A policy gradient approach for optimization of smooth risk measures

Abstract

We propose policy gradient algorithms for solving a risk-sensitive reinforcement learning problem in on-policy as well as off-policy settings. We consider episodic Markov decision processes, and model the risk using the broad class of smooth risk measures of the cumulative discounted reward. We propose two template policy gradient algorithms that optimize a smooth risk measure in on-policy and off-policy RL settings, respectively. We derive non-asymptotic bounds that quantify the rate of convergence to our proposed algorithms to a stationary point of the smooth risk measure. As special cases, we establish that our algorithms apply to optimization of mean-variance and distortion risk measures, respectively.

1 INTRODUCTION

Risk-sensitive reinforcement learning (RL) has received a lot of attention recently in the literature, and a few representative works are Tamar et al. [2012], Prashanth [2014], Tamar et al. [2015b], Borkar and Jain [2010], Chow et al. [2017], Prashanth and Ghavamzadeh [2016], Borkar [2010], Prashanth et al. [2016], Huang and Haskell [2017]. Mean-variance tradeoff Markowitz [1952], value at risk (VaR), conditional value at risk (CVaR) Rockafellar and Uryasev [2000], spectral risk measure Acerbi [2002], distortion risk measure Denneberg [1990], a risk measure based on cumulative prospect theory Tversky and Kahneman [1992] are some of the popular risk measures considered in the literature.

Policy gradients form a popular solution approach for traditional risk-neutral RL. The idea here is to considered a parameterized set of policies, usually in a continuous space, and perform a random search using stochastic gradient ascent to find a 'good-enough' policy that optimizes a certain performance criterion. Several risk-sensitive RL algorithms employ this approach to find policies that are risk-optimal, see Prashanth and Fu [2022] for a detailed survey of some of the recent developments in this research direction.

In this paper, we consider the problem of optimizing an abstract smooth risk measure (SRM) in a risk-sensitive RL context. SRMs constitute a broad class of risk measures that include mean-variance risk measure (MVRM) and distortion risk measure (DRM). Mean-variance tradeoff is a well-known risk measure that is closely related to exponential cost risk measure – a connection that can be seen using a Taylor series expansion (cf. Prashanth and Ghavamzadeh [2016]). Next, DRM is an expectation w.r.t. a distorted distribution that is arrived at using a distortion function that alters the underlying cumulative distribution function (CDF). Popular risk measures like VaR and CVaR can be seen as special cases of DRM using appropriate distortion functions. However, VaR is not a popular objective for risk-sensitive optimization since it is not coherent, while CVaR, though coherent[1], is not preferable, as it considers all losses beyond VaR equally, while ignoring all the losses below CVaR. A DRM can be written in an equivalent form as spectral risk measure with a risk-aversion function that attaches a weight to each loss value. In this form, it is easy to argue that DRMs are preferable over CVaR, since one can attach larger weight to higher losses, unlike CVaR which attached the same weight to all losses beyond VaR.

We employ the policy gradient approach for solving a risk-sensitive MDP, with an SRM as the objective. The goal in our formulation is to find a policy that maximizes the DRM of the cumulative reward in an episodic Markov decision process (MDP). We propose a template policy gradient to solve this problem for an abstract SRM. The template algorithm has the following crucial components: a risk estimation scheme and a gradient estimation scheme. The risk

[1] A risk measure is said to be coherent if it is translation invariant, sub-additive, positive homogeneous, and monotonic Artzner et al. [1999].

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

estimation scheme for an abstract SRM is assumed to guarantee a $O(\frac{1}{m})$ mean-square error (MSE), where m is the number of episodes. With an expected value objective in a risk-neutral setting, this MSE requirement is natural. For the case of mean-variance tradeoff and DRMs, we manifest such a bound for natural estimators. We would like to add that, unlike expected value where a sample mean was a good estimator, estimating a DRM is more challenging since the episodes are obtained using the CDF of the cumulative reward, while DRM is an expectation with a distorted distribution implying an estimate of the underlying CDF is necessary, or a sample mean is not sufficient for DRM estimation.

For the purpose of gradient estimation, we employ the smoothed functional (SF) approach. This scheme falls under the realm of simultaneous perturbation methods Bhatnagar et al. [2013], which estimate the gradient of a function given noisy observations. Simultaneous perturbation methods in general, and SF methods in particular, are efficient and easy to implement as they require only two function measurements for estimating the gradient, irrespective of the parameter dimension. The choice of the SF scheme for estimating the gradient of an abstract SRM is not arbitrary. For some risk measures, it is not possible to employ the likelihood ratio method to arrive at a policy gradient theorem. This is true for the mean variance risk measure, as shown in Prashanth and Ghavamzadeh [2016]. This is also unlike the classic expected value objective, for which one could use the policy gradient theorem to arrive at a gradient estimation scheme based on the likelihood ratio method.

We now summarize our contributions. First, we propose two template policy gradient algorithms with an SRM as the objective. The first algorithm operates in an on-policy RL setting, while the second caters to the off-policy RL setting. Second, we derive non-asymptotic bounds that quantify the rate of convergence to our proposed algorithms to a stationary point of an SRM. As special cases, we establish that our algorithms and associated theoretical guarantees apply to optimization of mean-variance and distortion risk measures, respectively, in a risk-sensitive RL context. To the best of our knowledge, policy gradient algorithm with non-asymptotic convergence guarantees are not available in the literature for SRMs in general, and for the special cases of mean-variance risk measure and DRMs in particular. Our non-asymptotic bound for the template algorithm can be used as a blackbox to characterize the convergence rate for SRMs beyond mean-variance and DRM. In particular, one can arrive at a $O(1/\epsilon^2)$ bound on the number of iterations for convergence to an ϵ-stationary point of the SRM, provided one verifies the necessary assumptions that guarantee smoothness of SRM and a MSE bound on the SRM estimators.

