Approximate gradient ascent methods for distortion risk measures

Nithia Vijayan and Prashanth L.A.

Abstract—We propose approximate gradient ascent algorithms for risk-sensitive reinforcement learning control problem in on-policy as well as off-policy settings. We consider episodic Markov decision processes, and model the risk using distortion risk measure (DRM) of the cumulative discounted reward. Our algorithms estimate the DRM using order statistics of the cumulative rewards, and calculate approximate gradients from the DRM estimates using a smoothed functional-based gradient estimation scheme. We derive non-asymptotic bounds that establish the convergence of our proposed algorithms to an approximate stationary point of the DRM objective.

I. INTRODUCTION

The objective in reinforcement learning (RL) is to find a policy which maximizes the mean of the cumulative reward. But risk-sensitive RL goes beyond the mean of the cumulative reward, and considers other aspects of the reward distribution such as variance, tail probabilities, and shape. Such attributes are quantified using a risk measure.

Though there are no scarcity for risk measures in literature, there is no consensus on an ideal risk measure. A risk measure is said to be coherent if it is translation invariant, sub-additive, positive homogeneous, and monotonic [1]. Coherent risk measures are very desirable as the aforementioned properties help to avoid inconsistent decisions. Later, [2] suggest that coherency may not be sufficient, and introduce a new smooth coherent risk measure. The risk measures like Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR) [3], and cumulative prospect theory (CPT) [4] are studied in RL literature. The popular risk measures like VaR and CVaR attained a lot of criticism in the past since these measures overlook any information from infrequent high severity outcomes. Though risk-neutral RL gives equal focus to all the outcomes, it is intuitive to emphasize desirable events, and de-emphasize undesirable events without ignoring infrequent extreme outcomes altogether.

A family of risk measures called distortion risk measures (DRM) [5], [6] uses a distortion function to distort the original distribution, and calculate the mean of the rewards with respect to the distorted distribution. A distortion function allows one to vary the emphasis on each possible reward value. The choice of the distortion function governs the risk measure. Further, choosing a concave distortion function ensures that the DRM is coherent [7]. The spectral risk functions are equivalent to distortion functions [8]. A DRM with an identity distortion function is simply the mean of the rewards. The popular risk measures like VaR and CVaR can be expressed as a DRM using appropriate distortion functions. But the distortion function is discontinuous for VaR, and though continuous, it is not differentiable at every point for CVaR. Hence in [2], the author disfavors such distortion functions and focuses on smooth distortion functions.

In this paper, we consider the family of DRMs with smooth distortion functions. Some examples of smooth distortion functions are dual-power function, quadratic function, square-root function, exponential function, and logarithmic function (see [9], [10] for more examples). In risk-neutral RL, the occasional extreme events get equal priority as other events. In a DRM, the distortion function is operating on the reward distribution without discarding any information. Hence, it is possible to emphasize frequent events, and still account for infrequent high severity events. As there is no universal ideal risk measure, it is intuitive to consider a risk measure which best fits the problem in hand. For DRMs, we may concentrate only on the distortion function.

In this paper, we consider optimizing the DRM in a risk-sensitive RL context. 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 consider this problem in on-policy as well as off-policy settings, and employ the gradient ascent solution approach. Solving a DRM-sensitive MDP is challenging for two reasons. First, DRM is a risk measure that focuses on the entire distribution of the cumulative reward, while the regular value function objective in a risk-neutral RL setting is concerned with only the mean of this distribution. This observation implies a sample average of the total reward across sample episodes would not be sufficient to estimate DRM. Secondly, a gradient ascent algorithm requires an estimate of the gradient of the DRM objective, and such gradient information is not directly available in a typical RL setting. For the risk-neutral case, one has the policy gradient theorem, which leads to a straightforward gradient estimate from sample episodes.

For estimating DRM from sample episodes, we use the empirical distribution function (EDF) as a proxy for the true distribution. We provide a non-asymptotic bound on the mean-square error of this estimator, and this may be of independent interest. Next, to estimate the DRM gradient, we employ the smoothed functional (SF) method [11], [12], [13]. We use a variant of SF which use two function measurements corresponding to two perturbed policies. An SF-based estimation scheme may be restrictive for some applications in an on-policy RL setting, since we need separate sets of episodes corresponding to two perturbed policies. But, in an off-policy RL context, we only need a single set of episodes corresponding to a behavior policy. We provide

Nithia Vijayan and Prashanth L.A. are with Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. nithiav@cse.iitm.ac.in, prashia@cse.iitm.ac.in
bounds on the bias and variance of the aforementioned gradient estimates. Using these bounds, we establish that our DRM gradient ascent algorithms require $O\left(1/\epsilon^2\right)$ iterations to find an $\epsilon$-stationary point of the DRM objective. To the best of our knowledge, non-asymptotic bounds have not been derived for an SF-based DRM gradient ascent algorithm in the current literature.

**Related work.** In [14], 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 [15], the authors survey policy gradient algorithms for optimizing different risk measures in a constrained as well as an unconstrained RL setting. In a non-RL context, the authors in [16] 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. In [17], authors study DRM, and derive a policy gradient theorem that caters to the DRM objective. They establish non-asymptotic bounds for their policy gradient algorithms which uses likelihood ratio (LR) based gradient estimation scheme. In [18] 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. In comparison to the aforementioned works, we would like to note the following aspects:

(i) For the DRM measure, we estimate the gradient using SF-based estimation scheme while [17] uses a LR-based gradient estimation scheme. Similar to our work, [17] establishes a convergence rate $O(1/\sqrt{N})$ that implies convergence to a stationary point of the DRM objective. Here $N$ denotes the number of iterations of the DRM gradient ascent algorithm. But the algorithms in [17] require $\sqrt{N}$ episodes per iteration for both on policy and off-policy RL settings, whereas our algorithm for off-policy RL setting requires only a constant $m$ episodes per iteration, though our algorithm for on-policy RL-setting require $2mN$ episodes per iteration. The algorithms in [17] directly estimate the gradient using order statistics. Our algorithms uses a two part estimation scheme, where we first estimate the DRM using order statistics, and then estimate its gradient using SF-based estimation scheme.

(ii) For a general coherent risk measure, [14] uses gradient estimation scheme which requires solving a convex optimization problem, whereas our algorithms can directly estimate the gradient from the samples without solving any optimization sub-problem.

(iii) In [18], the guarantees for a gradient ascent algorithm based on SPSA are asymptotic in nature, and is for CPT in an on-policy RL setting. CPT is also based on a distortion function, but the distortion function underlying CPT is neither concave nor convex, and hence, it is non-coherent.

(iv) In [15], the authors derive a non-asymptotic bound of $O(1/N^{1/3})$ for an abstract smooth risk measure. They use abstract gradient oracles which satisfies certain bias-variance conditions. In contrast, we provide concrete gradient estimation schemes in RL settings, and our bounds feature an improved rate of $O(1/\sqrt{N})$.

The rest of the paper is organized as follows: Section II describes the DRM-sensitive MDP. Section III introduces our algorithms, namely DRM-OnP-SF and DRM-OffP-SF. Section IV presents the non-asymptotic bounds for our algorithms. Finally, Section V provides the concluding remarks.

## II. Problem Formulation

### A. Distortion risk measure (DRM)

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

$$
\rho_g(X) = \int_0^\infty (g(1-F_X(x))-1)dx + \int_0^\infty g(1-F_X(x))dx.
$$

The distortion function $g : [0,1] \rightarrow [0,1]$ 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 given in Table I and their plots in Figure 1 (cf. [17]).

The DRMs are well studied from an ‘attitude towards risk’ perspective, and we refer the reader to [19], [20] for details. In this paper, we focus on ‘risk-sensitive decision making under uncertainty’, with DRM as the chosen risk measure. We incorporate DRMs into a risk-sensitive RL framework, and the following section describes our problem formulation.

### Table I: Examples of distortion functions

| Distortion Function      | Formula                  | Parameter(s) |
|-------------------------|--------------------------|--------------|
| 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    | $g(s) = \frac{\sqrt{r}rs - 1}{\log(1+rs)}$, $r > 0$ |              |
| Logarithmic function    | $g(s) = \frac{\log(1+rs)}{rs}$, $r > 0$ |              |

**Fig. 1:** Examples of distortion functions
Comparing (4) with (1), it is apparent that we have used the polynomial approximation to capture the cumulative discounted reward. Now, we form an estimate of the action selection distribution

\[ \pi_{\theta} = \mathbb{E}_{S_0 \sim \mathcal{D}}[\pi_{\theta}(S_0)] \]

where \( \mathcal{D} \) is the distribution of the initial state. This is done by sampling initial states from the distribution \( \mathcal{D} \) and estimating the action selection distribution at each state. The optimization problem in (2) can be solved by a gradient ascent algorithm. But, in a typical RL setting, we consider the expected reward in an off-policy RL setting.

