Policy Newton Methods for Distortion Riskmetrics

Soumen Pachal[1*], Mizhaan Prajit Maniyar [2*], Prashanth L. A.[3]

[1]Indian Institute of Technology Madras, and TCS Research,
[2]Google Deepmind,
[3]Indian Institute of Technology Madras
cs22d009@smail.iitm.ac.in, mizhaan@google.com, prashla@cse.iitm.ac.in

Abstract

We consider the problem of risk-sensitive control in a reinforcement learning (RL) framework. In particular, we aim to find a risk-optimal policy by maximizing the distortion riskmetric (DRM) of the discounted reward in a finite horizon Markov decision process (MDP). DRMs are a rich class of risk measures that include several well-known risk measures as special cases. We derive a policy Hessian theorem for the DRM objective using the likelihood ratio method. Using this result, we propose a natural DRM Hessian estimator from sample trajectories of the underlying MDP. Next, we present a cubic-regularized policy Newton algorithm for solving this problem in an on-policy RL setting using estimates of the DRM gradient and Hessian. Our proposed algorithm is shown to converge to an ϵ-second-order stationary point (ϵ-SOSP) of the DRM objective, and this guarantee ensures the escaping of saddle points. The sample complexity of our algorithms to find an ϵ-SOSP is $\mathcal{O}(\epsilon^{-3.5})$. Our simulation experiments on three popular RL benchmarks validate the theoretical findings. To the best of our knowledge, our is the first work to present convergence to an ϵ-SOSP of a risk-sensitive objective, while existing works in the literature have either shown convergence to a first-order stationary point of a risk-sensitive objective, or a SOSP of a risk-neutral one.

1 Introduction

Reinforcement learning (RL) has achieved tremendous success in several applications, e.g., finance, transportation, insurance, and supply chain management to name a few. In a traditional RL setting, the aim is to learn an optimal policy that maximizes the expected sum of discounted rewards. However, the expected value may not be a good objective in practical applications. To illustrate, consider a portfolio optimization application, where the goal is to find an optimal way to rebalance the portfolio that is spread over assets with varying risks (Cover 1991). In this application, an appealing strategy is to invest in assets with high risk but high return. The portfolio could involve safe stocks with low growth (and low risk), but (Cover 1991) showed that volatile stocks lead to great gains.

In this paper, we focus on distortion riskmetrics (DRMs) (Wang, Wei, and Willmot 2020) — a general class that covers several well-known risk measures as special cases, e.g., value at risk (VaR) (Jorion 1996), conditional value at risk (CVaR)(Rockafellar, Uryasev et al. 2000), Gini deviation (Gini 1912), Gini shortfall (Furman, Wang, and Zitikis 2017), rank-dependent expected utility (Quiggin 2012). DRMs distort the original distribution by using a distortion function, say $h : [0, 1] \to [0, 1]$, with $h(0) = 0$, and then calculate a distorted expected value. More importantly, DRMs include distortion risk measures (Wang 1996; Denneberg 1990), which additionally assume $h(1) = 1$ and also that h is monotone. DRMs also include deviation measures such as Gini deviation, Gini shortfall, Wang's right-tail, left-tail, and two-sided deviations (Jones and Zitikis 2003). DRMs are also equivalent to spectral risk measures (Acerbi 2002), see (Gzyl and Mayoral 2008). DRMs represent a rich class of risk measures, as shown recently by a characterization result involving a combination of DRM-type risk measures in (Fröhlich and Williamson 2024). The reader is referred to Figure 1 for several examples of distortion functions.

We consider a risk-sensitive RL problem with DRM as the objective in a finite horizon Markov decision process (MDP). The aim is to learn an optimal policy by maximizing the DRM of the cumulative discounted reward. We adopt the policy gradient solution approach for this problem. Risk-sensitive RL via policy gradients has received a lot of research attention recently, see (Prashanth and Fu 2022) for a recent survey. Previous works have explored specific DRMs such as Gini deviation (Luo et al. 2023), inter-expected shortfall range (Han, Wang, and Zhou 2022) and an abstract distortion risk measure (Vijayan and Prashanth 2021). However, distortion riskmetrics in their generality have not been considered earlier. The closest related previous work in (Vijayan and Prashanth 2021) has established convergence of a policy gradient algorithm to a first-order stationary point, which include spurious saddle points that do not result in a maxima for DRM. Moreover, there is no prior work that shows the practical utility of policy gradient type algorithms with a DRM objective that exhibits improved behavior from a risk-sensitivity viewpoint. Our work aims to fill the research gaps in theory as well as practice.

Our contributions. We summarize our contributions below, whereas Table 1 compares sample complexities of our algorithm to closely related works in the literature on risk-

*These authors contributed equally.

Algorithm	Objective	Sample Complexity	ϵ-FOSP	ϵ-SOSP
REINFORCE (Williams 1992)	Expected value	$\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$	✓	✗
HAPG(Shen et al. 2019)	Expected value	$\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$	✓	✗
(Yang, Zheng, and Pan 2021)	Expected value	$\mathcal{O}\left(\frac{1}{\epsilon^{4.5}}\right)$	✓	✓
CR-PN (Maniyar et al. 2024)	Expected value	$\mathcal{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	✓	✓
DRM-OnP-LR(Vijayan and Prashanth 2021)	DRM	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	✓	✗
Our work (CRPN-DRM)	DRM	$\mathcal{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	✓	✓

Table 1: Comparison of sample complexities for finding either a ϵ-first-order stationary point (ϵ-FOSP) or ϵ-second-order stationary point (ϵ-SOSP), see Definition 2.1. A ✓ indicates that the algorithm converges to the first or second-order stationary point, whereas ✗ implies the algorithm does not converge to the corresponding stationary point. The first four rows are risk-neutral RL algorithms, while the last two correspond to risk-sensitive RL algorithms with DRM as the risk measure.

neutral and DRM-sensitive RL.