Related work. In Tamar et al. [2015a], the authors propose a policy gradient algorithm for an abstract coherent risk measure, and derive a policy gradient theorem using the dual representation of a coherent risk measure. Their estimation scheme requires solving a convex optimization problem. Also, they establish asymptotic consistency of their proposed gradient estimate. In contrast, our estimation scheme is computationally inexpensive, and our theoretical guarantees are non-asymptotic in nature. In Prashanth and Fu [2022], the authors survey policy gradient algorithms for optimizing different risk measures in a constrained as well as an unconstrained RL setting. They provide a non-asymptotic bound of $O(1/N^{1/3})$ for an abstract smooth risk measure, assuming a gradient oracle that satisfies certain bias-variance conditions. In contrast, we provide concrete gradient estimation schemes in a risk-sensitive RL setting, and more importantly, we derive an improved non-asymptotic bound of order $O(1/\sqrt{N})$. In Prashanth et al. [2016] the authors consider a CPT-based objective in an RL setting, and they employ simultaneous perturbation stochastic approximation (SPSA) method for the gradient estimation, and provide asymptotic convergence guarantees for their algorithm. The optimization of a DRM is closely related to that of CPT. Under general conditions on the policy parameterization, which are usually employed in the analysis of policy gradient algorithms, we show that DRM is smooth, in turn leading a non-asymptotic bound of $O(1/\sqrt{N})$. This is unlike Prashanth et al. [2016], where the authors provide asymptotic guarantees assuming the policy parameterization ensures that the CPT-value is three times continuously differentiable — a condition that is hard to verify in practice. In a non-RL context, the authors in Glynn et al. [2021] study the sensitivity of DRM using an estimator that is based on the generalized likelihood ratio method, and establish a central limit theorem for their gradient estimator.

2 PRELIMINARIES

We consider an MDP with a state space $\mathcal{S}$ and an action space $\mathcal{A}$. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite spaces. Let $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [-r_{\max}, r_{\max}], r_{\max} \in \mathbb{R}^+$ be the single stage scalar reward, and $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ be the transition probability function. We consider episodic problems, where each episode starts at a fixed state S_0 and terminates at a special zero reward absorbing state 0. The action selection is based on parameterized stochastic policies $\{\pi_\theta : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to [0, 1], \theta \in \mathbb{R}^d\}$. We make the following assumptions on the parameterized policies $\{\pi_\theta, \theta \in \mathbb{R}^d\}$:

(A1). *The policies $\{\pi_\theta, \theta \in \mathbb{R}^d\}$ are proper, i.e.,* $\exists M > 0 : \max_{s \in \mathcal{S}} \mathbb{P}\left(S_M \neq 0 | S_0 = s, \pi_\theta\right) < 1, \forall \theta \in \mathbb{R}^d$.

(A2). $\exists M_d, M_h > 0 : \forall \theta \in \mathbb{R}^d, \forall a \in \mathcal{A}, s \in \mathcal{S}$, $\|\nabla \log \pi_\theta(a \mid s)\| \leq M_d$, *and* $\left\|\nabla^2 \log \pi_\theta(a \mid s)\right\| \leq M_h$, *where $\|\cdot\|$ is the d-dimensional Euclidean norm when the operand is a vector, and the operator norm when the operand is a matrix.*

Assumption (A1) is commonly used in the analysis of episodic MDPs (cf. Bertsekas and Tsitsiklis [1996]). An assumption like (A2) is common for analyzing policy gradient algorithms (cf. Zhang et al. [2020], Papini et al. [2018]).

We denote by S_t and A_t, the state and the action at time $t \in \{0, 1, \cdots\}$ respectively. The cumulative discounted reward R^θ is defined by

$$R^\theta = \sum_{t=0}^{T-1} \gamma^t r(S_t, A_t, S_{t+1}), \forall \theta \in \mathbb{R}^d, \tag{1}$$

where $A_t \sim \pi_\theta(\cdot, S_t)$, $S_{t+1} \sim p(\cdot, S_t, A_t)$, $\gamma \in (0, 1)$ is the discount factor, and T is the random length of an episode. We can see that $\forall \theta \in \mathbb{R}^d, |R^\theta| < \frac{r_{\max}}{1-\gamma} = M_r$ a.s. From (A1), we infer that $\mathbb{E}[T] < \infty$. This fact in conjunction with $T \geq 0$ implies the following bound:

$$\exists M_e > 0 : T \leq M_e \text{ a.s.} \tag{2}$$

On-policy learning is a scheme where a policy parameter θ is optimized using the data collected by the same policy π_θ. In contrast, off-policy learning is a scheme where we optimize θ using data collected by a different behavior policy b.

In an off-policy setting, we collect episodes from b and estimate the values of π_θ, using importance sampling ratios. We require the behavior policy b to be proper, i.e.,

(A3). $\exists M > 0 : \max_{s \in \mathcal{S}} \mathbb{P}(S_M \neq 0 \mid S_0 = s, b) < 1.$

We also assume that the target policy π_θ is absolutely continuous w.r.t. the behavior policy b, i.e.,

(A4). $\forall \theta \in \mathbb{R}^d, b(a|s) = 0 \Rightarrow \pi_\theta(a|s) = 0, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}.$

Assumption (A4) is standard in an off-policy RL setting (cf. Sutton et al. [2009]).