Our goal is to find the parameterized policies \( \pi_{\theta} \) that maximize the expected reward. The optimization problem in (2) can be solved by a gradient ascent algorithm. However, in a typical RL setting, we consider the expected reward in an off-policy RL setting.

We generate \( m \) episodes using the policy \( \pi_{\theta} \), and estimate the CDF of the expected reward using sample averages. We denote by \( \hat{\rho}_g(\theta) \) the cumulative reward of the episode \( i \). We form the estimate \( \hat{\rho}_g(\theta) \) of the expected reward of the episode \( i \) as follows:

\[ \hat{\rho}_g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{S_0 \sim \mathcal{D}}[r(S_0)] \]

where \( m \) is the number of episodes needed in each iteration of our algorithm. The gradient estimate is averaged over \( n \) unit vectors to reduce the variance.

The update iteration in DRM-OnP-SF is as follows:

\[ \theta_{k+1} = \theta_k + \alpha \hat{\nabla}_{\mu,n} \hat{\rho}_g(\theta_k) \]

where \( \theta_0 \) is set arbitrarily, and \( \alpha \) is the step-size. Algorithm 1 presents the pseudocode of DRM-OnP-SF.
We can simplify (13) in terms of order statistics as
\[
\hat{\rho}^H_g(\theta) = R^b_{(1)} + \sum_{i=2}^{m} R^b_{(i)} g \left( 1 - \min \left\{ 1, \frac{1}{m} \sum_{k=1}^{i-1} \psi^\theta_{(k)} \right\} \right) - \sum_{i=1}^{m-1} R^b_{(i)} g \left( 1 - \min \left\{ 1, \frac{1}{m} \sum_{k=1}^{i} \psi^\theta_{(k)} \right\} \right),
\]
where \( R^b_{(i)} \) is the \( i \)-th smallest order statistic of the samples \( \{ R^b_1, \ldots, R^b_m \} \), and \( \psi^\theta_{(i)} \) is the importance sampling ratio of \( R^b_{(i)} \). The reader is referred to Lemma 2 in Appendix I for a proof.

2) DRM gradient estimation: We use the SF-based gradient estimation scheme as in Section III-A.2 and an estimate \( \hat{\nabla}_{\mu,n} \hat{\rho}^H_g(\theta) \) of the gradient \( \nabla \rho_g(\cdot) \) is formed as follows:
\[
\hat{\nabla}_{\mu,n} \hat{\rho}^H_g(\theta) = \frac{d}{n} \sum_{i=1}^{n} \hat{\rho}^H_g(\theta + \mu \psi_i) - \hat{\rho}^H_g(\theta - \mu \psi_i) \frac{1}{2\mu}, \tag{15}
\]

The update iteration in DRM-OffP-SF is as follows:
\[
\theta_{k+1} = \theta_k + \alpha \hat{\nabla}_{\mu,n} \hat{\rho}^H_g(\theta_k). \tag{16}
\]

The pseudocode of DRM-offP-SF algorithm is similar to Algorithm 1 but in each iteration, we get \( m \) episodes from policy \( b \). Then, we generate \( n \) DRM estimates \( \{ \hat{\rho}^H_g(\theta_k \pm \mu \psi_i), i \in \{ 1, \ldots, n \} \} \) using (14). We estimate the gradient using (15), and use the policy parameter update rule (16). The reader is referred to Algorithm 2 in Appendix VI.

IV. MAIN RESULTS

Our non-asymptotic analysis establishes a bound on the number of iterations of our proposed algorithms to find an \( \epsilon \)-stationary point of the DRM, which is defined below.

**Definition 1 (\( \epsilon \)-stationary point):** Let \( \theta_R \) be the output of an algorithm. Then, \( \theta_R \) is called an \( \epsilon \)-stationary point of problem (3), if \( \mathbb{E} \| \nabla \rho_g(\theta_R) \|^2 \leq \epsilon \).

The study of convergence of the policy gradient algorithms to an \( \epsilon \)-stationary point is common for non-asymptotic analysis in RL, since the objective is non-convex (cf. [24], [25]).

A. Non-asymptotic bounds for DRM-OffP-SF

We make the following assumptions to ensure the Lipschitzness, and smoothness of the DRM \( \rho_g \).

(A4): \( \exists M_d, M_h > 0 : \forall \theta \in \mathbb{R}^d, \forall \alpha \in A, s \in S, \| \nabla \log \pi_\theta(\alpha | s) \| \leq M_d \), and \( \| \nabla^2 \log \pi_\theta(\alpha | s) \| \leq M_h \.

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

An assumption like (A4) is common in the literature for the non-asymptotic analysis of policy gradient algorithms (cf. [26], [25]). 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 I. Since \( g(\cdot) \) is bounded by definition, we can see that any \( g(\cdot) \) whose second derivative is bounded, will have a bounded first derivative also.

Letting \( T \) denote the (random) episode length of a proper policy \( \pi_\theta \), we have \( \mathbb{E}|T| < \infty \) from (A1) This fact in conjunction with \( T \geq 0 \) implies
\[
\exists M_e > 0 : T \leq M_e \ a.s. \tag{17}
\]

The main result that establishes a non-asymptotic bound for DRM-OffP-SF is given below. This result is for a random
iterate \( \theta_R \), that is chosen uniformly at random from the policy parameters \( \{ \theta_1, \ldots, \theta_N \} \). Such a randomized stochastic gradient algorithm has been studied earlier in an stochastic optimization setting in [27].

**Theorem 1:** (DRM-OnP-SF) Assume \( \text{(A1)(A4)(A5)} \). Let \( \{ \theta_i, i = 1, \ldots, N \} \) be the policy parameters generated by DRM-OnP-SF, and let \( \theta_R \) be chosen uniformly at random from this set. Set \( \alpha = \frac{1}{N}, \mu = \frac{1}{N}, n = N \), and \( m > 0 \). Then

\[
\mathbb{E} \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] \leq \frac{2 \rho^* - \rho_g(\theta_0)}{\sqrt{N}} + \frac{d_2 L^2}{N} + \frac{2 \rho \mu}{N} \sqrt{N} + \frac{16 d^2 L^2 M^2 g}{mN}.
\]

The above, \( \rho^*_g = \max_{\theta \in \Theta} \rho_g(\theta) \), and \( M_r = \frac{m}{\sqrt{N}} \). The constants \( L_r = 2 M_r M_e (M_b M_{\theta} + M_r M_{\theta}) \), and \( L_\rho = 2 M_r M_r M_g M_d, \) with \( M_e, M_b, M_{\theta}, M_{\mu} \) as in (17). The constants \( M_{\theta}, M_{\mu}, M_d, \) and \( M_h \) are as in \( \text{(A4)(A5)} \).

**Remark 1:** The result above shows that after \( N \) iterations of [9], DRM-OnP-SF returns an iterate that satisfies \( \mathbb{E} \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] = O \left( \frac{1}{\sqrt{N}} \right) \). To put it differently, to find an \( \epsilon \)-stationary point of the DRM objective, an order \( O(\frac{1}{\epsilon^2}) \) iterations of DRM-OnP-SF are enough.