First, we derive a policy Hessian theorem for the DRM of cumulative discounted reward in a finite horizon MDP. Using this result, we employ the likelihood ratio method to arrive an estimator of the Hessian of DRM using sample trajectories. We establish MSE bounds of $\mathcal{O}(\frac{1}{b^{3/2}})$ for our DRM Hessian estimator, where b number of trajectories. Second, we propose a cubic-regularized policy Newton algorithm in the on-policy RL setting for maximizing the DRM of cumulative discounted reward in a finite horizon MDP. Third, we provide a non-asymptotic bound in expectation for convergence to an ϵ-second-order stationary point (ϵ-SOSP) of our algorithm. The sample complexity of our proposed algorithm is $\mathcal{O}(\epsilon^{-3.5})$. More importantly, our algorithm escapes saddle points. To the best of our knowledge, we are the first to present convergence of a policy gradient-type algorithm to an ϵ-SOSP convergence of a risk-sensitive objective. Finally, we conduct simulation experiments on the three environments, namely cliff walk, cart pole and humanoid. In each case, we find that our DRM-sensitive policy Newton algorithm finds a risk-seeking policy with a higher expected return than risk-neutral variant.

Related work. Various riskmetrics have been proposed in the literature. A popular class of riskmetrics is coherent risk measures (Artzner et al. 1999). A coherent risk measure is sub-additive, homogeneous, translation invariant, and monotonic. Value-at-risk (VaR) is a popular risk measure that is widely used in financial applications. However, VaR is not coherent. A closely related risk measure is Conditional value at risk (CVaR) (Rockafellar, Uryasev et al. 2000), known variously as tail VaR, and expected shortfall. Unlike VaR, CVaR is a coherent risk measure. Distortion risk measures generalize VaR and CVaR, and under a concave distortion function, these measures are coherent. Distortion riskmetrics (DRMs), as mentioned before, generalize distortion risk measures further by including non-monotone distortion functions. Examples of non-monotone DRMs include Gini deviation, Gini shortfall, mean-median deviation, inter-quantile range, inter-ES range etc. A popular non-coherent risk measure called cumulative prospect theory (CPT) was proposed in (Tversky and Kahneman 1992; Prashanth et al. 2016). CPT has been shown to model human preferences well through an inverted S-shaped distortion function. A risk measure that is closely related to CPT is rank-dependent expected utility (RDEU), see (Quiggin 2012).

Risk-sensitive RL using the policy gradient approach has received a lot of research attention recently, and a variety of risk measures including CVaR, mean-variance tradeoff, cumulative prospect theory, percentile criteria, have been explored, see (Prashanth and Fu 2022) for a survey. In (Anantharam and Borkar 2017), the authors consider an exponential sum of rewards in this example, and aim to find a risk-sensitive re-balancing strategy. By approximating the exponential sum of rewards as a sum of mean and a constant multiple of variance, we can intuitively reason that the aforementioned risk-sensitive strategy prefers volatile stocks with high returns. A similar interpretation in a closely related RL setting can be seen in (Chow et al. 2020).

In (Maniyar et al. 2024), authors proposed a risk-neutral cubic-regularized policy Newton algorithm in a finite horizon MDP and show the convergence to a ϵ-SOSP. In comparison, we consider a risk-sensitive objective based on DRMs. Unlike the risk-neutral case, our algorithm requires estimation of the DRM Hessian, which in turn requires empirical distribution functions (or sample means are not enough). Our DRM policy Hessian theorem and the Hessian estimate analysis (for an MSE bound) involve significant deviations from the risk-neutral case considered in (Maniyar et al. 2024).

In (Vijayan and Prashanth 2021), the authors propose a policy gradient algorithm for distortion risk measures and establish the convergence to an ϵ-FOSP. However, such a convergence guarantee is insufficient, as FOSPs often include saddle points. In contrast, our policy Newton algorithm avoids saddle points. More importantly, we consider a larger class of distortion riskmetrics that subsume distortion risk measures, and unlike (Vijayan and Prashanth 2021), we also demonstrate the practical utility of DRM-sensitive algorithm in three well-known RL benchmarks.

In (Tamar et al. 2015), the authors propose a policy gradient approach for a coherent risk measure and this subsumes DRMs under some conditions on the distortion function. However, their approach requires the solution of an optimization problem for risk estimation and gradient computation, and they establish asymptotic convergence to a

FOSP. In contrast, we have gradient estimates that are easy to implement, and we establish a non-asymptotic bound that shows convergence to an ϵ-SOSP.

The rest of the paper is organized as follows: In Section 2, we formulate the DRM-sensitive MDP. In Section 3, we present the policy Hessian theorem for DRM. In Section 4, we describe the DRM gradient and Hessian estimation schemes, and in Section 5, we present the DRM policy Newton algorithm. In Section 6, we establish convergence to an approximate SOSP for our proposed algorithm. In Section 7, we present the simulation experiments. Finally, in Section 8 we provide the concluding remarks. Owing to space constraints, the proofs of all the theoretical claims in this paper are provided in a longer version of this paper, which is available in (Pachal, Maniyar, and Prashanth 2025).

2 Problem Formulation

Let X be a random variable with cumulative distribution function (CDF) F. The distortion riskmetric (DRM) of X is defined by

$$\rho_h(X) = \int_{-\infty}^{0} [h(1-F(x)) - h(1)]dx + \int_{0}^{\infty} h(1-F(x))dx, \tag{1}$$

where the distortion function h is a function of bounded variation with $h(0) = 0$. DRM ρ_h is defined only for functions h that ensure both integrals in (1) are finite. For sub-Gaussian distributions, it is easy to see that $\rho_h(\cdot)$ is finite. Further, if $h(t) = t$, then DRM coincides with the expected value of X. Figure 1 presents a few popular DRMs.

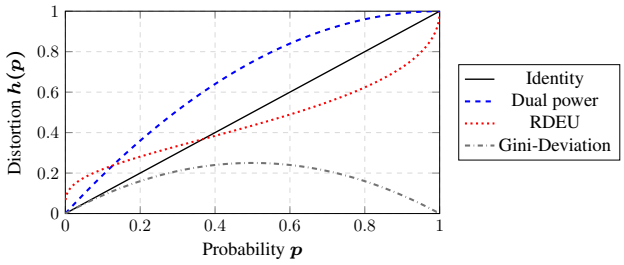


Figure 1: Examples of distortion functions. The choices $h(t) = t$, $h(t) = 1-(1-t)^2$, $h(t) = \exp(-\sqrt{-\ln t})$, $h(t) = t - t^2$ correspond to identity, dual power DRM, RDEU, and Gini deviation, respectively.