The cumulative discounted reward R^b is defined by

$$R^b = \sum_{t=0}^{T-1} \gamma^t r(S_t, A_t, S_{t+1}), \tag{3}$$

where $A_t \sim b(\cdot, S_t)$, $S_{t+1} \sim p(\cdot, S_t, A_t)$, $\gamma \in (0, 1)$, and T is the random length of an episode. As before, (A3) implies $\mathbb{E}[T] < \infty$, and the following bound:

$$\exists M_e > 0 : T \leq M_e, \text{ a.s.} \tag{4}$$

The importance sampling ratio ψ^θ is defined by

$$\psi^\theta = \prod_{t=0}^{T-1} \frac{\pi_\theta(A_t \mid S_t)}{b(A_t \mid S_t)}. \tag{5}$$

From (A2) and (A4), we obtain $\forall \theta \in \mathbb{R}^d, \pi_\theta(a|s) > 0$ and $b(a|s) > 0, \forall a \in \mathcal{A},$ and $\forall s \in \mathcal{S}$. This fact in conjunction with (4) implies the following bound for ψ^θ:

$$\exists M_s > 0 : \forall \theta \in \mathbb{R}^d, \psi^\theta \leq M_s, \text{ a.s.} \tag{6}$$

Table 1: Examples of distortion functions

Dual-power function	$g(s) = 1 - (1-s)^r, r \geq 2$
Quadratic function	$g(s) = (1+r)s - rs^2, 0 \leq r \leq 1$
Exponential function	$g(s) = \frac{1-\exp(-rs)}{1-\exp(-r)}, r > 0$
Square-root function	$g(s) = \frac{\sqrt{1+rs}-1}{\sqrt{1+r}-1}, r > 0$
Logarithmic function	$g(s) = \frac{\log(1+rs)}{\log(1+r)}, r > 0$

3 SMOOTH RISK MEASURES

The cumulative reward is a random quantity as there is randomness in state transition as modeled by the transition probability function as well as in the action selection in the case of stochastic policies. We consider a smooth risk measure ρ as an objective function, which provides a numerical value that represents certain aspects of this random quantity.

Definition 1. *A risk measure is smooth if it satisfies the following condition:*

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d, \|\nabla \rho(\theta_1) - \nabla \rho(\theta_2)\| \leq L_{\rho'} \|\theta_1 - \theta_2\|. \tag{7}$$

Our goal is to find a policy parameter θ^* that maximizes the objective function ρ, i.e,

$$\theta^* \in \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \rho(\theta). \tag{8}$$

As examples of smooth risk measures, we consider mean-variance and the class of distortion risk measures. We describe these risk measures in the following sections.

3.1 DISTORTION RISK MEASURE

The DRM of a random variable X is the expected value of X under a distortion of the CDF F_X, attained using a given distortion function $g(\cdot)$. A DRM is defined using a Choquet integral as follows:

$$\rho_g(X) = \int_{-\infty}^{0} (g(1 - F_X(x)) - 1) dx + \int_{0}^{\infty} g(1 - F_X(x)) dx.$$

The distortion function $g : [0, 1] \to [0, 1]$ is non-decreasing, with $g(0) = 0$ and $g(1) = 1$. We can see that $\rho_g(X) = \mathbb{E}[X]$, if $g(\cdot)$ is the identity function. A few examples of $g(\cdot)$ are available in Table 1 and their plots are in Figure 1.

The DRM $\rho_g(\theta)$ of the cumulative reward R^θ, defined in (1), is given by

$$\rho_g(\theta) = \int_{-M_r}^{0} (g(1 - F_{R^\theta}(x)) - 1) dx + \int_{0}^{M_r} g(1 - F_{R^\theta}(x)) dx, \tag{9}$$

where F_{R^θ} is the CDF of R^θ.

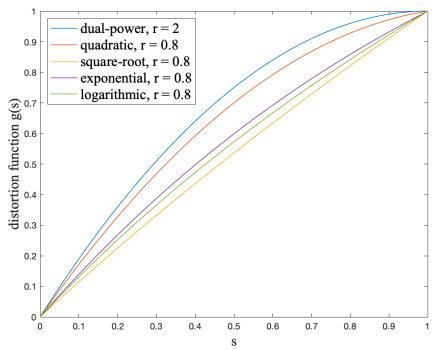

Figure 1: Examples of distortion functions

3.2 MEAN-VARIANCE RISK MEASURE

The MVRM $\rho_\lambda(\theta)$ of R^θ, defined in (1), is given by

$$\rho_\lambda(\theta) = J(\theta) - \lambda V(\theta), \text{ where}$$
$$J(\theta) = \mathbb{E}\left[R^\theta\right]; \; V(\theta) = \mathbb{E}\left[\left(R^\theta\right)^2\right] - J(\theta)^2. \quad (10)$$

In the above, $J(\theta)$ is the value function, which is the objective in a risk-neutral RL setting. Further, $V(\theta)$ is the variance of the cumulative reward, and λ is a scalar that is used to tradeoff between mean and variance. A popular risk measure in control literature is exponential utility, where the objective is $-\frac{1}{\lambda}\log\mathbb{E}[e^{-\lambda R^\theta}]$, with R^θ denoting the cumulative reward. Using a first-order Taylor expansion, it is apparent that

$$-\frac{1}{\lambda}\log\mathbb{E}[e^{-\lambda R^\theta}] = \mathbb{E}[R^\theta] - \frac{\lambda}{2}\text{Var}[R^\theta] + O(\lambda^2).$$

Thus, the MVRM risk measure defined above can be seen as an approximation to the exponential utility risk measure. Optimizing the latter risk measure in an RL context is challenging, and to the best of our knowledge, there is no RL algorithm with compact parameterization for this problem, see Section 7.1 of Prashanth and Fu [2022] for the details. In contrast, we present a policy gradient algorithm for MVRM with a provable bound on the rate for stationary convergence.

4 POLICY GRADIENT TEMPLATE FOR SMOOTH RISK MEASURES

We propose two policy gradient algorithms for optimizing a smooth risk measure. The first algorithm operates in an on-policy RL setting, and Algorithm 1 presents the pseudocode. The second algorithm caters to an off-policy RL setting, with a pseudocode that follows the template in Algorithm 1 with variations in estimation. There are two crucial ingredients in each of these policy gradient algorithms:

1. Risk estimation: This refers to the problem of estimating the value of a smooth risk measure for a given policy parameter, say θ. In an on-policy setting, the estimation scheme has access to a mini-batch of episodes from the policy θ itself. On the other hand, in an off-policy setting, the estimation scheme has to use the episodes simulated using a behavior policy.
2. Gradient estimation: This refers to the estimation of the policy gradient $\nabla\rho(\theta)$ for a given parameter θ. Such an estimate would be used to perform stochastic gradient ascent in the policy parameter.