**Proof:** [Theorem 1] We provide a proof sketch below. For a detailed proof, the reader is referred to Appendix [IV].

The proof uses the following results related to our on-policy estimation scheme:

1. DRM and its gradient are Lipschitz, i.e., \( \forall \theta_1, \theta_2 \in \mathbb{R}^d \),

\[
\| \rho_g(\theta_1) - \rho_g(\theta_2) \| \leq L_\rho \| \theta_1 - \theta_2 \|,
\]

(18)

\[
\| \nabla \rho_g(\theta_1) - \nabla \rho_g(\theta_2) \| \leq L_\rho \| \theta_1 - \theta_2 \|.
\]

(19)

2. The DRM estimation error satisfies

\[
\mathbb{E} \left[ \| \rho_g(\theta) - \rho^*_g(\theta) \|^2 \right] \leq \frac{16 M^2 g^2}{m^2}.
\]

(20)

3. The bias of the DRM gradient estimate satisfies

\[
\mathbb{E} \left[ \| \hat{\nabla}_{\mu, \rho}^G(\theta) - \nabla \rho_g(\theta) \| \right] \leq \mu^2 d^2 L^2 \rho + \frac{4 d^2 L^2 \rho}{n} + \frac{16 d^2 M^2 g}{\mu^2 mn}.
\]

(21)

4. The variance of the DRM gradient estimate is bounded by

\[
\mathbb{E} \left[ \| \hat{\nabla}_{\mu, \rho}^G(\theta) \| \right] \leq \frac{2 d^2 L^2 \rho}{n} + \frac{16 d^2 M^2 g}{\mu^2 mn}.
\]

(22)

We now turn to proving the main result. Using the fundamental theorem of calculus, we obtain

\[
\rho_g(\theta_k) - \rho_g(\theta_{k+1}) = \left( \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \right)
\]

\[
+ \int_0^1 \left( \nabla \rho_g(\theta_{k+1} + \tau(\theta_k - \theta_{k+1}))- \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \right) d\tau
\]

\[
\leq \| \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \|
\]

\[
+ \int_0^1 \| \nabla \rho_g(\theta_{k+1} + \tau(\theta_k - \theta_{k+1}))- \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \| d\tau
\]

\[
\leq \left( \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \right) + L_\rho \| \theta_k - \theta_{k+1} \| \int_0^1 (1 - \tau) d\tau
\]

(23)

Rearranging and taking expectations on both sides of (23), we obtain

\[
\alpha \mathbb{E} \left[ \| \nabla \rho_g(\theta_k) \|^2 \right] \]

\[
\leq 2\mathbb{E} [\rho_g(\theta_{k+1}) - \rho_g(\theta_k)] + L_\rho \alpha^2 \mathbb{E} \left[ \| \hat{\nabla}_{\mu, \rho}^G(\theta_k) \| \right]
\]

\[
+ \alpha \mathbb{E} \left[ \| \nabla \rho_g(\theta_k) - \hat{\nabla}_{\mu, \rho}^G(\theta_k) \| \right].
\]

(24)

Using (21) and (22), we simplify (24) as

\[
\alpha \mathbb{E} \left[ \| \nabla \rho_g(\theta_k) \|^2 \right] \leq 2\mathbb{E} [\rho_g(\theta_{k+1}) - \rho_g(\theta_k)]
\]

\[
+ L_\rho \alpha^2 \left( \frac{2 d^2 L^2 \rho}{n} + \frac{16 d^2 M^2 g}{\mu^2 mn} \right)
\]

\[
+ \alpha \left( \frac{4 d^2 L^2 \rho}{n} + \mu^2 d^2 L^2 \rho + \frac{16 d^2 M^2 g}{\mu^2 mn} \right).
\]

(25)

Summing up (25) from \( k = 0, \ldots, N - 1 \), we obtain

\[
\alpha \sum_{k=0}^{N-1} \mathbb{E} \left[ \| \nabla \rho_g(\theta_k) \|^2 \right] \leq 2 \mathbb{E} [\rho_g(\Theta) - \rho_g(\Theta)]
\]

\[
+ N L_\rho \alpha^2 \left( \frac{2 d^2 L^2 \rho}{n} + \frac{16 d^2 M^2 g}{\mu^2 mn} \right)
\]

\[
+ N \alpha \left( \frac{4 d^2 L^2 \rho}{n} + \mu^2 d^2 L^2 \rho + \frac{16 d^2 M^2 g}{\mu^2 mn} \right).
\]

Since \( \theta_R \) is chosen uniformly at random from the policy iterations \( \{ \theta_1, \ldots, \theta_N \} \), we obtain

\[
\mathbb{E} \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] \leq \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E} \left[ \| \nabla \rho_g(\theta_k) \|^2 \right]
\]

\[
\leq \frac{2 (\rho^*_g - \rho_g(\theta_0))}{N \alpha} + L_\rho \alpha \left( \frac{2 d^2 L^2 \rho}{n} + \frac{16 d^2 M^2 g}{\mu^2 mn} \right)
\]

\[
+ \frac{4 d^2 L^2 \rho}{n} + \mu^2 d^2 L^2 \rho + \frac{16 d^2 M^2 g}{\mu^2 mn}
\]

\[
\leq \frac{2 (\rho^*_g - \rho_g(\theta_0))}{\sqrt{N}} + \frac{d^2 L^2 \rho}{\sqrt{N}} + \frac{2 d^2 L^2 \rho L^2}{N \sqrt{N}}.
\]

(19)
where last inequality follows since \( \alpha = \frac{1}{\sqrt{N}}, \mu = \frac{1}{\sqrt{N}}, \) and \( n = N \).

\[ \text{B. Non-asymptotic bounds for DRM-OffP-SF} \]

The main result that establishes a non-asymptotic bound for our algorithm DRM-OffP-SF is given below.

**Theorem 2:** (DRM-OffP-SF) Assume \((A1)-(A5)\). Let \((\theta_i, i = 1, \ldots, N)\) be the policy parameters generated by DRM-OffP-SF, and let \(\theta_R\) be chosen uniformly at random from this set. Set \( \alpha = \frac{1}{\sqrt{N}}, \mu = \frac{1}{\sqrt{N}}, \) \( n = N, \) and \( m > 0 \).

Then
\[
\begin{align*}
\mathbb{E} \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] & \leq \frac{2}{m} \left( \rho_g^* - \rho_g(\theta_0) \right) + \frac{d^2 L_g^2}{N} + \frac{2d^2 L_g^2}{N} + \frac{2d^2 L_g^2}{N} \\
& + \frac{16d^2 L_g^2 M_g^2 M_g^2}{mN} + \frac{4d^2 L_g^2}{N} + \frac{16d^2 M_g^2 M_g^2}{mN} \\
& + \frac{16d^2 M_g^2 M_g^2}{mN}.
\end{align*}
\]

In the above, \( \rho_g^* = \max_{\theta \in \mathbb{R}} \rho_g(\theta) \), and \( M_g = \frac{\max_{\theta \in \mathbb{R}} g(\theta)}{\rho_g(\theta)} \). The constants \( L_g = 2M_gM_gM_gM_g, \) \( M_g = M_g(M_g + M_g) \), and \( L_g = 2M_gM_gM_gM_g^2 \), with \( M_g \) as in (17). The constants \( M_g, M_g, M_g, \) and \( M_g \) are as in (A4)-(A5) while \( M_g \) is an uniform upper bound on the importance sampling ratio \( \psi(\theta) \).

**Proof:** (Sketch) For establishing the main result, we follow the technique employed in the proof of Theorem [1] and use the following results related to our off-policy estimation scheme in place of their on-policy counterparts:

1. The estimation error of the DRM satisfies
   \[ \mathbb{E} \left[ \| \rho_g(\theta) - \hat{\rho}_g(\theta) \|^2 \right] \leq \frac{16\rho^2 M_g^4 M_g^2}{m} \]
2. The bias of the DRM gradient estimate is bounded by
   \[ \mathbb{E} \left[ \left\| \nabla \rho_g(\theta) - \nabla \hat{\rho}_g(\theta) \right\|^2 \right] \leq \frac{4d^2 L_g^2}{m} + \frac{16d^2 M_g^2 M_g^2}{m} \]
3. The variance of the DRM gradient estimate is bounded by
   \[ \mathbb{E} \left[ \left\| \nabla \rho_g(\theta) - \nabla \hat{\rho}_g(\theta) \right\|^2 \right] \leq \frac{4d^2 L_g^2}{m} + \frac{16d^2 M_g^2 M_g^2}{m} \]

The reader is referred to Appendix [V] for the detailed proof.

\[ \text{V. CONCLUSIONS} \]

We proposed DRM-based approximate gradient algorithms for risk sensitive RL control. We employed SF-based gradient estimation schemes in on-policy as well as off-policy RL settings, and provided non-asymptotic bounds that establish convergence to an approximate stationary point of the DRM.

As future work, it would be interesting to study DRM optimization in a risk-sensitive RL setting with feature-based representations, and function approximation.