We integrate DRM into risk-sensitive RL problems to learn an optimal policy in the finite-horizon Markov decision process (MDP). We consider a finite-horizon MDP with finite state space $\mathbb{S}$, finite action space $\mathbb{A}$. A single stage reward r is defined as $r : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to [-r_{\max}, r_{\max}]$, where $r_{\max} \geq 0$. Let p be the probability transition function. Let the horizon or episode length be T, i.e., all episodes end after T steps. We parameterize the stochastic policies $\{\pi_\theta : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to [0,1]\}$ using $\theta \in \mathbb{R}^d$.

The cumulative discounted reward R^θ is given by $R^\theta = \sum_{t=0}^{T-1} \gamma^t r(S_t, A_t, S_{t+1})$, where $A_t \sim \pi_\theta(., S_t)$ is the action chosen at time t, $S_{t+1} \sim p(., S_t, A_t)$ is the next sate and $\gamma \in (0,1)$ is the discounting factor.

The optimization objective in a DRM-sensitive MDP for a given policy parameter θ is the DRM of R^θ. For notational simplicity, we shall use $\rho_h(\theta)$ to denote $\rho_h(R^\theta)$. Our aim is to find an optimal θ^* such that

$$\theta^* \in \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \rho_h(\theta). \tag{2}$$

A point θ^* is called a first-order stationary point (FOSP) if $\nabla \rho_h(\theta^*) = 0$. FOSPs do not coincide with local maxima as they include saddle points. To mitigate this issue, the notion of a second-order stationary point (SOSP) is considered in optimization literature, cf. (Nesterov and Polyak 2006; Jin et al. 2021). At an SOSP, the gradient vanishes and the Hessian is positive semi-definite.

Finding an FOSP/SOSP directly in an RL setting is not feasible as the model information is not available. Moreover, running a policy gradient-type algorithm for a finite number of steps would ensure convergence to an approximate FOSP or SOSP. We make the latter notions precise in the definition below.

Definition 2.1 (ϵ**-first and second-order stationary points).** Let $\epsilon > 0$. An output $\bar{\theta}$ of a stochastic iterative algorithm to solve (2) is said to be ϵ-first-order stationary point if $\mathbb{E}\left[\left\|\nabla \rho_h(\bar{\theta})\right\|\right] \leq \epsilon$ and an ϵ-second-order stationary point if $\max\left\{\sqrt{\mathbb{E}[\|\nabla \rho_h(\bar{\theta})\|]}, \frac{1}{\sqrt{\rho}}\mathbb{E}[\lambda_{\max}(\nabla^2 \rho_h(\bar{\theta}))]\right\} \leq \sqrt{\epsilon}$, for some $\rho > 0$.

We are interested in finding an ϵ-SOSP of $\rho_h(\cdot)$. Convergence to ϵ-SOSP would imply escaping saddle points if the Hessian is not degenerate, see (Anandkumar and Ge 2016).

3 DRM Policy Hessian Theorem

For deriving gradient and Hessian expressions for the DRM objective, we make the following assumptions:

(A1). *There exists $M_d > 0$ such that $\|\nabla \log \pi_\theta(a|s)\| \leq M_d$ for all $\theta \in \mathbb{R}^d, a \in \mathcal{A}, s \in S$.*

(A2). *There exists $M_h > 0$ such that $\|\nabla^2 \log \pi_\theta(a|s)\| \leq M_h$ for all $\theta \in \mathbb{R}^d, a \in \mathcal{A}, s \in S$.*

(A3). *There exists $M_{h'}, M_{h''}, M_{h'''} > 0$ such that $|h'(t)| \leq M_{h'}$, $|h''(t)| \leq M_{h''}$, and $|h'''(t)| \leq M_{h'''}$.*

(A4). *For any pair of parameters (θ_1, θ_2) and any state-action (s,a) there exists a L_2 such that*

$$\left\|\nabla^2 \log \pi_{\theta_1}(a|s) - \nabla^2 \log \pi_{\theta_2}(a|s)\right\| \leq L_2 \left\|\theta_1 - \theta_2\right\|.$$

Assumption (A1) and (A2) are frequently made for analysis in policy gradient and actor-critic algorithms, as shown in (Shen et al. 2019). Assumption (A3) is required for the boundedness of the DRM policy gradient and Hessian. Assumption (A4) is required for the analysis of second-order policy search algorithms (Zhang et al. 2020).

Now we present the gradient and Hessian expressions of the CDF $F_{R^\theta}(.)$ for policy parameter θ. These are derived using the following identity: $F_{R^\theta}(x) = \mathbb{E}\left[\mathbf{1}\{R^\theta \leq x\}\right]$, followed by differentiation w.r.t. θ on both sides. Gradient $\nabla F_{R^\theta}(\cdot)$ is required in arriving at $\nabla \rho_h(\theta)$, while $\nabla^2 F_{R^\theta}(\cdot)$ is needed for working out the expression for $\nabla^2 \rho_h(\theta)$.

Lemma 1. *Suppose assumptions (A1) -(A2) hold. Then for all $x \in (-M_r, M_r)$ with $M_r = \frac{r_{max}}{1-\gamma}$, we have*

$$\nabla F_{R^\theta}(x) = \mathbb{E}\left[\mathbf{1}\{R^\theta \le x\}\Phi_t\right], \text{ and}$$
$$\nabla^2 F_{R^\theta}(x) = \mathbb{E}\left[\mathbf{1}\{R^\theta \le x\}\sum_{t=0}^{T-1}\nabla^2 \log \pi_\theta(A_t|S_t)\right] + \mathbb{E}\left[\mathbf{1}\{R^\theta \le x\}\left(\Phi_t\right)\left(\Phi_t\right)^\intercal\right],$$

where $\Phi_t = \sum_{t=0}^{T-1}\nabla \log \pi_\theta(A_t|S_t)$.