The estimation scheme is specific to the risk measure considered. For the theoretical guarantees in the next section, we require the following bound on the estimate $\hat{\rho}_m(\theta)$ of the risk $\rho(\theta)$, given m episodes:

$$\mathbb{E}\left[\left|\hat{\rho}(\theta) - \rho(\theta)\right|^2\right] \leq \frac{C_1}{m}. \quad (11)$$

The condition above relates to the mean-square error of the risk estimator, and the rate of $O(1/m)$ is natural, considering such a bound is reasonable even for the case of an expected value objective. For the two applications with mean-variance and distortion risk measures, we shall establish later that the estimators of these risk measures satisfy the condition specified above.

For handling the problem of gradient estimation, both proposed algorithms use an SF-based estimation scheme. The choice of this gradient estimation scheme is not arbitrary. For some risk measures, it is not possible to employ the likelihood ratio method to arrive at a policy gradient theorem. This is true for the mean-variance risk measure, as shown in Prashanth and Ghavamzadeh [2016], since the policy gradient expression for the square value $\mathbb{E}\left[(R^\theta)^2\right]$ has the factor $\nabla J(\theta, x')$, in turn requiring an estimate of the gradient of the value function for every possible start state. In the aforementioned reference, the authors employed SPSA, a popular simultaneous perturbation method to workaround the policy gradient expression. In our work, we use SF, which also falls under the realm of simultaneous perturbation methods for gradient estimation. Moreover, unlike Prashanth and Ghavamzadeh [2016], we consider a broad class of smooth risk measures, and more importantly, we establish non-asymptotic bounds that quantify the rate of convergence of our proposed SF-based policy gradient algorithms.

The SF-based gradient estimation is a zeroth-order gradient estimation scheme, where the gradient is estimated from perturbed function values (cf. Nesterov and Spokoiny [2017], Bhatnagar et al. [2013], Shamir [2017]). The SF-based methods form a smoothed version of the objective function $\rho(\cdot)$ as $\rho_\mu(\cdot)$ and use the gradient $\nabla\rho_\mu$ as an approximation for $\nabla\rho$. The smoothed functional $\rho_\mu(\theta)$ is defined as

$$\rho_\mu(\theta) = \mathbb{E}_{u\in\mathbb{B}^d}\left[\rho(\theta + \mu u)\right], \quad (12)$$

Algorithm 1 OnP-SF

1: **Input**: Parameterized form of the policy π, iteration limit N, step-size α, perturbation parameter μ, and batch sizes m and n;
2: **Initialize**: Target policy parameter $\theta_0 \in \mathbb{R}^d$, and the discount factor $\gamma \in (0,1)$;
3: **for** $k = 0, \ldots, N-1$ **do**
4: **for** $i = 1, \ldots, n$ **do**
5: Get $[v_i^1, \ldots, v_i^d] \in \mathbb{S}^{d-1}$;
6: Generate m episodes each using $\pi_{(\theta_k \pm \mu v_i)}$;
7: Estimate $\hat{\rho}(\theta_k \pm \mu v_i)$;
8: **end for**
9: Use (15) to estimate $\widehat{\nabla}_{\mu,n}\hat{\rho}(\theta_k)$;
10: Use (16) to calculate θ_{k+1};
11: **end for**
12: **Output**: Policy θ_R, where $R \sim \mathcal{U}\{0, N-1\}$

where u is sampled uniformly at random from the unit ball $\mathbb{B}^d = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$, and $\mu \in (0,1]$ is the smoothing parameter. From [Flaxman et al., 2005, Lemma 2.1], we obtain the gradient of $\rho_\mu(\theta)$ as follows:

$$\nabla \rho_\mu(\theta) = \mathbb{E}_{v \in \mathbb{S}^{d-1}}\left[\frac{d}{\mu}\rho(\theta + \mu v)v\right], \tag{13}$$

where v is sampled uniformly at random from the unit sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$. The gradient $\nabla\rho_\mu(\theta)$ is estimated as follows:

$$\widehat{\nabla}_{\mu,n}\rho(\theta) = \frac{d}{n}\sum_{i=1}^{n}\frac{\rho(\theta + \mu v_i) - \rho(\theta - \mu v_i)}{2\mu}v_i, \tag{14}$$

where $\forall i, v_i$ is sampled uniformly at random from $\mathbb{S}^{d-1}$. The gradient estimate is averaged over n unit vectors to reduce the variance. Using the proof technique from Vijayan and Prashanth [2021a], we show that $\widehat{\nabla}_{\mu,n}\rho(\theta)$ is an unbiased estimator of $\nabla\rho_\mu(\theta)$, see Appendix A for the details. In a typical RL setting, we may not have direct measurements of $\rho(\cdot)$, which need to be estimated using sample episodes. Let $\hat{\rho}(\cdot)$ be the estimator for $\rho(\cdot)$, then we use a gradient estimator as given below:

$$\widehat{\nabla}_{\mu,n}\hat{\rho}(\theta) = \frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}(\theta + \mu v_i) - \hat{\rho}(\theta - \mu v_i)}{2\mu}v_i. \tag{15}$$

We solve (8) using the following update iteration:

$$\theta_{k+1} = \theta_k + \alpha\widehat{\nabla}_{\mu,n}\hat{\rho}(\theta_k), \tag{16}$$

where θ_0 is set arbitrarily, and α is the step-size.

We consider two algorithms, both armed with a risk estimator $\hat{\rho}(\cdot)$ and a risk gradient estimate using SF. In our first algorithm OnP-SF, $\hat{\rho}(\cdot)$ uses an on-policy evaluation. Algorithm 1 presents the pseudocode of OnP-SF.