**REFERENCES**

[1] P. Artzner, F. Delbaen, J. Eber, and D. Heath, “Coherent measures of risk,” *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.
[2] S. Wang, “A risk measure that goes beyond coherence,” 2002.
[3] R. T. Rockafellar and S. Uryasev, “Optimization of conditional value-at-risk,” *Journal of risk*, vol. 2, pp. 21–42, 2000.
[4] A. Tversky and D. Kahneman, “Advances in prospect theory: Cumulative representation of uncertainty,” *J. Risk Uncertain.*, vol. 5, 1992.
[5] D. Denneberg, “Distorted probabilities and insurance premiums,” *Methods of Operations Research*, vol. 63, no. 3, pp. 3–5, 1990.
[6] S. S. Wang, V. R. Young, and H. H. Panjer, “Axiomatic characterization of insurance prices,” *Insur. Math. Econ.*, vol. 21, no. 2, 1997.
[7] J. Wirch and M. Hardy, “Distortion risk measures: Coherence and stochastic dominance,” *Insur. Math. Econ.*, vol. 32, pp. 168–168, 2003.
[8] H. Gzyl and S. Mayoral, “On a relationship between distorted and spectral risk measures,” 2006.
[9] B. Jones and R. Zitikis, “Empirical estimation of risk measures and related quantities,” *North American Actuarial Journal*, vol. 7, 2003.
[10] S. Wang, “Premium calculation by transforming the layer premium density,” *ASTIN Bulletin*, vol. 26, no. 1, pp. 71–92, 1996.
[11] V. Katkovnik and Y. Kulchitsky, “Convergence of a class of random search algorithms,” *Autom. Remote. Control.*, vol. 33, 1972.
[12] Y. Nesterov and V. Spokoiny, “Random gradient-free minimization of convex functions,” *Found. Comput. Math.*, vol. 17, pp. 527–566, 2017.
[13] S. Bhatnagar, H. Prasad, and L. A. Prashanth, “Stochastic recursive algorithms for optimization. simultaneous perturbation methods,” *Lecture Notes in Control and Inform. Sci.*, vol. 434, 2013.
[14] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, “Policy gradient for coherent risk measures,” in *Adv. Neural Inf. Process. Syst.*, 2015.
[15] L. A. Prashanth and M. Fu, “Risk-sensitive reinforcement learning,” *arXiv preprint arXiv:1510.09126*, 2021.
[16] P. Glynn, Y. Peng, M. Fu, and J. Hu, “Computing sensitivities for distortion risk measures,” *INFORMS J. Comp.*, 2021.
[17] N. Vijayan and L. A. Prashanth, “Policy gradient methods for distortion risk measures,” *arXiv preprint arXiv:2107.04422*, 2021.
[18] L. A. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvari, “Cumulative prospect theory meets reinforcement learning: Prediction and control,” in *ICML*, vol. 48, 2016, pp. 1406–1415.
[19] K. Dowd and B. Blake, “After var: The theory, estimation, and insurance applications of quantile-based risk measures,” *The Journal of Risk and Insurance*, vol. 73, no. 2, pp. 193–229, 2006.
[20] A. Balbás, J. Garrido, and S. Mayoral, “Properties of distortion risk measures,” *Methodology and Computing in Applied Probability*, vol. 11, no. 3, pp. 385–399, 2009.
[21] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, 1st ed. Athena Scientific, 1996.
[22] 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*, 2005, pp. 385–394.
[23] 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.*, vol. 21, 2009.
[24] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli, “Stochastic variance-reduced policy gradient,” in *ICML*, 2018.
[25] Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi, “Hessian aided proximal policy optimization,” *Adv. Neural Inf. Process. Syst.*, vol. 63, no. 3, pp. 3–5, 1990.
[26] S. Ghadimi and G. Lan, “Stochastic first- and zeroth-order methods for nonconvex stochastic programming,” *SIAM J. Optim.*, vol. 23, 2013.
[27] J. Kim, “Bias correction for estimated distortion risk measure using the bootstrap,” *Insur.: Math. Econ.*, vol. 47, pp. 198–205, 2010.
[28] Y. E. Nesterov, *Introductory Lectures on Convex Optimization - A Basic Course*, ser. Applied Optimization, 2004, vol. 87.
[29] X. Gao, B. Jiang, and S. Zhang, “On the information-adaptive variants of the admm: An iteration complexity perspective,” *Insur. Math. Econ.*, vol. 76, no. 1, pp. 327–363, 2018.
[30] O. Shamir, “An optimal algorithm for bandit and zero-order convex approximation,” in *ICML*, vol. 18, no. 1, pp. 1703–1713, 2017.
The following lemma estimate the DRM in an off-policy RL setting.

**Lemma 1:** $\hat{\rho}_g^C(\theta) = \sum_{i=1}^m R_{(i)}^\theta \left( g \left( 1 - \frac{i-1}{m} \right) - g \left( 1 - \frac{i}{m} \right) \right)$.

*Proof:* Our proof follows the technique from [28]. We rewrite (3) as

$$G^m_{R^\theta}(x) = \begin{cases} 0, & \text{if } x < R_{(1)}^\theta \\ \frac{i}{m}, & \text{if } R_{(i)}^\theta \leq x < R_{(i+1)}^\theta, \ i \in \{1, \cdots, m-1\} \\ 1, & \text{if } x \geq R_{(m)}^\theta, \end{cases} \quad (26)$$

where $R_{(i)}^\theta$ is the $i^{th}$ smallest order statistic from the samples $R_1^\theta, \cdots R_m^\theta$.

We assume without loss of generality that $R_{(j)}^\theta < R_{(j+1)}^\theta$, and obtain,

$$\hat{\rho}_g^C(\theta) = \int_0^{R_{(1)}^\theta} (g(1-G^m_{R^\theta}(x))-1)dx + \int_{R_{(1)}^\theta}^{R_{(j)}^\theta} g(1-G^m_{R^\theta}(x))dx + \int_{R_{(j)}^\theta}^{R_{(j+1)}^\theta} (g(1-G^m_{R^\theta}(x))-1)dx + \int_{R_{(j+1)}^\theta}^{R_{(m)}^\theta} g(1-G^m_{R^\theta}(x))dx$$

$$= \sum_{i=2}^j \int_{R_{(i-1)}^\theta}^{R_{(i)}^\theta} \left( g \left( 1 - \frac{i-1}{m} \right) - 1 \right)dx + \int_{R_{(j)}^\theta}^{R_{(j+1)}^\theta} \left( g \left( 1 - \frac{j}{m} \right) - 1 \right)dx + \int_{0}^{R_{(j+1)}^\theta} g \left( 1 - \frac{j}{m} \right)dx + \sum_{i=j+1}^{m-1} \int_{R_{(i)}^\theta}^{R_{(i+1)}^\theta} g \left( 1 - \frac{i}{m} \right)dx$$

$$= \sum_{i=2}^j \left( R_{(i)}^\theta - R_{(i-1)}^\theta \right) \left( g \left( 1 - \frac{i-1}{m} \right) - 1 \right) - \left( R_{(j)}^\theta \right) \left( g \left( 1 - \frac{j}{m} \right) - 1 \right) + R_{(j+1)}^\theta \left( g \left( 1 - \frac{j}{m} \right) \right)$$

$$+ \sum_{i=j+1}^{m-1} \left( R_{(i+1)}^\theta - R_{(i)}^\theta \right) g \left( 1 - \frac{i}{m} \right)$$

$$= \sum_{i=2}^j \left( R_{(i)}^\theta - R_{(i-1)}^\theta \right) g \left( 1 - \frac{i-1}{m} \right) + R_{(1)}^\theta + \sum_{i=j}^{m-1} \left( R_{(i+1)}^\theta - R_{(i)}^\theta \right) g \left( 1 - \frac{i}{m} \right)$$

$$= \sum_{i=1}^m R_{(i)}^\theta g \left( 1 - \frac{i-1}{m} \right) - \sum_{i=1}^{m-1} R_{(i)}^\theta g \left( 1 - \frac{i}{m} \right)$$

$$= \sum_{i=1}^m R_{(i)}^\theta \left( g \left( 1 - \frac{i-1}{m} \right) - g \left( 1 - \frac{i}{m} \right) \right)$$

The following lemma estimate the DRM in an off-policy RL setting.

**Lemma 2:** $\hat{\rho}_g^H(\theta) = R_{(1)}^\theta + \sum_{i=2}^m R_{(i)}^\theta \left( 1 - \min \left\{ 1, \frac{i}{m} \sum_{k=1}^{i-1} \psi_{(k)}^\theta \right\} \right) - \sum_{i=1}^{m-1} R_{(i)}^\theta \left( 1 - \min \left\{ 1, \frac{1}{m} \sum_{k=1}^{i} \psi_{(k)}^\theta \right\} \right)$.