We next state the policy gradient and Hessian theorem that provides expressions for the DRM gradient and Hessian, which in turn involves $\nabla F_{R^\theta}(\cdot)$ and $\nabla^2 F_{R^\theta}(\cdot)$.

Theorem 1 (DRM policy gradient and Hessian theorem). *Suppose Assumptions (A1) to (A2) hold. Then the gradient and Hessian of the DRM are given by*

$$\nabla \rho_h(\theta) = -\int_{M_r}^{M_r} h'(1 - F_{R^\theta}(x))\nabla F_{R^\theta}(x)dx, \text{ and}$$
$$\nabla^2 \rho_h(\theta) = \int_{-M_r}^{M_r} h''(1 - F_{R^\theta}(x))\nabla F_{R^\theta}(x)\nabla F_{R^\theta}(x)^\intercal dx - \int_{-M_r}^{M_r} h'(1 - F_{R^\theta}(x))\nabla^2 F_{R^\theta}(x)dx.$$

Next, we establish first as well as second-order smoothness of the DRM objective under the assumptions listed above. This result is essential for the analysis of the DRM policy Newton algorithm that we propose.

Lemma 2 (Smoothness of DRM). *Suppose Assumptions (A1) to (A4) hold. Then, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\nabla \rho_h(\theta_1) - \nabla \rho_h(\theta_2)\| \le G_{\mathcal{H}}\|\theta_1 - \theta_2\|, \quad (3)$$
$$\|\nabla^2 \rho_h(\theta_1) - \nabla^2 \rho_h(\theta_2)\| \le L_{\mathcal{H}}\|\theta_1 - \theta_2\|, \quad (4)$$

where $G_{\mathcal{H}} = 2M_r T(M_h M_{h'} + TM_d^2(M_{h'} + M_{h''}))$,
$\nu = (TM_h + T^2M_d^2)$, $L_{\mathcal{H}} = \xi_1 + \xi_2$ with
$\xi_1 = 2M_r(2M_{h''}TM_d\nu + T^3M_d^3M_{h'''})$ and
$\xi_2 = 2M_r(M_{h''}TM_d\nu + M_{h'}(TL_2 + 2TM_dM_h))$.

The expressions $\nabla \rho_h(.)$ and $\nabla^2 \rho_h(.)$ presented in Theorem 1 cannot be evaluated directly in a typical RL setting due to absence of model information. Instead, one can form a sample-based gradient and Hessian estimates for the DRM objective. We present these estimates in the next section.

4 DRM Policy Gradient and Hessian Estimation

In this section, we estimate ∇F_{R^θ} and $\nabla^2 F_{R^\theta}$ by using m and b episodes, respectively. Let R_i^θ be the cumulative reward in i-th episode. Here, A_t^i and S_t^i denote the action and state at time t in i-th episode.

DRM gradient estimation. Given independent and identically distributed (i.i.d.) samples $\{R_i^\theta, i = 1, \ldots, m\}$ of R^θ, let $G_{R^\theta}^m(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{R_i^\theta \le x\}, \forall x \in \mathbb{R}$, denote the EDF of $F_{R^\theta}(.)$. Following (Vijayan and Prashanth 2021), we form the estimate $\widehat{\nabla} G_{R^\theta}^m(.)$ of $\nabla F_{R^\theta}(.)$ as follows: $\forall x \in \mathbb{R}$,

$$\widehat{\nabla} G_{R^\theta}^m(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{R_i^\theta \le x\}\sum_{t=0}^{T-1}\nabla \log \pi_\theta(A_t^i|S_t^i). \quad (5)$$

Using $\widehat{\nabla} G_{R^\theta}^m(\cdot)$, we form the gradient estimate $\widehat{\nabla}\rho_h(\theta)$ as follows:

$$\widehat{\nabla}\rho_h(\theta) = -\int_{-M_r}^{M_r} h'(1 - G_{R^\theta}^m(x))\widehat{\nabla} G_{R^\theta}^m(x)dx. \quad (6)$$

DRM Hessian estimation. Given i.i.d. samples $\{R_i^\theta, i = 1, \ldots, b\}$ of R^θ, we form an estimate $\widehat{\nabla}^2 G_{R^\theta}^b(.)$ of $\nabla^2 F_{R^\theta}(.)$ as follows: $\forall x \in \mathbb{R}$,

$$\widehat{\nabla}^2 G_{R^\theta}^b(x) = \frac{1}{b}\sum_{i=1}^b \mathbf{1}\{R_i^\theta \le x\}\sum_{t=0}^{T-1}\nabla^2 \log \pi_\theta(A_t^i|S_t^i) + \frac{1}{b}\sum_{i=1}^b \mathbf{1}\{R_i^\theta \le x\}\left[\sum_{t=0}^{T-1}\nabla \log \pi_\theta(A_t^i|S_t^i)\right] \times \left[\sum_{t=0}^{T-1}\nabla^\intercal \log \pi_\theta(A_t^i|S_t^i)\right] \quad (7)$$

Using $\widehat{\nabla}^2 G_{R^\theta}^b(\cdot)$, we form the Hessian estimate $\widehat{\nabla}^2\rho_g(\theta)$ is as follows:

$$\widehat{\nabla}^2\rho_h(\theta) = \int_{-M_r}^{M_r} h''(1 - G_{R^\theta}^m(x))\widehat{\nabla} G_{R^\theta}^m(x)\widehat{\nabla} G_{R^\theta}^m(x)^\intercal dx - \int_{-M_r}^{M_r} h'(1 - G_{R^\theta}^m(x))\widehat{\nabla}^2 G_{R^\theta}^b(x)dx. \quad (8)$$

The DRM Hessian estimate in (8) can be computed in a computationally efficient fashion using order statistics of the b samples $\{R_i^\theta\}_{i=1}^b$, see Appendix D of (Pachal, Maniyar, and Prashanth 2025) for further details.