Each iteration of OnP-SF requires $2mn$ episodes corresponding to $2n$ perturbed policies. In some practical applications, it may not be feasible to generate system trajectories corresponding to different perturbed policies. In our second algorithm OffP-SF, we overcome the aforementioned problem by performing the off-policy evaluation. Using the off-policy setting, the number of episodes needed in each iteration of our algorithm can be reduced to m. The pseudocode of OffP-SF is similar to Algorithm 1 with the following deviations: The estimate $\hat{\rho}(\theta_k \pm \mu v_i)$ in step 7 is performed in a off-policy fashion, and for this purpose m episodes are generated only once using the behavior policy. In contrast, step 6 in Algorithm 1 requires simulation of m episodes in each iteration using the current policy parameter $\pi_{\theta_k \pm \mu v_i}$.

5 APPLICATION: DRM

Recall that the optimization problem in (8) is solved using stochastic gradient ascent, and for each update iteration, we require estimates of $\rho_g(\cdot)$. In the following sections, we describe our algorithms that estimate DRM in on-policy and off-policy RL settings, respectively.

5.1 ON-POLICY DRM ESTIMATION

We generate m episodes using the policy π_θ, and estimate the CDF $F_{R^\theta}(\cdot)$ using sample averages. We denote by R_i^θ the cumulative reward of the episode i. We form the estimate $G_{R^\theta}^m(\cdot)$ of $F_{R^\theta}(\cdot)$ as follows:

$$G_{R^\theta}^m(x) = \frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbb{1}\{R_i^\theta \leq x\}. \tag{17}$$

Now, we form an estimate $\hat{\rho}_g^G(\theta)$ of $\rho_g(\theta)$ as follows:

$$\hat{\rho}_g^G(\theta) = \int_{-M_r}^{0}(g(1 - G_{R^\theta}^m(x)) - 1)dx + \int_0^{M_r} g(1 - G_{R^\theta}^m(x))dx. \tag{18}$$

Comparing (18) with (9), it is apparent that we have used the EDF $G_{R^\theta}^m$ in place of the true cdf F_{R^θ}.

We simplify (18) in terms of order statistics as follows:

$$\hat{\rho}_g^G(\theta) = \sum_{i=1}^{m} R_{(i)}^\theta\left(g\left(1 - \frac{i-1}{m}\right) - g\left(1 - \frac{i}{m}\right)\right), \tag{19}$$

where $R_{(i)}^\theta$ is the ith smallest order statistic of the samples $\{R_1^\theta, \cdots R_m^\theta\}$. The reader is referred to Lemma 13 in Appendix B for proof. If we choose the distortion function as the identity function, then the estimator in (19) is merely the sample mean.

We make the following assumptions to ensure the Lipschitzness, and smoothness of the DRM ρ_g.

(A5). $\exists M_{g'}, M_{g''} > 0 : \forall t \in (0,1)$, $|g'(t)| \leq M_{g'}$, *and* $|g''(t)| \leq M_{g''}$.

The assumption (A5) helps us establish that the distortion functions and its derivative are Lipschitz continuous. A few examples of distortion functions, which satisfy (A5) are given in Table 1.

A critical requirement for establishing convergence guarantee is a bound on the MSE of the risk estimation scheme – see the discussion around (11). The result below shows that this MSE requirement is met by the DRM estimator (18).

Lemma 1. *Assume (A1)-(A2) and (A5). Then,*

$$\mathbb{E}\left[\left|\rho_g(\theta) - \hat{\rho}_g^G(\theta)\right|^2\right] \leq \frac{16 M_r^2 M_{g'}^2}{m}.$$

Proof. See Appendix B. □

5.2 OFF-POLICY DRM ESTIMATION

We generate m episodes using the policy b to estimate the CDF $F_{R^\theta}(\cdot)$ using importance sampling. We denote by R_i^b the cumulative reward, and ψ_i^θ the importance sampling ratio of the episode i. We form the estimate $H_{R^\theta}^m(\cdot)$ of $F_{R^\theta}(\cdot)$ as follows:

$$H_{R^\theta}^m(x) = \min\{\hat{H}_{R^\theta}^m(x), 1\}, \text{ where} \tag{20}$$

$$\hat{H}_{R^\theta}^m(x) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{R_i^b \leq x\}\psi_i^\theta. \tag{21}$$

In the above, $\hat{H}_{R^\theta}^m(x)$ is an empirical estimate of $F_{R^\theta}(x)$ as $F_{R^\theta}(x) = \mathbb{E}\left[\mathbb{1}\{R^b \leq x\}\psi^\theta\right]$. Because of the importance sampling ratio, $\hat{H}_{R^\theta}^m(x)$ can get a value above 1. Since we are estimating a CDF, we restrict $\hat{H}_{R^\theta}^m(x)$ to $H_{R^\theta}^m(x)$.

Now we form an estimate $\hat{\rho}_g^H(\theta)$ of $\rho_g(\theta)$ as

$$\hat{\rho}_g^H(\theta) = \int_{-M_r}^0 (g(1-H_{R^\theta}^m(x))-1)dx + \int_0^{M_r} g(1-H_{R^\theta}^m(x))dx. \tag{22}$$

We can simplify (22) in terms of order statistics as

$$\hat{\rho}_g^H(\theta) = R_{(1)}^b + \sum_{i=2}^m R_{(i)}^b g\left(1-\min\left\{1, \frac{1}{m}\sum_{k=1}^{i-1}\psi_{(k)}^\theta\right\}\right) - \sum_{i=1}^{m-1} R_{(i)}^b g\left(1-\min\left\{1, \frac{1}{m}\sum_{k=1}^{i}\psi_{(k)}^\theta\right\}\right), \tag{23}$$

where $R_{(i)}^b$ is the ith smallest order statistic of the samples $\{R_1^b, \cdots R_m^b\}$, and $\psi_{(i)}^\theta$ is the importance sampling ratio of $R_{(i)}^b$. The reader is referred to Lemma 14 in Appendix B for proof.

A result in the spirit of Lemma 1 for the off-policy setting is given below.

Lemma 2. *Assume (A1)-(A5). Then,*

$$\mathbb{E}\left[\left|\rho_g(\theta) - \hat{\rho}_g^H(\theta)\right|^2\right] \leq \frac{16 M_r^2 M_{g'}^2 M_s^2}{m}.$$

Proof. See Appendix B. □

6 APPLICATION: MVRM

As in the case of DRMs, we describe the estimation of the MVRM in on-policy and off-policy settings, respectively.