*Proof:* We rewrite (11) as

$$H^m_{R^\theta}(x) = \begin{cases} 0, & \text{if } x < R_{(1)}^\theta \\ \min \{1, \frac{1}{m} \sum_{j=1}^i \psi_{(j)}^\theta \}, & \text{if } R_{(i)}^\theta \leq x < R_{(i+1)}^\theta, \ i \in \{1, \cdots, m-1\} \\ 1, & \text{if } x \geq R_{(m)}^\theta, \end{cases} \quad (27)$$

where $R_{(i)}^\theta$ is the $i^{th}$ smallest order statistic from the samples $R_1^\theta, \cdots R_m^\theta$, and $\psi_{(i)}^\theta$ is the importance sampling ratio of $R_{(i)}^\theta$. 
We assume without loss of generality that $R_{(j)}^b < 0 < R_{(j+1)}^b$, and obtain,
\[
\hat{\psi}_A^b(\theta) = \int_{-M_r}^0 (g(1 - H_{R^b}^m(x)) - 1) dx + \sum_{i=2}^j \int_{R_{(i)}^b}^{M_r} g(1 - H_{R^b}^m(x)) dx
\]
\[
R_{(j)}^b = \int_{-M_r}^0 (g(1 - H_{R^b}^m(x)) - 1) dx + \sum_{i=2}^j \int_{R_{(i)}^b}^{M_r} g(1 - H_{R^b}^m(x)) dx + \int_{R_{(j)}^b}^{R_{(j+1)}^b} g(1 - H_{R^b}^m(x)) dx + \int_{R_{(j+1)}^b}^{0} g(1 - H_{R^b}^m(x)) dx
\]
\[
\text{Lemma 3: } \forall \theta \in \Theta, g(\cdot) = \frac{1}{m} \sum_{k=1}^{i-1} \psi(\theta(k)) - 1 - R_{(j)}^b \left( 1 - \min \left\{ \frac{1}{m} \sum_{k=1}^{i} \psi(\theta(k)) \right\} \right)
\]
\[
\text{Lemma 4: } \forall \theta \in \Theta, g(\cdot) = \frac{1}{m} \sum_{k=1}^{i-1} \psi(\theta(k)) - 1 - R_{(j)}^b \left( 1 - \min \left\{ \frac{1}{m} \sum_{k=1}^{i} \psi(\theta(k)) \right\} \right)
\]

\[\text{Proof:}\]
\[
\text{APPENDIX II}
\]
\[\text{LIPSCHITZ PROPERTIES OF THE DRM AND ITS GRADIENT}\]

A. Results related to the distortion function

The following two lemma establish Lipschitzness of the $g(\cdot)$, and $g'(\cdot)$. We require this result to establish the smoothness of the DRM.

**Lemma 3:** $\forall t, t' \in (0,1), |g(t) - g(t')| \leq M_g |t - t'|$, and $|g'(t) - g'(t')| \leq M_g' |t - t'|$.

**Proof:** Using mean value theorem, we obtain $g(t) - g(t') = g'(\tilde{t})(t - t')$, where $\tilde{t} \in (t, t')$. From $[A5]$, we obtain $|g'(\tilde{t})| \leq M_g', \forall \tilde{t} \in (0, 1)$. Hence, $|g(t) - g(t')| \leq M_g |t - t'| \forall t, t' \in (0, 1)$.

Similarly, using mean value theorem, we obtain $g'(t) - g'(t') = g''(\tilde{t})(t - t')$, where $\tilde{t} \in (t, t')$. From $[A5]$, we obtain $|g''(\tilde{t})| \leq M_{g''}, \forall \tilde{t} \in (0, 1)$. Hence, $|g'(t) - g'(t')| \leq M_{g''} |t - t'| \forall t, t' \in (0, 1)$.

B. Lipschitz properties of the CDF

The following two lemmas establish an upper bound for the gradient and the Hessian of the CDF. These lemmas are similar to lemmas in [17]. For the sake of completeness, we provide the detailed proof.

**Lemma 4:** $\forall x \in [-M_r, M_r], \nabla F_{R^b}(x) = \mathbb{E} \left[ \mathbb{I} \{ R^b \leq x \} \sum_{i=0}^{T-1} \nabla \log \pi(\theta(A_t|S_t)) \right]$, and

\[
\nabla^2 F_{R^b}(x) = \mathbb{E} \left[ \mathbb{I} \{ R^b \leq x \} \left( \sum_{t=0}^{T-1} \nabla^2 \log \pi(\theta(A_t|S_t)) + \left[ \sum_{t=0}^{T-1} \nabla \log \pi(\theta(A_t|S_t)) \right] \cdot \left[ \sum_{t=0}^{T-1} \nabla \log \pi(\theta(A_t|S_t)) \right] \right) \right].
\]

**Proof:** Let $\Omega$ denote the set of all sample episodes. For any episode $\omega \in \Omega$, we denote by $T(\omega)$, its length, and $S_t(\omega)$ and $A_t(\omega)$, the state and action at time $t \in \{0, 1, 2, \cdots\}$ respectively. Let $R(\omega) = \sum_{t=0}^{T(\omega)-1} \gamma^t r(S_t(\omega), A_t(\omega), S_{t+1}(\omega))$ be the cumulative discounted reward of the episode $\omega$. Let
\[ \mathbb{P}_\theta(\omega) = \prod_{t=0}^{T(\omega)-1} \pi\theta(A_t(\omega) \mid S_t(\omega))p(S_{t+1}(\omega), S_t(\omega), A_t(\omega)). \]

From \( \nabla \frac{\partial \mathbb{P}_\theta(\omega)}{\partial \mathbb{P}_\theta(\omega)} \) = \( \sum_{t=0}^{T(\omega)-1} \nabla \log \pi\theta(A_t(\omega) \mid S_t(\omega)) \), we obtain
\[
\nabla F_{R^\theta}(x) = \nabla \mathbb{E}[\mathbb{1}\{R^\theta \leq x\}] = \nabla \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \mathbb{P}_\theta(\omega)
\]
\[
= \sum_{\omega \in \Omega} \nabla (\mathbb{1}\{R(\omega) \leq x\} \mathbb{P}_\theta(\omega))
\]
\[
= \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \nabla \mathbb{P}_\theta(\omega)
\]
\[
= \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \frac{\nabla \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} \mathbb{P}_\theta(\omega) = \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \sum_{t=0}^{T(\omega)-1} \nabla \log \pi\theta(A_t(\omega) \mid S_t(\omega)) \mathbb{P}_\theta(\omega)
\]
\[
= \mathbb{E}
\left[
\left.
\left[
\nabla \log \pi\theta(A_t \mid S_t)
\right]^{T}
\right|_1
\right].
\]

In the above, the equality in (28) follows by an application of the dominated convergence theorem to interchange the differentiation and the expectation operation. The aforementioned application is allowed since (i) \( \Omega \) is finite and the underlying measure is bounded, as we consider an MDP where the state and actions spaces are finite, and the policies are proper, (ii) \( \nabla \log \pi\theta(A_t \mid S_t) \) is bounded from [A4]. The equality in (29) follows, since for a given episode \( \omega \), the cumulative reward \( R(\omega) \) does not depend on \( \theta \).

Similarly,
\[
\nabla^2 F_{R^\theta}(x) = \mathbb{E}
\left[
\left[
\nabla \log \pi\theta(A_t \mid S_t)
\right]^{T}
\right|_1
\right].
\]

**Lemma 5:** \( \forall x \in [-M_r, M_r], \| \nabla F_{R^\theta}(x) \| \leq M_c M_d, \) and \( \| \nabla^2 F_{R^\theta}(x) \| \leq M_c M_h + M_c^2 M_d^2. \)

Proof: Recall that \( T \) denotes the (random) episode length of a proper policy \( \pi\theta. \) From (A1) we infer that \( \mathbb{E}[T] < \infty. \) This fact in conjunction with \( T \geq 0 \) implies the following bound:
\[
\exists M_c > 0 \text{ s.t. } T \leq M_c \text{ a.s.}
\]

From (A4) and (30), for any \( x \in [-M_r, M_r], \) we have
\[
\| \mathbb{E} \left[ \nabla \log \pi\theta(A_t \mid S_t) \right] \| \leq M_c M_d \text{ a.s.}
\]

and
\[
\left\| \mathbb{E} \left[ \nabla \log \pi\theta(A_t \mid S_t) \right] \right\| \leq M_c M_h + M_c^2 M_d^2 \text{ a.s.}
\]

From Lemma 4, for any \( x \in [-M_r, M_r], \) we have
\[
\| \nabla F_{R^\theta}(x) \| \leq \mathbb{E} \left[ \| \mathbb{E} \left[ \nabla \log \pi\theta(A_t \mid S_t) \right] \| \right] \leq M_c M_d,
\]

and
\[
\| \nabla^2 F_{R^\theta}(x) \| \leq \mathbb{E} \left[ \| \nabla \log \pi\theta(A_t \mid S_t) \| \right] \leq M_c M_h + M_c^2 M_d^2,
\]

where these inequalities follows from (31), (32), and the assumption that the state and action spaces are finite.
The following lemma establishes Lipschitzness of the CDF and its gradient.