5 DRM-Sensitive Policy Newton Algorithm

In general, the classical Newton method fails to escape saddle points for non-convex objective functions. In (Nesterov and Polyak 2006), authors proposed a cubic-regularized Newton algorithm for escaping saddle points by adding a cubic term in the auxiliary objective. We adopt this cubic regularization technique in a risk-sensitive RL framework with DRM as the objective. In particular, our proposed cubic-regularized policy Newton algorithm for DRM (CRPN-DRM) performs the following update:

$$\theta_{k+1} = \underset{\theta \in \mathbf{R}^d}{\text{argmax}}\left\{\left\langle \widehat{\nabla}\rho_h(\theta_k), \theta - \theta_k\right\rangle + \frac{1}{2}\left\langle \widehat{\nabla}^2\rho_h(\theta_k)(\theta - \theta_k), \theta - \theta_k\right\rangle - \frac{\alpha}{6}\left\|\theta - \theta_k\right\|^3\right\}, \quad (9)$$

where α is a regularization parameter, $\widehat{\nabla}\rho_h$ and $\widehat{\nabla}^2\rho_h$ are the gradient and Hessian estimates, defined in (6) and (8), respectively. We re-write these gradient and Hessian estimates in an alternate form, by first defining the following quantities:

$$c_i'' := \begin{cases} (R^\theta_{(i+1)} - R^\theta_{(i)})h''(1-\frac{i}{\tau}), & i \in \{1,\ldots,\tau-1\}, \\ (M_r - R^\theta_{(\tau)})h''_+(0), & i = \tau, \end{cases}$$
$$c_i' := \begin{cases} (R^\theta_{(i+1)} - R^\theta_{(i)})h'(1-\frac{i}{\tau}), & i \in \{1,\ldots,\tau-1\} \\ (M_r - R^\theta_{(\tau)})h'_+(0), & i = \tau, \end{cases} \tag{10}$$

where $\psi_i' = \sum_{j=i}^{\tau} c_j'$, $l^\theta_{(i)} = \sum_{t=0}^{T-1} \log \pi_\theta(A^i_t|S^i_t)$, and $s^\theta_i = \sum_{j=1}^{i} l^\theta_{(j)}$. Then, $\widehat{\nabla}\rho_h(\theta) = -\frac{1}{\tau}\sum_{i=1}^{\tau} \psi_i' \nabla l^\theta_{(i)}$, and

$$\widehat{\nabla}^2\rho_h(\theta) = \frac{1}{\tau}\sum_{i=1}^{\tau} \frac{c_i''}{b} \nabla s^\theta_i \nabla^\intercal s^\theta_i - \psi_i'[\nabla^2 l^\theta_{(i)} + \nabla l^\theta_{(i)} \nabla^\intercal l^\theta_{(i)}].$$

Here $\tau \in \mathbb{N}$ denotes the number of episodes simulated and can be different for the gradient and Hessian estimates. The reader is referred to Appendix D of (Pachal, Maniyar, and Prashanth 2025) for a proof of the above equivalence for gradient and Hessian expressions. Note that the above simplified expressions make it suitable for practical implementations as it avoids redundant gradient computations. Inspired by (Markowitz et al. 2023), we further propose a variance-induced estimate by ignoring the cross-episode terms, as these terms in expectation have zero mean. The result below provides alternative expressions for DRM gradient and Hessian, while incorporating variance reduction.

Lemma 3 (Variance-reduced DRM estimates)**.** *Suppose Assumptions (A1) to (A4) hold. Then the DRM gradient* (6) *and Hessian estimates* (8) *form can be re-written as follows:*

$$\widehat{\nabla}\rho_h(\theta) = \frac{1}{m}\sum_{i=1}^{m} R^\theta_{(i)} h'\left(1-\frac{i}{m}\right)\nabla l^\theta_{(i)},$$

$$\widehat{\nabla}^2\rho_h(\theta) = \frac{1}{b}\sum_{i=1}^{b}\Bigg[\psi_i''\nabla l^\theta_{(i)}\nabla^\intercal l^\theta_{(i)} + R^\theta_{(i)} h'(1-\frac{i}{b})\left(\nabla^2 l^\theta_{(i)} + \nabla l^\theta_{(i)}\nabla^\intercal l^\theta_{(i)}\right)\Bigg],$$

where $\psi_i'' = \frac{1}{b}\sum_{j=i}^{b} c_j''$, *and* $l^\theta_{(i)} = \sum_{t=0}^{T-1}\log\pi_\theta(A^i_t|S^i_t)$, *with* m *and* b *episodes simulated for gradient and Hessian estimation, respectively.*

6 Main Results

Before presenting the main ϵ-SOSP convergence result, we provide the error bounds for the Hessian estimate $\widehat{\nabla}^2\rho_h(\theta)$ in terms of number of episodes m and b, respectively. An error bound in similar spirit for $\widehat{\nabla}\rho_h(\theta)$ was derived in (Vijayan and Prashanth 2021).

Lemma 4. *Suppose assumptions (A1) - (A4) hold. Let the Hessian estimate* $\widehat{\nabla}^2\rho_h(\theta)$ *be computed by* (8) *with* b *number of trajectories. Suppose* $m \geq b$ *and* $b \geq C(d)$, *where* $C(d) = 4(1+2\log 2d)$. *Then,*

$$\mathbb{E}\left[\left\|\widehat{\nabla}^2\rho_h(\theta) - \nabla^2\rho_h(\theta)\right\|^2\right] \leq \frac{t_1+\kappa_2}{b}, \text{ and}$$
$$\mathbb{E}\left[\left\|\widehat{\nabla}^2\rho_h(\theta) - \nabla^2\rho_h(\theta)\right\|^3\right] \leq \frac{\sqrt{(\kappa_3+t_2)(\kappa_2+t_1)}}{b^{\frac{3}{2}}}, \tag{11}$$

where $\nu = (TM_h + T^2M_d^2)$,
$t_1 = 32M_r^2M_{h'}^2C(d)\nu^2$, $t_2 = 1920M_r^2M_{h'}^4d^2\nu^4$,
$\kappa_2 = 64M_r^2(3e^2M_{h''}^2T^4M_d^4 + 2T^4M_d^4M_{h'''}^2 + M_{h''}^2\nu^2)$,
$\kappa_3 = 4096M_r^2[T^8M_d^8(9e^2M_{h''}^4 + 8M_{h'''}^4) + M_{h''}^4\nu^4]$. *The constants* $M_h, M_d, M_{h'}, M_{h''}, M_{h'''}$ *are specified in Assumptions (A1) to (A4).*

The main result that establishes the convergence of our algorithm to an ϵ-SOSP is given below.