6.1 ON-POLICY MVRM ESTIMATION

We generate m episodes using the policy π_θ, and estimate $J(\theta)$ and $V(\theta)$ using sample averages. We denote by R_i^θ the cumulative reward of the episode i. The estimators $\hat{J}_m^\pi$ of $J(\theta)$ and $\widehat{V}_m^\pi$ of $V(\theta)$ is defined as follows:

$$\hat{J}_m^\pi = \frac{1}{m}\sum_{i=1}^m R_i^\theta; \tag{24}$$

$$\widehat{V}_m^\pi = \frac{1}{m-1}\sum_{i=1}^m \left(R_i^\theta - \hat{J}_m^\pi\right)^2. \tag{25}$$

Using Theorem 2-3 in [Mood et al., 1974, chapter V1], we can see that the above estimates are unbiased.

Now using (24) and (25), we estimate $\rho_\lambda(\theta)$ as follows:

$$\hat{\rho}_\lambda^\pi(\theta) = \hat{J}_m^\pi(\theta) - \lambda \widehat{V}_m^\pi(\theta). \tag{26}$$

As in the case of DRM, we show in the result below that the mean-variance estimator (26) satisfies an MSE bound of order $O(1/m)$.

Lemma 3. *Assume (A1)-(A2) and let $m > 2$. Then*

$$\mathbb{E}\left[\left|\hat{\rho}_\lambda^\pi(\theta) - \rho_\lambda(\theta)\right|^2\right] \leq \frac{8M_r^2 + 32\lambda^2 M_r^4}{m}.$$

Proof. See Appendix C. □

6.2 OFF-POLICY MVRM ESTIMATION

We generate m episodes using the policy b to estimate $J(\theta)$ using importance sampling. We denote by R_i^b the cumulative reward, and ψ_i^θ the importance sampling ratio of the episode i. Since $J(\theta) = \mathbb{E}\left[R^b\psi^\theta\right]$, we estimate it using sample average as follows:

$$\hat{J}_m^b = \frac{1}{m}\sum_{i=1}^m R_i^b\psi_i^\theta; \tag{27}$$

$$\widehat{V}_m^b = \frac{1}{m-1}\sum_{i=1}^m \left(R_i^b\psi_i^\theta - \hat{J}_m^b\right)^2. \tag{28}$$

As in the on-policy setting, these estimates are unbiased.

Now using (27) and (28), we estimate $\rho_\lambda(\theta)$ as follows:

$$\hat{\rho}^b_\lambda(\theta) = \hat{J}^b_m(\theta) - \lambda \widehat{V}^b_m(\theta). \tag{29}$$

A result in the spirit of Lemma 3 for the off-policy setting is given below.

Lemma 4. *Assume (A1)-(A4) and let $m > 2$. Then*

$$\mathbb{E}\left[\left|\hat{\rho}^b_\lambda(\theta) - \rho_\lambda(\theta)\right|^2\right] \leq \frac{8M_r^2 M_s^2 + 32\lambda^2 M_r^4 M_s^4}{m}.$$

Proof. See Appendix C. □

7 MAIN RESULTS

Our non-asymptotic analysis establishes a bound on the number of iterations of our proposed algorithms to find an ϵ-stationary point of the smooth risk measure, which is defined below.

Definition 2 (ϵ-stationary point)**.** *Let θ_R be the output of an algorithm. Then, θ_R is called an ϵ-stationary point of problem (8), if $\mathbb{E}\left[\|\nabla \rho(\theta_R)\|^2\right] \leq \epsilon$.*

For a non-convex objective function, it is common in optimization literature to establish a convergence rate result to an ϵ-stationary point. Such a convergence notion is used in the analysis of policy gradient algorithms as well, cf. Papini et al. [2018], Shen et al. [2019], Zhang et al. [2020].

7.1 BOUNDS FOR ONP-SF/OFFP-SF

We make the following assumptions.

(A6). $\forall \theta \in \mathbb{R}^d$, *$\hat{\rho}(\theta)$ and $\rho(\theta)$ are bounded.*

(A7). $\forall \theta \in \mathbb{R}^d$, $\mathbb{E}\left[|\hat{\rho}(\theta) - \rho(\theta)|^2\right] \leq \frac{C_1}{m}$.

(A8). $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $|\rho(\theta_1) - \rho(\theta_2)| \leq L_\rho \|\theta_1 - \theta_2\|$.

(A9). $\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla\rho(\theta_1) - \nabla\rho(\theta_2)\| \leq L_{\rho'} \|\theta_1 - \theta_2\|$.

We present bounds for an iterate θ_R that is chosen uniformly at random from $\{\theta_0, \cdots, \theta_{N-1}\}$. The bound that we present below applies to the template algorithm for on-policy as well as off-policy RL settings, and the proof is available in Appendix A. Moreover, the bound below is for a general step-size, smoothing parameter and batch size parameters. Subsequently, we specialize this result to arrive at a $O(N^{-1/2})$ bound on $\mathbb{E}\left[\|\nabla \rho(\theta_R)\|^2\right]$.

Proposition 1. *(OnP-SF/OffP-SF) Assume (A6)-(A9). Let $\{\theta_i, i = 0, \cdots, N-1\}$ be the policy parameters generated by OnP-SF/OffP-SF, and let θ_R be chosen uniformly at random from this set. Let $\rho^* = \max_{\theta \in \mathbb{R}^d} \rho(\theta)$. Then*

$$\mathbb{E}\left[\|\nabla\rho(\theta_R)\|^2\right] \leq \frac{2\left(\rho^* - \rho(\theta_0)\right)}{N\alpha} + \mu^2 d^2 L_{\rho'}^2 + L_{\rho'}\alpha\left(\frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}\right) + \frac{4d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}, \tag{30}$$

where L_ρ, L'_ρ, and C_1 are as in (A6)-(A9).