**Lemma 6:** \( \forall x \in [-M_r, M_r] \), \(| F_{R_{F_1}}(x) - F_{R_{F_2}}(x) | \leq M_r M_d \| \theta_1 - \theta_2 \|, \) and \( \| \nabla F_{R_{F_1}}(x) - \nabla F_{R_{F_2}}(x) \| \leq (M_r M_h + M^2_r M^3_d) \| \theta_1 - \theta_2 \| \)

**Proof:** The result follows by Lemma 5 and Lemma 1.2.2 in [29].

---

**C. Gradient of the DRM**

The following lemma derives an expression for the gradient of the DRM. This lemma is similar to Theorem 1 in [17]. For the sake of completeness, we provide the detailed proof.

**Lemma 7:** \( \nabla \rho_g(\theta) = - \int_{-M_r}^{M_r} g'(1 - F_{R_{F}}(x)) \nabla F_{R_{F}}(x) dx \).

**Proof:** Notice that

\[
\nabla \rho_g(\theta) = \nabla \int_{-M_r}^{M_r} (g(1-F_{R_{F}}(x)) - 1) \, dx + \int_{-M_r}^{M_r} g(1-F_{R_{F}}(x)) \, dx \\
= \int_{-M_r}^{M_r} \nabla (g(1-F_{R_{F}}(x)) - 1) \, dx + \int_{0}^{M_r} \nabla g(1-F_{R_{F}}(x)) \, dx \\
= - \int_{-M_r}^{M_r} g'(1 - F_{R_{F}}(x)) \nabla F_{R_{F}}(x) dx.
\]

In the above, the equality in (35) follows by an application of the dominated convergence theorem to interchange the differentiation and the expectation operation. The aforementioned application is allowed since (i) \( \rho_g(\theta) \) is finite for any \( \theta \in \mathbb{R}^d \); (ii) \( |g'(\cdot)| \leq M_g' \) from (A5) and \( \nabla F_{R_{F}}(\cdot) \) is bounded from (33). The bounds on \( g' \) and \( \nabla F_{R_{F}} \) imply \( \int_{-M_r}^{M_r} \| g'(1-F_{R_{F}}(x)) \nabla F_{R_{F}}(x) \| dx \leq 2M_r M_g'M_c M_d \).

---

**D. Lipschitz properties of the DRM and its gradient**

The following two lemmas establish the Lipschitzness of the DRM and its gradient.

**Lemma 8:** \( \forall \theta_1, \theta_2 \in \mathbb{R}^d \), \( | \rho_g(\theta_1) - \rho_g(\theta_2) | \leq L_\rho \| \theta_1 - \theta_2 \| \), where \( L_\rho = 2M_r M_g'M_c M_d \).

**Proof:**

\[
| \rho_g(\theta_1) - \rho_g(\theta_2) | \leq \int_{-M_r}^{M_r} | g(1-F_{R_{F_1}}(x)) - g(1-F_{R_{F_2}}(x)) | dx \\
\leq M_g' \int_{-M_r}^{M_r} | F_{R_{F_1}}(x) - F_{R_{F_2}}(x) | dx \leq 2M_r M_g'M_c M_d \| \theta_1 - \theta_2 \| \) (from lemmas 5 and 6).

The result follows since \( L_\rho = 2M_r M_g'M_c M_d \).

**Lemma 9:** \( \forall \theta_1, \theta_2 \in \mathbb{R}^d \), \( \| \nabla \rho_g(\theta_1) - \nabla \rho_g(\theta_2) \| \leq L_{\rho'} \| \theta_1 - \theta_2 \| \), where \( L_{\rho'} = 2M_r M_c (M_h M_g' + M_c M^2_d (M_g' + M_g'')) \).

**Proof:** From Lemma 7 we obtain

\[
\| \nabla \rho_g(\theta_1) - \nabla \rho_g(\theta_2) \| \\
\leq \int_{-M_r}^{M_r} \| g'(1-F_{R_{F_1}}(x)) \nabla F_{R_{F_1}}(x) - g'(1-F_{R_{F_2}}(x)) \nabla F_{R_{F_2}}(x) \| dx \\
\leq \int_{-M_r}^{M_r} \| g'(1-F_{R_{F_1}}(x)) \nabla F_{R_{F_1}}(x) - g'(1-F_{R_{F_2}}(x)) \nabla F_{R_{F_2}}(x) + g'(1-F_{R_{F_1}}(x)) \nabla F_{R_{F_2}}(x) \| dx \\
\leq \int_{-M_r}^{M_r} \| g'(1-F_{R_{F_1}}(x)) \| \| \nabla F_{R_{F_1}}(x) - \nabla F_{R_{F_2}}(x) \| + \| \nabla F_{R_{F_2}}(x) \| \| g'(1-F_{R_{F_1}}(x)) - g'(1-F_{R_{F_2}}(x)) \| dx \\
\leq \int_{-M_r}^{M_r} M_g' \| \nabla F_{R_{F_1}}(x) - \nabla F_{R_{F_2}}(x) \| dx + M_c M_d M_g' \| F_{R_{F_1}}(x) - F_{R_{F_2}}(x) \| dx \) (from A5 and Lemmas 5) \\
\leq \int_{-M_r}^{M_r} M_g' (M_c M_h + M^2_c M^2_d) \| \theta_1 - \theta_2 \| + M^2_c M^2_d M_g' \| \theta_1 - \theta_2 \| dx \) (from Lemma 6) \\
\leq 2M_r M_c (M_h M_g' + M_c M^2_d (M_g' + M_g'')) \| \theta_1 - \theta_2 \| .
\]

The result follows since \( L_{\rho'} = 2M_r M_c (M_h M_g' + M_c M^2_d (M_g' + M_g'')) \).
APPENDIX III
SF-BASED GRADIENT ESTIMATE

Some of the lemmas below use the following SF-based gradient estimate of \( \rho_g(\theta) \):
\[
\hat{\nabla}_{\mu,n} \rho_g(\theta) = \frac{d}{n} \sum_{i=1}^{n} \rho_g(\theta + \mu v_i) - \rho_g(\theta - \mu v_i) v_i, \tag{36}
\]

The difference between the equation (36) above and the one defined in (8) is that (36) uses the risk measure \( \rho \), while (8) uses the estimate \( \hat{\rho}^G(\cdot) \).

The following lemmas establish some results related to the SF-based gradient estimate.

\textbf{Lemma 10:} \( \| \nabla_{\rho_{g,\mu}}(\theta) - \nabla_{\hat{\rho}_g(\theta)} \| \leq \frac{\mu d L_{\rho}}{2} \).

\textit{Proof:} The result follows from [30, Proposition 7.5] along with Lemma 9. \( \blacksquare \)

\textbf{Lemma 11:} \( \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu,n} \rho_g(\theta) \right\|^2 \right] \leq \frac{d^2 L^2_\rho}{n} \).

\textit{Proof:} Since \( v_1 \in \mathbb{S}^{d-1}, \|v\| = 1 \), from (36), we have
\[
\mathbb{E}_{v_{1:n}} \left[ \left\| \hat{\nabla}_{\mu,n} \rho_g(\theta) \right\|^2 \right] \leq \frac{d^2}{4 \mu^2 n^2} \sum_{i=1}^{n} \mathbb{E}_v \left[ \left\| (\rho_g(\theta + \mu v) - \rho_g(\theta - \mu v)) v \right\|^2 \right]
\]
(since \( v_{1:n} \) are i.i.d zero r.v.s, and \( \rho_g(\cdot) \) is bounded)
\[
\leq \frac{d^2}{4 \mu^2 n} \mathbb{E}_v \left[ \left\| (\rho_g(\theta + \mu v) - \rho_g(\theta - \mu v)) \|v\|^2 \right\| \right]
\leq \frac{d^2 L^2_\rho}{n} \|v\|^4 \tag{from Lemma 8}.
\]

Finally, \( \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu,n} \rho_g(\theta) \right\|^2 \right] = \mathbb{E} \left[ \mathbb{E}_{v_{1:n}} \left[ \left\| \hat{\nabla}_{\mu,n} \rho_g(\theta) \right\|^2 \right] \right] \leq \frac{d^2 L^2_\rho}{n} \).