Theorem 2 (Convergence to ϵ-SOSP). *Let the assumptions (A1) - (A4) hold. Let* $\{\theta_1,\ldots,\theta_N\}$ *denote the iterates obtained by running CRPN-DRM algorithm for* N*-iterations with the following parameters:*

$$\alpha_k = 3L_{\mathcal{H}}, \quad N = \frac{12\sqrt{L_{\mathcal{H}}}\,(\rho(\theta_0)-\rho^*)}{\epsilon^{\frac{3}{2}}},$$
$$m_k = \frac{25\kappa_1}{4\epsilon^2}, \quad b_k = \frac{9\sqrt[3]{2(\kappa_3+t_2)(\kappa_2+t_1)}}{L_{\mathcal{H}}\epsilon}, \tag{12}$$

where $\kappa_1 = 32M_r^2T^2M_d^2(e^2M_{h'}^2 + M_{h''}^2)$ *and* $\kappa_2, \kappa_3, t_1, t_2$ *are defined in Lemma 4. Let* $\bar{\theta}$ *be chosen from* $\{\theta_1,\ldots,\theta_N\}$ *uniformly at random. Then, for any* $0 < \epsilon \leq \frac{25L_{\mathcal{H}}\kappa_1}{36\sqrt[3]{2(\kappa_3+t_2)(\kappa_2+t_1)}}$, *we have*

$$5\sqrt{\epsilon} \geq \max\left\{\sqrt{\mathbb{E}[\|\nabla\rho_h(\bar{\theta})\|]}, \frac{5}{6\sqrt{L_{\mathcal{H}}}}\mathbb{E}[\lambda_{\max}(\nabla^2\rho_h(\bar{\theta}))]\right\}. \tag{13}$$

As a consequence of Theorem 2, the sample complexity of our CRPN-DRM algorithm to converge a ϵ-SOSP is bounded by $\mathcal{O}(\epsilon^{-3.5})$.

7 Simulation Experiments

The implementation is available at https://github.com/mizhaan23/drm_rl. We compare the performance of the following algorithms on three environments, namely Cart-pole, Humanoid, and cliff walking.

1. **ACRPN:** This is the risk-neutral cubic-regularized policy Newton algorithm from (Maniyar et al. 2024). We implement the approximate variant of their policy Newton algorithm, as described in Section 5 of (Maniyar et al. 2024).
2. **DRMACRPN:** This is the algorithm described in Section 5. We implement DRMACRPN on the three environments using the variance-reduced gradient and Hessian estimates as shown in Lemma 3. We consider the following two DRMs: (i) Dual-power with $h(t) = 1-(1-t)^2$; and (ii) Gini-deviation with $h(t) = t - t^2$. Appendix F of (Pachal, Maniyar, and Prashanth 2025) describes DRMACRPN implementation using only Hessian vector products, making it computationally comparable with first-order algorithms for deep RL.
3. **REINFORCE-DRM:** Uses the gradient estimate in

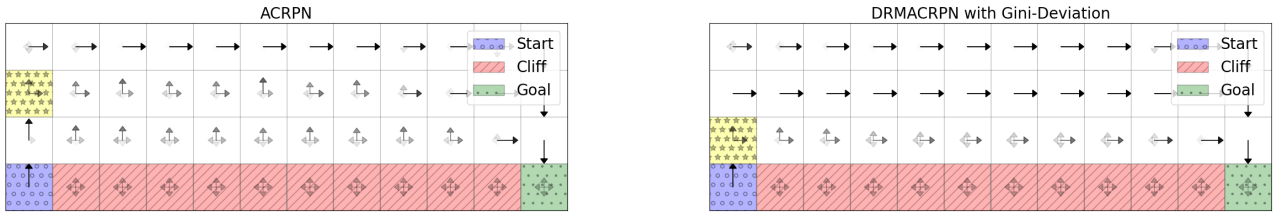


Figure 2: A visual illustration of the policies found by ACRPN and DRMACRPN algorithms on cliff walk environment. The length of the arrows is proportional to the probability of taking that action by the policy. The bottom left and right cells denote the start and goal states, respectively.

Lemma 3 and performs the traditional single step gradient ascent with the step-size $\lambda = \sqrt{\frac{2}{\alpha||\bar{g}||}}$, as suggested in (Maniyar et al. 2024). This step-size is obtained by solving the cubic sub-problem in (9) while assuming $\mathcal{H} = 0$, i.e., without having access to second-order information.

For DRMACRPN, we set the cubic-penalty parameter $\alpha = 10^5$ as that gave the most stable learning behavior across DRMs. The batch sizes for gradient and Hessian estimation are set as follows: $m_k = b_k = 200$. The algorithm was run for $N = 100$ iterations, using the same episodes for the gradient as well as Hessian estimation. For ease of implementation, we assume $M_r = R^{\theta}_{(m)}$. The policy parameterization is tabular, linear and deep neural network on cliff walk, cart pole and humanoid environments, respectively. For the humanoid case, the policy parameterization involved a deep neural network consisting of two hidden layers of 64 neurons each, with softmax activation. All the policies use a Boltzmann distribution in the final layer for the probabilities to sum up to one. The results reported are averages over ten independent replications. Experiments were conducted on a personal laptop (Windows 11 Home) with NVIDIA RTX 3070Ti GPU (8GB GDDR6 VRAM), Intel(R) Core(TM) i7-12700H CPU with 64GB of DDR5 RAM.