A straightforward specialization of the bound in (30) with specific choices for the step-size α, smoothing parameter μ, and batch sizes m and n leads to following bounds for OnP-SF and OffP-SF algorithms, respectively.

Theorem 1. *(OnP-SF) Set $\alpha = \frac{1}{\sqrt{N}}$, $\mu = \frac{1}{\sqrt[4]{N}}$, $n = \sqrt{N}$, and $m = \sqrt{N}$. Then, under the conditions of Proposition 1, we have*

$$\mathbb{E}\left[\|\nabla\rho(\theta_R)\|^2\right] \leq \frac{2\left(\rho^* - \rho(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{4d^2L_\rho^2 + d^2C_1L_{\rho'}}{N} + \frac{d^2C_1 + d^2L_{\rho'}^2}{\sqrt{N}}.$$

Theorem 2. *(OffP-SF) Set $\alpha = \frac{1}{\sqrt{N}}$, $\mu = \frac{1}{\sqrt[4]{N}}$, $n = N$, and $m = C_2 > 0$. Then, under the conditions of Proposition 1, we have*

$$\mathbb{E}\left[\|\nabla\rho(\theta_R)\|^2\right] \leq \frac{2\left(\rho^* - \rho(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{C_2 4d^2L_\rho^2 + d^2C_1L_{\rho'}}{C_2 N} + \frac{d^2C_1 + C_2 d^2L_{\rho'}^2}{C_2\sqrt{N}}.$$

Remark 1. *The results above show that after N iterations of (16), OnP-SF/OffP-SF return an iterate that satisfies $\mathbb{E}\left[\|\nabla\rho(\theta_R)\|^2\right] = O\left(\frac{1}{\sqrt{N}}\right)$. To put it differently, to find an ϵ-stationary point of the smooth risk measure objective, an order $O(\frac{1}{\epsilon^2})$ iterations of OnP-SF/OffP-SF are enough.*

Remark 2. *The bounds obtained for OnP-SF and OffP-SF are $O(N^{-1/2})$, but with different choices for the parameters α, μ, n, m. One could vary the parameters n, m for OnP-SF such that their product remains N, and still obtain the $O(N^{-1/2})$ bound. The implication is that OnP-SF requires $\Theta(N)$ episodes to achieve the aforementioned rate. On the other hand, one can choose a constant batch size in OffP-SF and increase the parameter n to be $\Theta(N)$ to arrive at a overall convergence rate of $O(N^{-1/2})$. In the off-policy setting, with a fixed dataset, one could increase the parameter n for higher averaging in the gradient estimate, and such a scheme would not entail simulation of additional episodes.*

Remark 3. *Typical results in risk-sensitive RL literature are produces asymptotic in nature. An exception is a result from Prashanth and Fu [2022], where non-asymptotic bounds of $O\left(N^{-1/3}\right)$ are presented. In contrast, we derive $O\left(N^{-1/2}\right)$ bounds for SRMs.*

7.2 CASE 1: DRM

First we show that the assumptions (A8)-(A9) are satisfied for the DRM using the results from the following lemmas (see Appendix B for the proofs).

Lemma 5. *$\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $|\rho_g(\theta_1) - \rho_g(\theta_2)| \leq L_\rho \|\theta_1 - \theta_2\|$, where $L_\rho = 2M_r M_{g'} M_e M_d$.*
$\|\nabla \rho_g(\theta_1) - \nabla \rho_g(\theta_2)\| \leq L_{\rho'} \|\theta_1 - \theta_2\|$, where $L_{\rho'} = 2M_r M_e \left(M_h M_{g'} + M_e M_d^2 (M_{g'} + M_{g''})\right)$.

The main result that establishes a non-asymptotic bound for Algorithm 1 with DRM as the risk measure is given below.

Corollary 1. *(DRM-OnP-SF) Assume (A1)-(A2) and (A5). Then, under the conditions of Theorem 1, we have*

$$\mathbb{E}\left[\|\nabla \rho_g(\theta_R)\|^2\right] \leq \frac{2\left(\rho_g^* - \rho_g(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2 L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{4d^2 L_\rho^2 + d^2 C_1 L_{\rho'}}{N} + \frac{d^2 C_1 + d^2 L_{\rho'}^2}{\sqrt{N}}.$$

In the above, $\rho_g^* = \max_{\theta \in \mathbb{R}^d} \rho_g(\theta)$. The constants C_1, L_ρ, and $L_{\rho'}$ are as in Lemmas 1 and 5, respectively.

Proof. Lemma 1 implies (A7) holds for DRM estimator. Lemma 5 implies the conditions in (A8) and (A9) hold for DRM. The conditions in (A6) can be easily verified for DRM (see Appendix B for details). The main claim now follows by an application of Theorem 1. □

For the off-policy case, a non-asymptotic bound can be inferred from Theorem 2 in a similar fashion as the on-policy case, with Lemma 2 in place of Lemma 1.

Corollary 2. *(DRM-OffP-SF) Assume (A1)-(A5). Then, under the conditions of Theorem 2, we have*

$$\mathbb{E}\left[\|\nabla \rho_g(\theta_R)\|^2\right] \leq \frac{2\left(\rho_g^* - \rho_g(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2 L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{C_2 4d^2 L_\rho^2 + d^2 C_1 L_{\rho'}}{C_2 N} + \frac{d^2 C_1 + C_2 d^2 L_{\rho'}^2}{C_2\sqrt{N}}.$$

In the above, ρ_g^*, L_ρ, and $L_{\rho'}$ are as in Corollary 1. The constant C_1 is as in Lemma 2.

7.3 CASE 2: MVRM

We specialize the result in Proposition 1 to MVRM. Though MVRM is previously analyzed in Tamar et al. [2012], Prashanth and Ghavamzadeh [2013], they only provide asymptotic convergence results. In the following lemma, we show that the assumptions (A8)-(A9) are satisfied for the MVRM. The proof is available in Appendix C.