\textbf{Lemma 12:} \( \mathbb{E} \left[ \hat{\nabla}_{\mu,n} \rho_g(\theta) \right| \theta = \nabla_{\rho_{g,\mu}}(\theta). \)

\textit{Proof:} We follow the technique from [31]. Since \( v_{1:n} \) are i.i.d r.v.s, and have symmetric distribution around the origin, we obtain
\[
\mathbb{E} \left[ \hat{\nabla}_{\mu,n} \rho_g(\theta) \right| \theta = \mathbb{E}_{v_{1:n}} \left[ \hat{\nabla}_{\mu,n} \rho_g(\theta) \right] = \frac{d}{2 \mu n} \sum_{i=1}^{n} \mathbb{E}_v \left[ (\rho_g(\theta + \mu v) - \rho_g(\theta - \mu v)) v \right]
= \frac{d}{2 \mu} \left( \mathbb{E}_v [\rho_g(\theta + \mu v)v] + \mathbb{E}_v [\rho_g(\theta + \mu(-v))(-v)] \right) = \frac{d}{2 \mu} \mathbb{E}_v [\rho_g(\theta + \mu v)v] = \nabla_{\rho_{g,\mu}}(\theta),
\]
where last equality follows from [22, Lemma 2.1]. \( \blacksquare \)

APPENDIX IV
CONVERGENCE ANALYSIS: DRM-OnP-SF

A. The estimation error of the DRM

In the following lemma, we bound the estimation error of the DRM.

\textbf{Lemma 13:} \( \mathbb{E} \left[ \left| \rho_g(\theta) - \hat{\rho}^G(\theta) \right|^2 \right] \leq \frac{16 M^2 R^2}{m}. \)

\textit{Proof:} Since \( \forall x \in [-M, M], |\{R^o \leq x\}| \leq 1 \ a.s., \) using Hoeffdng’s inequality, we obtain \( \forall x \in [-M, M], \)
\[
\mathbb{P} \left( \left| G^m_{R^o}(x) - F_{R^o}(x) \right| > c \right) \leq 2 \exp \left( -\frac{m c^2}{2} \right), \quad \text{and}
\]
\[
\mathbb{E} \left[ \left| G^m_{R^o}(x) - F_{R^o}(x) \right|^2 \right] = \int_0^\infty \mathbb{P} \left( \left| G^m_{R^o}(x) - F_{R^o}(x) \right| > \sqrt{c} \right) dc \leq \int_0^\infty 2 \exp \left( -\frac{m c^2}{2} \right) dc = \frac{4}{m}. \tag{37}
\]

Now,
\[
\mathbb{E} \left[ \left| \rho_g(\theta) - \hat{\rho}^G(\theta) \right|^2 \right] = \mathbb{E} \left[ \left\| \int_{-M}^{M} (g(1 - F_{R^o}(x)) - g(1 - G^m_{R^o}(x))) dx \right\|^2 \right]
\leq 2 M \mathbb{E} \left[ \int_{-M}^{M} \left| (g(1 - F_{R^o}(x)) - g(1 - G^m_{R^o}(x))) \right|^2 dx \right] \quad \text{(from Cauchy-Schwarz inequality)}
\leq 2 M \int_{-M}^{M} \mathbb{E} \left[ \left| (g(1 - F_{R^o}(x)) - g(1 - G^m_{R^o}(x))) \right|^2 dx \right] \quad \text{(from Fubini's theorem)}
\]
where the last inequality follows from \([37]\).
C. Non-asymptotic bound for DRM-OnP-SF

Finally, we provide a non-asymptotic bound for DRM-OnP-SF.

**Proof:** (Theorem 11 DRM-OnP-SF) Using the fundamental theorem of calculus, we obtain

\[
\rho_g(\theta_k) - \rho_g(\theta_{k+1}) = \langle \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \rangle + \int_0^1 \langle \nabla \rho_g(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})), \theta_k - \theta_{k+1} \rangle \, d\tau
\]

\[
\leq \langle \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \rangle + \int_0^1 \|\nabla \rho_g(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})) - \nabla \rho_g(\theta_k)\| \, d\tau
\]

\[
\leq \langle \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \rangle + L_{\rho'} \|\theta_k - \theta_{k+1}\|^2 \int_0^1 (1 - \tau) \, d\tau \quad \text{(from Lemma 9)}
\]

\[
= \langle \nabla \rho_g(\theta_k), \theta_k - \theta_{k+1} \rangle + \frac{L_{\rho'}}{2} \|\theta_k - \theta_{k+1}\|^2
\]

\[
= \alpha \left( \langle \nabla \rho_g(\theta_k), \nabla \rho_g(\theta_k) \rangle - \nabla \rho_g(\theta_k) \right) - \alpha \|\nabla \rho_g(\theta_k)\|^2 + \frac{L_{\rho'}}{2} \alpha^2 \|\nabla \rho_g(\theta_k)\|^2
\]

\[
\leq \alpha \left( \langle \nabla \rho_g(\theta_k), \nabla \rho_g(\theta_k) \rangle - \nabla \rho_g(\theta_k) \right) - \alpha \|\nabla \rho_g(\theta_k)\|^2 + \frac{L_{\rho'}}{2} \alpha^2 \|\nabla \rho_g(\theta_k)\|^2
\]

\[
= \frac{\alpha}{2} \left( \|\nabla \rho_g(\theta_k)\|^2 - \|\nabla \rho_g(\theta_k)\|^2 \right) + \frac{L_{\rho'}}{2} \alpha^2 \|\nabla \rho_g(\theta_k)\|^2.
\]

Rearranging and taking expectations on both sides of (40), we obtain

\[
\alpha \mathbb{E} \|\nabla \rho_g(\theta_k)\|^2 \leq 2 \mathbb{E} \rho_g(\theta_{k+1}) - \rho_g(\theta_k) \quad \text{for } k = 0, \ldots, N - 1,
\]

where the last inequality follows from lemmas 15 and 16.

Summing up (41) from \( k = 0, \ldots, N - 1 \), we obtain

\[
\alpha \sum_{k=0}^{N-1} \mathbb{E} \|\nabla \rho_g(\theta_k)\|^2 \leq 2 \mathbb{E} \rho_g(\theta_N) - \rho_g(\theta_0) + N L_{\rho'} \alpha^2 \left( \frac{2d^2 L^2 }{n} + \frac{16d^2 M_2^2 M_{g'}^2}{\mu^2 mn} \right) + N \alpha \left( \frac{4d^2 L^2 }{n} + \frac{2d^2 L^2 }{n} + \frac{16d^2 M_2^2 M_{g'}^2}{\mu^2 mn} \right).
\]

Since \( \theta_R \) is chosen uniformly at random from the policy iterates \( \{\theta_1, \ldots, \theta_N\} \), we obtain

\[
\mathbb{E} \|\nabla \rho_g(\theta_R)\|^2 \leq \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E} \|\nabla \rho_g(\theta_k)\|^2 \leq 2 \frac{\rho_g(\theta_0) - \rho_g(\theta_N)}{N \alpha} + L_{\rho'} \alpha \left( \frac{2d^2 L^2 }{n} + \frac{16d^2 M_2^2 M_{g'}^2}{\mu^2 mn} \right) + \frac{4d^2 L^2 }{n} + \frac{d^2 L^2 }{n} + \frac{16d^2 M_2^2 M_{g'}^2}{\mu^2 mn}.
\]

Since \( \alpha = \frac{1}{\sqrt{N}}, \mu = \frac{1}{\sqrt{N}} \), and \( n = N \), we obtain

\[
\mathbb{E} \|\nabla \rho_g(\theta_R)\|^2 \leq \frac{2 \rho_g(\theta_0) - \rho_g(\theta_N)}{\sqrt{N}} + \frac{d^2 L^2 }{\sqrt{N}} + \frac{2d^2 L^2 }{N \sqrt{N}} + \frac{16d^2 L^2 }{mN} + \frac{4d^2 L^2 }{N^2} + \frac{16d^2 M_2^2 M_{g'}^2}{m \sqrt{N}}.
\]

**APPENDIX V**

**CONVERGENCE ANALYSIS: DRM-OffP-SF**

A. Importance sampling ratio

The following Lemma establish an upper bound on the importance sampling ratio.

**Lemma 17:** For any episode generated using \( b \), the importance sampling ratio \( \psi^b \leq M_s, \forall \theta \in \mathbb{R}^d \) a.s.

**Proof:** From (A3) and (A4) we obtain \( \forall \theta \in \mathbb{R}^d, \pi_\theta(\alpha|s) > 0 \) and \( b(\alpha|s) > 0 \), \( \forall \alpha \in \mathcal{A}, \forall s \in \mathcal{S} \). From (A2) we obtain that the episode length is bounded for \( b \). So the importance sampling ratio \( \psi^b \) is bounded for any episode. Hence WLOG, we say \( \psi^b \leq M_s, \forall \theta \in \mathbb{R}^d \) a.s., for some constant \( M_s > 0 \).

\[\blacksquare\]
B. The estimation error of the DRM

In the following lemma, we bound the estimation error of the DRM.