Cliff walk. We implement a variant of this environment with a modified reward scheme which incentivizes the agent to move towards the goal state. The modified reward incorporates the distance from the goal state, and is given by

$$r_{\texttt{modified}}(s, a, s') = r_{\texttt{default}}(s, a, s') - c \cdot ||s' - s_g||_1,$$

where (s, a, s') are the current state, current action and next state, respectively, s_g is the coordinate of the goal state and $||\cdot||_1$ denotes the ℓ_1 norm. In our experiments, we set $c = 0.5$. Note that, for the purposes of understanding the results, we only consider $r_{\texttt{default}}$ while the agent is trained on $r_{\texttt{modified}}$. Cliff walk serves as a good environment to study the performance for the risk-neutral and DRM policies owing to the risk involved in the path close to the cliff, which can result in very high negative rewards.

We present the mean, standard deviation, minimum and maximum of the episodic return in Section 7. It is apparent that the risk-neutral algorithms have lower standard deviation and their *best* episode – with a cumulative reward of -16 – follows the path that walks along the border of the grid, farthest away from the cliff. On the other hand, risk-seeking algorithms follow the shortest "riskier" path – with a cumulative reward of -12 – closer to the cliff, and this path has a higher risk of falling off the cliff. It is also interesting to note that risk-seeking policies may tend to fall off the cliff, as is evident by the minimum values of -123.

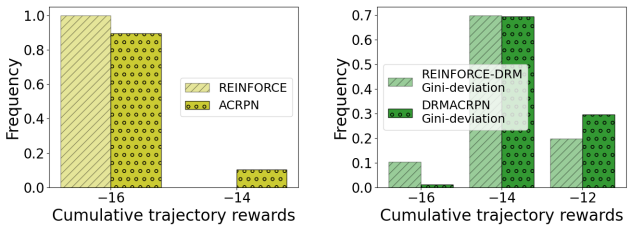
Algorithm	**Mean $\pm$ std**	**Min**	**Max**
REINFORCE	-16.2 ± 0.5	**-22**	-16
ACRPN	-16.0 ± 0.9	-24	-16
REINFORCE-DRM	$-14.1 \pm \mathbf{5.4}$	-123	**-12**
DRMACRPN	$\mathbf{-13.6 \pm 4.9}$	-123	**-12**

Table 2: Episodic return for risk-neutral algorithms (REINFORCE, ACRPN) and their DRM counterparts with Gini deviation on cliff walking environment.

To get deeper insights into the policies' actions on the grid, we present a visualization in Figure 2. It is interesting to note that the less risky ACRPN policy, tends to move away from the cliff even in the first row (the row second-closest to the cliff), while the DRMACRPN policy are highly determined to move towards the goal with shorter path lengths. In the second row (the row closest to the cliff), DRMACRPN policy is more likely to move towards the goal than move away from the cliff. This highlights the urgency of the DRM-sensitive policy to get to the goal state despite the low but non-zero chance of falling off the cliff. Furthermore, noticing the cells highlighted in yellow (star), we see that the DRMACRPN's risk-seeking policy tends to go towards the goal earlier than the risk-neutral polices, and DRMACRPN, which is a second-order algorithm, tends to be more likely to do so than its first-order counterpart (REINFORCE-DRM), see (Pachal, Maniyar, and Prashanth 2025) for more details.

Second-order vs first-order algorithms. We plot the reward distributions for cliff walk in Figure 3 after removing outliers, i.e. cumulative rewards with very low frequencies. From Figure 3, it is apparent that the second-order algorithms converge to the optimal policy more often than their first-order counterparts.

Humanoid. Figure 5 presents the following quantities for each of the algorithms mentioned above: (i) the histogram of the policies obtained after N iterations; and (ii) the learning curves, which show the change in episodic return. Section 7 tabulates the mean and variance of the episodic return for



(a) Risk-neutral policies. (b) Risk-seeking policies.

Figure 3: Distributions of the policies comparing first-order vs second-order. Generated using 10 independent runs for each policy after sufficient convergence.

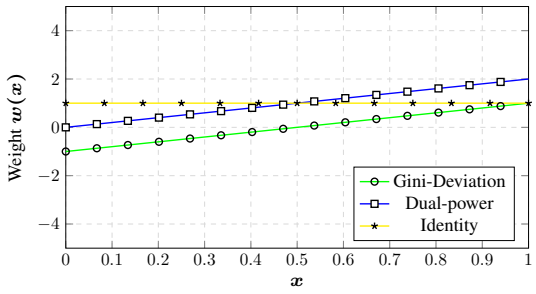


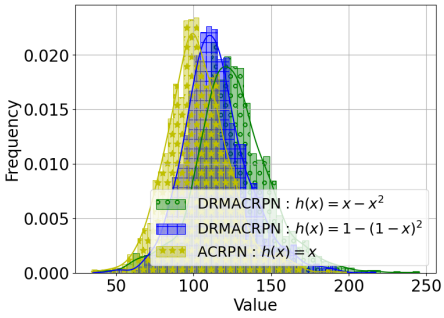
Figure 4: Weight coefficient $w(x) = h'(1-x)$ for $x \in [0,1]$. The return of the i-th episode in the gradient estimate is scaled by $w(\frac{i}{m})$, see Lemma 3.

our DRMACRPN algorithm with different distortion functions. Further, Table 6 of (Pachal, Maniyar, and Prashanth 2025) compares the performance of our proposed algorithm with the risk-neutral baseline. From these results, it is apparent that our algorithm performs better than the risk-neutral counterpart w.r.t. the DRM objective. To put it differently, the risk-neutral objective is not a surrogate for DRM, and we need a specialized algorithm to optimize the latter objective. Figure 5 presents the performance of DRM-sensitive and risk-neutral algorithms on the humanoid environment for the dual-power and Gini-deviation DRMs. We observe that our algorithm outperforms the risk-neutral policy Newton, both from the learning curves viewpoint as well as the reward distributions shown in Figure 5b.

