889–1897. Lille, France: Proceedings of Machine Learning Research.

Schulman, J., P. Moritz, S. Levine, M. Jordan and P. Abbeel. 2015. "High-Dimensional Continuous Control Using Generalized Advantage Estimation". *arXiv preprint arXiv:1506.02438*.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford and O. Klimov. 2017. "Proximal Policy Optimization Algorithms". *arXiv preprint arXiv:1707.06347*.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche *et al*. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature* 529(7587):484–489.

Spall, J. C. 1992. "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation". *IEEE Transactions on Automatic Control* 37(3):332–341.

Tamar, A., Y. Glassner, and S. Mannor. 2014. "Policy Gradients Beyond Expectations: Conditional Value-at-Risk". *arXiv preprint arXiv:1404.3862*.

Tversky, A. and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty". *Journal of Risk and Uncertainty* 5:297–323.

Wand, M. P. and M. C. Jones. 1994. *Kernel Smoothing*. Chemical Rubber Company Press.

Wang, S. S. 2000. "A Class of Distortion Operators for Pricing Financial and Insurance Risks". *Journal of Risk and Insurance*:15–36.

Williams, R. J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". *Machine Learning* 8:229–256.

AUTHOR BIOGRAPHIES

JINYANG JIANG is a PhD candidate in the Department of Management Science and Information Systems in Guanghua School of Management at Peking University, Beijing, China. He received the BS degree in information and computational sciences from School of Mathematics and Statistics, Wuhan University, and the double BS degree in computer science and technology from School of Computer Science, Wuhan University, in 2021. His research interests include machine learning and simulation optimization. His email address is jinyang.jiang@stu.pku.edu.cn.

BERND HEIDERGOTT is the professor of Stochastic Optimization at the Department of Operations Analytics at the Vrije Universiteit Amsterdam, the Netherlands. He received his PhD degree from the University of Hamburg, Germany, in 1996, and held postdoc positions at various universities before joining the Vrije Universiteit. Bernd is a research fellow of the Tinbergen Institute and a board member of the Amsterdam Business Research Institute. His research interests are optimization and control of discrete event systems, perturbation analysis, Markov chains, max-plus algebra, and social networks. His email address is b.f.heidergott@vu.nl.

JIAQIAO HU is an associate professor in the Department of Applied Mathematics and Statistics at the State University of New York, Stony Brook. He received the BS degree in automation from Shanghai Jiao Tong University, the MS degree in applied mathematics from the University of Maryland, Baltimore County, and the PhD degree in electrical engineering from the University of Maryland, College Park. His research interests include Markov decision processes, applied probability, and simulation optimization. He is currently Department Editor for IISE Transactions and Associate Editor for Operations Research. His email address is jqhu@ams.stonybrook.edu.

YIJIE PENG is an associate professor in Guanghua School of Management at Peking University. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare. He is a member of INFORMS and IEEE and serves as an Associate Editor of the Asia-Pacific Journal of Operational Research and the Conference Editorial Board of the IEEE Control Systems Society. His email address is pengyijie@pku.edu.cn.