**Lemma 18:** \( \mathbb{E} \left[ \left| \hat{\rho}_g(\theta) - \hat{\rho}_g^H(\theta) \right|^2 \right] \leq \frac{16 M^2 \gamma^2 \lambda^2}{m}. \)

**Proof:** We use parallel arguments to the proof of Lemma [13].

From Lemma [17] we obtain \( \forall x \in [-M_r, M_r], \ |(R^\theta \leq x)\psi^\theta| \leq M_s \) a.s. From Hoeffding inequality, we obtain \( \forall x \in [-M_r, M_r], \)

\[
P \left( \left| \hat{H}^m_{R^\theta}(x) - F_{R^\theta}(x) \right| > \epsilon \right) \leq 2 \exp \left( -\frac{m \epsilon^2}{2\lambda^2} \right).
\] (42)

From (11) and (12), we observe that \( P \left( |H^m_{R^\theta}(x) - F_{R^\theta}(x)| > \epsilon \right) \leq P \left( |\hat{H}^m_{R^\theta}(x) - F_{R^\theta}(x)| > \epsilon \right). \) Hence, we obtain \( \forall x \in [-M_r, M_r], \)

\[
P \left( |H^m_{R^\theta}(x) - F_{R^\theta}(x)| > \epsilon \right) \leq 2 \exp \left( -\frac{m \epsilon^2}{2\lambda^2} \right).
\] (43)

Using similar arguments as in (38) along with (43), we obtain \( \forall x \in [-M_r, M_r], \)

\[
\mathbb{E} \left[ |H^m_{R^\theta}(x) - F_{R^\theta}(x)|^2 \right] \leq \frac{4 M^2 \gamma^2}{m}, \forall x.
\] (44)

Using similar arguments as in (38) along with (44), we obtain

\[
\mathbb{E} \left[ \left| \hat{\rho}_g(\theta) - \hat{\rho}_g^H(\theta) \right|^2 \right] = \frac{16 M^2 \gamma^2 \lambda^2}{m}.
\]

C. Bias and variance of the DRM gradient estimate

The following lemmas establish bounds for the bias and variance of the DRM gradient estimate.

**Lemma 19:** \( \mathbb{E} \left[ \frac{d}{n} \sum_{i=1}^n \frac{\hat{\rho}_g^H(\theta + \mu v_i) - \rho_g(\theta + \mu v_i)}{2\mu} v_i \right]^2 \leq \frac{4d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn}. \)

**Proof:** Using similar arguments as in Lemma [14] along with Lemma [18], we obtain

\[
\mathbb{E} \left[ \frac{d}{n} \sum_{i=1}^n \frac{\hat{\rho}_g^H(\theta + \mu v_i) - \rho_g(\theta + \mu v_i)}{2\mu} v_i \right]^2 \leq \frac{4d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn}.
\]

**Lemma 20:** \( \mathbb{E} \left[ \left\| \nabla_{\mu,v} \hat{\rho}_g^H(\theta) - \nabla \rho_g(\theta) \right\|^2 \right] \leq \frac{4d^2 L^2}{n} + \mu^2 \left( \frac{L^2}{L^2} + \frac{16d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn} \right). \)

**Proof:** The result follows using similar arguments as in Lemma [15] along with Lemma [19].

**Lemma 21:** \( \mathbb{E} \left[ \left\| \nabla_{\mu,v} \hat{\rho}_g^H(\theta) \right\|^2 \right] \leq \frac{2d^2 L^2}{n} + \frac{16d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn} \).

**Proof:** The result follows using similar arguments as in Lemma [16] along with Lemma [19].

D. Non-asymptotic bound for DRM-Off-P-SF

Finally, we provide a non-asymptotic bound for DRM-Off-P-SF.

**Proof:** (Theorem 2: DRM-Off-P-SF) By using a completely parallel argument to the initial passage in the proof of Theorem [1] leading up to (41), we obtain

\[
\alpha \mathbb{E} \left[ \left\| \nabla \rho_g(\theta_k) \right\|^2 \right] \leq 2 \mathbb{E} \left[ \rho_g(\theta_{k+1}) - \rho_g(\theta_k) \right] + L_\rho \alpha^2 \mathbb{E} \left[ \left\| \nabla_{\mu,v} \hat{\rho}_g^H(\theta_k) \right\|^2 \right] + \alpha \mathbb{E} \left[ \left\| \nabla \rho_g(\theta_k) - \nabla_{\mu,v} \hat{\rho}_g^H(\theta_k) \right\|^2 \right]
\]

\[
\leq 2 \mathbb{E} \left[ \rho_g(\theta_{k+1}) - \rho_g(\theta_k) \right] + L_\rho \alpha^2 \left( \frac{2d^2 L^2}{n} + \frac{16d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn} \right) + \alpha \left( \frac{4d^2 L^2}{n} + \mu^2 \left( \frac{L^2}{L^2} + \frac{16d^2 M^2 \gamma^2 \lambda^2}{\mu^2 mn} \right) \right), \tag{45}
\]

where the last inequality follows from lemmas 20, 21.

Summing up (45) from \( k = 0, \ldots, N - 1 \), we obtain

\[
\alpha \sum_{k=0}^{N-1} \mathbb{E} \left[ \left\| \nabla \rho_g(\theta_k) \right\|^2 \right]
\]
\[ E \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] = \frac{1}{N} \sum_{k=0}^{N-1} E \left[ \| \nabla \rho_g(\theta_k) \|^2 \right] \leq 2 \left( \rho_g^* - \rho_g(\theta_0) \right) \frac{\rho_{\mu} \alpha^2}{N' \alpha} + N' L \rho' \alpha \left( \frac{2d^2 L_{\rho}^2}{n} + \frac{16d^2 M^2_{\rho} M^2_g M^2_s}{\mu^2 mn} \right) + \frac{4d^2 L_{\rho}^2}{n} + \mu^2 d^2 L_{\rho}^2 + \frac{16d^2 M^2_{\rho} M^2_g M^2_s}{\mu^2 mn} \] 

Since \( \theta_R \) is chosen uniformly at random from the policy iterates \( \{\theta_1, \ldots, \theta_N\} \), we obtain

\[ E \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] \leq 2 \left( \rho_g^* - \rho_g(\theta_0) \right) \frac{\rho_{\mu} \alpha^2}{N' \alpha} + N' L \rho' \alpha \left( \frac{2d^2 L_{\rho}^2}{n} + \frac{16d^2 M^2_{\rho} M^2_g M^2_s}{\mu^2 mn} \right) + \frac{4d^2 L_{\rho}^2}{n} + \mu^2 d^2 L_{\rho}^2 + \frac{16d^2 M^2_{\rho} M^2_g M^2_s}{\mu^2 mn} \] 

Since \( \alpha = \frac{1}{\sqrt{N}} \), \( \mu = \frac{1}{\sqrt{N}} \), and \( n = N \), we obtain

\[ E \left[ \| \nabla \rho_g(\theta_R) \|^2 \right] \leq \frac{2 \left( \rho_g^* - \rho_g(\theta_0) \right)}{\sqrt{N}} + \frac{d^2 L_{\rho}^2}{\sqrt{N}} + 2 \frac{d^2 L_{\rho}^2 L_{\rho}^2}{N \sqrt{N}} + \frac{16d^2 L_{\rho}^2 M^2_{\rho} M^2_g M^2_s}{mN} + \frac{4d^2 L_{\rho}^2}{N^2} + \frac{16d^2 M^2_{\rho} M^2_g M^2_s}{m \sqrt{N}} \] 

**APPENDIX VI**

**PSEUDOCODE OF DRM-offP-SF**

Algorithm 2 presents the pseudocode of DRM-OffP-SF.

**Algorithm 2** DRM-OffP-SF

1. **Input**: Parameterized form of the policy \( \pi \), behavior policy \( b \), iteration limit \( N \), step-size \( \alpha \), perturbation parameter \( \mu \), 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. Get \( m \) episodes from \( b \);
5. for \( i = 1, \ldots, n \) do
6. Get \( [v_1^i, \ldots, v_d^i] \in S^{d-1} \);
7. Use (14) to estimate \( \hat{\rho}_g^H(\theta_k \pm \mu v_i) \);
8. end for
9. Use (15) to estimate \( \hat{\nabla}_{\mu,n} \rho_g^H(\theta_k) \);
10. Use (16) to calculate \( \theta_{k+1} \);
11. end for
12. **Output**: Policy \( \theta_R \), where \( R \) is chosen uniformly at random from \( \{1, \ldots, N\} \).