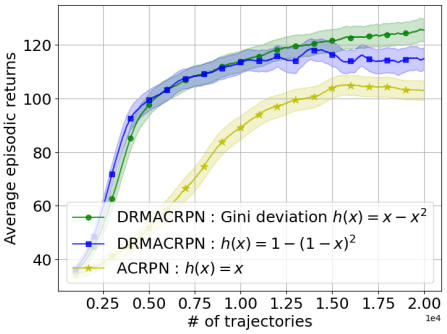
We believe the superior performance of dual-power and Gini deviation is due to the fact that these DRMs weigh episodes with higher returns more as is evident from the gradient expression in Lemma 3 and from the plot of weight coefficients given in Figure 4. In other words, each step of DRMACRPN would be in the direction of policies that achieve higher returns, owing to higher weights from DRM gradient/Hessian estimates, potentially leading to faster learning by *focusing on the best experiences*. Figure 5a shows the distribution of the policies obtained for the different algorithms. We also note that although Gini-deviation gives us better mean, it has thicker tails as compared to the Dual-power DRM. This result is owing to the fact that Gini deviation is closely related to variance, and the algorithm learns to

DRM	Mean	Standard deviation
Identity	103.0	20.6
Dual-power	114.4	21.1
Gini deviation	**124.8**	**24.7**

Table 3: Sample mean and variance of the episodic return for DRMACRPN with three different distortion functions on Humanoid-v4.



(a) Episodic return distributions of the DRM policies



(b) Learning curves of the DRM policies.

Figure 5: Performance comparison of risk-neutral policy Newton and DRMACRPN with two different DRMs, namely dual-power and Gini deviation on the Humanoid-v4 environment with $\alpha = 10^5$.

maximize the same. Similar inferences can be made from the experiments on the Cart-pole environment, see Appendix G of (Pachal, Maniyar, and Prashanth 2025) for further details.

8 Conclusions

We proposed a policy Newton algorithm for maximizing distortion riskmetrics. We derived the policy Hessian theorem for a DRM objective, and used it to form DRM gradient/Hessian estimates using episodes. We established the convergence of our policy Newton algorithm to an ϵ-SOSP for the DRM objective with sample complexity $\mathcal{O}(\epsilon^{-3.5})$. To the best of our knowledge, this is the first work to present convergence to an ϵ-SOSP in a risk-sensitive RL framework. Finally, we showed the superiority of our DRM-sensitive policy Newton algorithm over its risk-neutral counterpart and REINFORCE through simulation experiments.

References

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

Anandkumar, A.; and Ge, R. 2016. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on learning theory*, 81–102. PMLR.

Anantharam, V.; and Borkar, V. S. 2017. A variational formula for risk-sensitive reward. *SIAM Journal on Control and Optimization*, 55(2): 961–988.

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

Chow, Y.; Cui, B.; Ryu, M.; and Ghavamzadeh, M. 2020. Variational model-based policy optimization. *arXiv preprint arXiv:2006.05443*.

Cover, T. M. 1991. Universal portfolios. *Mathematical finance*, 1(1): 1–29.

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

Fröhlich, C.; and Williamson, R. C. 2024. Risk Measures and Upper Probabilities: Coherence and Stratification. *Journal of Machine Learning Research*, 25(207): 1–100.

Furman, E.; Wang, R.; and Zitikis, R. 2017. Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. *Journal of Banking & Finance*, 83: 70–84.

Gini, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche [Fasc. I.]*. Bologna, Italy: Tipogr. di P. Cuppini.

Gzyl, H.; and Mayoral, S. 2008. On a relationship between distorted and spectral risk measures. *Revista de Economía Financiera*, 15: 8–21.

Han, X.; Wang, R.; and Zhou, X. Y. 2022. Choquet regularization for reinforcement learning. *arXiv:2208.08497*.

Jin, C.; Netrapalli, P.; Ge, R.; Kakade, S. M.; and Jordan, M. I. 2021. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2): 1–29.

Jones, B. L.; and Zitikis, R. 2003. Empirical estimation of risk measures and related quantities. *North American Actuarial Journal*, 7(4): 44–54.

Jorion, P. 1996. Risk2: Measuring the risk in value at risk. *Financial analysts journal*, 52(6): 47–56.

Luo, Y.; Liu, G.; Poupart, P.; and Pan, Y. 2023. An Alternative to Variance: Gini Deviation for Risk-averse Policy Gradient. *arXiv:2307.08873*.

Maniyar, M. P.; Prashanth, L. A.; Mondal, A.; and Bhatnagar, S. 2024. A Cubic-regularized Policy Newton Algorithm for Reinforcement Learning. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 238: 4708–4716.

Markowitz, J.; Gardner, R. W.; Llorens, A.; Arora, R.; and Wang, I.-J. 2023. A Risk-Sensitive Approach to Policy Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 15019–15027.

Nesterov, Y.; and Polyak, B. T. 2006. Cubic regularization of Newton method and its global performance. *Mathematical programming*, 108(1): 177–205.

Pachal, S.; Maniyar, M. P.; and Prashanth, L. A. 2025. Policy Newton methods for Distortion Riskmetrics. arXiv:2508.07249.

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

Prashanth, L. A.; Jie, C.; Fu, M. C.; Marcus, S. I.; and Szepesvári, C. 2016. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, 1406–1415. PMLR.

Quiggin, J. 2012. *Generalized Expected Utility Theory: The Rank-dependent Model*. Springer Science & Business Media.

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

Shen, Z.; Ribeiro, A.; Hassani, H.; Qian, H.; and Mi, C. 2019. Hessian aided policy gradient. In *International conference on machine learning*, 5729–5738. PMLR.

Tamar, A.; Chow, Y.; Ghavamzadeh, M.; and Mannor, S. 2015. Policy Gradient for Coherent Risk Measures. In *Advances in Neural Information Processing Systems*, volume 28, 1468–1476.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5: 297–323.

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

Wang, R.; Wei, Y.; and Willmot, G. E. 2020. Characterization, Robustness, and Aggregation of Signed Choquet Integrals. *Mathematics of Operations Research*, 45(3): 993–1015.

Wang, S. 1996. Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1): 71–92.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8: 229–256.

Yang, L.; Zheng, Q.; and Pan, G. 2021. Sample complexity of policy gradient finding second-order stationary points. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10630–10638.

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