Lemma 6. *$\forall \theta_1, \theta_2 \in \mathbb{R}^d$, $|\rho_\lambda(\theta_1) - \rho_\lambda(\theta_1)| \leq L_\rho \|\theta_1 - \theta_2\|$, $L_\rho = M_r M_e M_d + 3\lambda M_r^2 M_e M_d$, and $\|\nabla \rho_\lambda(\theta_1) - \nabla \rho_\lambda(\theta_1)\| \leq L_{\rho'} \|\theta_1 - \theta_2\|$, where $L_{\rho'} = M_r M_e \left(M_h + M_e M_d^2\right) + \lambda M_r^2 M_e \left(3M_h + 5M_e M_d^2\right)$.*

Corollary 3. *(MVRM-OnP-SF) Assume (A1)-(A2). Let the batch size $m = \sqrt{N} > 2$. Then, under the conditions of Theorem 1, we have*

$$\mathbb{E}\left[\|\nabla \rho_\lambda(\theta_R)\|^2\right] \leq \frac{2\left(\rho_\lambda^* - \rho_\lambda(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2 L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{4d^2 L_\rho^2 + d^2 C_1 L_{\rho'}}{N} + \frac{d^2 C_1 + d^2 L_{\rho'}^2}{\sqrt{N}}.$$

In the above, $\rho_\lambda^* = \max_{\theta \in \mathbb{R}^d} \rho_\lambda(\theta)$. The constants C_1, L_ρ, and $L_{\rho'}$ are as in Lemmas 3 and 6, respectively.

Proof. Lemma 3 implies (A7) holds for DRM estimator. Lemma 6 implies the conditions in (A8) and (A9) hold for MVRM. The conditions in (A6) can be easily verified for MVRM (see Appendix C for details). The main claim now follows by an application of Theorem 1. □

The bound for the off-policy variant follows by using Lemma 4 in place of Lemma 3.

Corollary 4. *(MVRM-OffP-SF) Assume (A1)-(A4). Let $C_2 > 2$. Then, under the conditions of Theorem 2, we have*

$$\mathbb{E}\left[\|\nabla \rho_\lambda(\theta_R)\|^2\right] \leq \frac{2\left(\rho_\lambda^* - \rho_\lambda(\theta_0)\right)}{\sqrt{N}} + \frac{2d^2 L_\rho^2 L_{\rho'}}{N\sqrt{N}} + \frac{C_2 4d^2 L_\rho^2 + d^2 C_1 L_{\rho'}}{C_2 N} + \frac{d^2 C_1 + C_2 d^2 L_{\rho'}^2}{C_2\sqrt{N}}.$$

In the above, ρ_λ^*, L_ρ, and $L_{\rho'}$ are as in Corollary 3. The constant C_1 is as in Lemma 4.

8 CONCLUSIONS

We proposed two policy gradient algorithms that cater to the broad class of smooth risk measures. Both algorithms employed an SF-based gradient estimation scheme, and were shown to work in on-policy as well as off-policy RL settings. We derived non-asymptotic bounds that quantify the rate of convergence to our proposed algorithms to a stationary point of the smooth risk measure. As special cases, we showed that our theory and algorithms apply to optimization of mean-variance and distortion risk measures, respectively. To the best of our knowledge, policy gradient algorithm with non-asymptotic convergence guarantees are not available in the literature for smooth risk measures in general, and the special cases in particular.

References

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

S. Bhatnagar, H. Prasad, and L. A. Prashanth. Stochastic recursive algorithms for optimization. simultaneous perturbation methods. *Lecture Notes in Control and Inform. Sci.*, 434, 2013.

V. S. Borkar. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

V. S. Borkar and R. Jain. Risk-constrained Markov decision processes. In *IEEE Conference on Decision and Control*, pages 2664–2669, 2010.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *J. Mach. Learn. Res.*, 18(1):6070–6120, 2017.

D. Denneberg. Distorted probabilities and insurance premiums. *Methods of Operations Research*, 63(3):3–5, 1990.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005.

X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: An iteration complexity perspective. *J. Sci. Comput.*, 76(1):327–363, 2018.

P. Glynn, Y. Peng, M. Fu, and J. Hu. Computing sensitivities for distortion risk measures. *INFORMS J. Comp.*, 2021.

W. Huang and W. B. Haskell. Risk-aware Q-learning for Markov decision processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4928–4933. IEEE, 2017.

J. Kim. Bias correction for estimated distortion risk measure using the bootstrap. *Insur.: Math. Econ.*, 47:198–205, 2010.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

A. M. Mood, , F. A. Graybill, and D. C. Boes. Introduction to the theory of statistics. 1974.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comut. Math.*, 17: 527– 566, 2017.

Y. E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. 2004.

M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. Stochastic variance-reduced policy gradient. In *ICML*, 2018.

L. A. Prashanth. Policy gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory (ALT)*, pages 155–169, 2014.

L. A. Prashanth and M. Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning*, 105(3):367–417, Dec 2016.

L.A. Prashanth and M. Fu. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends in Machine Learning*, 15(5):537–693, 2022.

L.A. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In *Adv. Neural Inf. Process. Syst.*, volume 26, 2013.

L.A. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvari. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *ICML*, volume 48, pages 1406–1415, 2016.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, 2017.

Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In *ICML*, pages 5729–5738, 2019.

R. S. Sutton, H. Maei, and C. Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Adv. Neural Inf. Process. Syst.*, volume 21, 2009.

A. Tamar, D. D. Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pages 387–396, 2012.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Adv. Neural Inf. Process. Syst.*, 2015a.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, volume 28, pages 1468–1476, 2015b.

A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.*, 5, 1992.

N. Vijayan and L.A. Prashanth. Smoothed functional-based gradient algorithms for off-policy reinforcement learning: A non-asymptotic viewpoint. *Systems & Control Letters*, 155:104988, 2021a. ISSN 0167-6911.

N. Vijayan and L.A. Prashanth. Policy gradient methods for distortion risk measures. *arXiv preprint arXiv:2107.04422*, 2021b.

K. Zhang, A. Koppel, H. Zhu, and T. Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control. Optim.*, 58(6):3586–3612, 2020.