A policy gradient approach for optimization of smooth risk measures (Supplementary Material)

Nithia Vijayan[1] **Prashanth L.A.**[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras, India.

A RESULTS FOR THE POLICY GRADIENT TEMPLATE

A.1 RESULTS WITH TRUE OBJECTIVE FUNCTION $\rho(\cdot)$

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

Lemma 7. $\mathbb{E}\left[\widehat{\nabla}_{\mu,n}\rho(\theta) \mid \theta\right] = \nabla\rho_\mu(\theta)$.

Proof. We follow the technique from Shamir [2017]. Since $v_{1:n}$ are i.i.d r.v.s, and have symmetric distribution around the origin, we obtain

$$\begin{aligned}\mathbb{E}\left[\widehat{\nabla}_{\mu,n}\rho(\theta) \mid \theta\right] = \mathbb{E}_{v_{1:n}}\left[\widehat{\nabla}_{\mu,n}\rho(\theta)\right] &= \frac{d}{2\mu n}\sum_{i=1}^{n}\mathbb{E}_v\left[\left(\rho(\theta+\mu v)-\rho(\theta-\mu v)\right)v\right] \\ &= \frac{d}{2\mu}\left(\mathbb{E}_v\left[\rho(\theta+\mu v)v\right]+\mathbb{E}_v\left[\rho(\theta+\mu(-v))(-v)\right]\right) = \frac{d}{\mu}\mathbb{E}_v\left[\rho(\theta+\mu v)v\right] = \nabla\rho_\mu(\theta),\end{aligned}$$

where last equality follows from [Flaxman et al., 2005, Lemma 2.1]. □

Lemma 8. *Suppose* $\forall\theta_1,\theta_2\in\mathbb{R}^d$, $\|\nabla\rho(\theta_1)-\nabla\rho(\theta_2)\| \le L_{\rho'}\|\theta_1-\theta_2\|$. *Then* $\|\nabla\rho_\mu(\theta)-\nabla\rho(\theta)\| \le \frac{\mu d L_{\rho'}}{2}$.

Proof. The result follows from [Gao et al., 2018, Proposition 7.5]. □

Lemma 9. *Suppose* $\forall\theta\in\mathbb{R}^d$, $\rho(\theta)$ *is bounded and* $\forall\theta_1,\theta_2\in\mathbb{R}^d$, $|\rho(\theta_1)-\rho(\theta_2)| \le L_\rho\|\theta_1-\theta_2\|$. *Then* $\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta)\right\|^2\right] \le \frac{d^2L_\rho^2}{n}$.

Proof. Since $\forall v\in\mathbb{S}^{d-1}$, $\|v\|=1$, from (12), we have

$$\begin{aligned}\mathbb{E}_{v_{1:n}}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta)\right\|^2\right] &\overset{(a)}{\le} \frac{d^2}{4\mu^2n^2}\sum_{i=1}^{n}\mathbb{E}_v\left[\left\|\left(\rho(\theta+\mu v)-\rho(\theta-\mu v)\right)v\right\|^2\right] \\ &\le \frac{d^2}{4\mu^2 n}\mathbb{E}_v\left[\left\|\left(\rho(\theta+\mu v)-\rho(\theta-\mu v)\right)\right\|^2\|v\|^2\right] \\ &\le \frac{d^2L_\rho^2}{n}\|v\|^4 = \frac{d^2L_\rho^2}{n},\end{aligned}$$

where (a) follows from the fact that $v_{1:n}$ are i.i.d mean zero r.v.s, and $\rho(\cdot)$ is bounded. Finally,

$$\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta)\right\|^2\right] = \mathbb{E}\left[\mathbb{E}_{v_{1:n}}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta)\right\|^2\right]\right] \le \frac{d^2L_\rho^2}{n}.$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

□

A.2 RESULTS WITH APPROXIMATE OBJECTIVE FUNCTION $\hat{\rho}_m(\cdot)$

The following lemmas establish bounds for the bias and variance of the gradient estimate in (13).

Lemma 10. *Suppose* $\forall \theta \in \mathbb{R}^d$, $\rho(\theta)$ *and* $\hat{\rho}_m(\theta)$ *are bounded, and* $\mathbb{E}\left[|\rho(\theta) - \hat{\rho}_m(\theta)|^2\right] \leq \frac{C_1}{m}$. *Then*

$$\mathbb{E}\left[\left\|\frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}_m(\theta \pm \mu v_i) - \rho(\theta \pm \mu v_i)}{2\mu}v_i\right\|^2\right] \leq \frac{d^2 C_1}{4\mu^2 mn}.$$

Proof. Notice that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}_m(\theta \pm \mu v_i) - \rho(\theta \pm \mu v_i)}{2\mu}v_i\right\|^2\right] \leq \frac{d^2}{4n^2\mu^2}\mathbb{E}\left[\mathbb{E}_{v_{1:n}}\left\|\sum_{i=1}^{n}\left(\hat{\rho}_m(\theta \pm \mu v_i) - \rho(\theta \pm \mu v_i)\right)v_i\right\|^2\right] \\
&\overset{(a)}{\leq} \frac{d^2 n}{4\mu^2 n^2}\mathbb{E}\left[\mathbb{E}_v\left[\|\hat{\rho}_m(\theta \pm \mu v) - \rho(\theta \pm \mu v)v\|^2\right]\right] \\
&\overset{(b)}{\leq} \frac{d^2}{4\mu^2 n}\mathbb{E}\left[\mathbb{E}_v\left[\left(\hat{\rho}_m(\theta \pm \mu v) - \rho(\theta \pm \mu v)\right)^2\right]\right] \\
&= \frac{d^2}{4\mu^2 n}\mathbb{E}\left[\left(\hat{\rho}_m(\theta \pm \mu v) - \rho(\theta \pm \mu v)\right)^2\right] \leq \frac{d^2 C_1}{4\mu^2 mn},
\end{aligned}$$

where (a) follows from the fact that $v_{1:n}$ are i.i.d mean zero r.v.s, and $\hat{\rho}_m(\cdot)$ and $\rho(\cdot)$ are bounded, and (b) follows since $\|v\| = 1$. □

Lemma 11. $\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta) - \nabla\rho(\theta)\right\|^2\right] \leq \frac{4d^2 L_\rho^2}{n} + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2 C_1}{\mu^2 mn}$

Proof. Notice that

$$\begin{aligned}
\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta) &= \frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}_m(\theta + \mu v_i) - \hat{\rho}_m(\theta - \mu v_i)}{2\mu}v_i \\
&= \frac{d}{n}\sum_{i=1}^{n}\frac{\rho(\theta + \mu v_i) - \rho(\theta - \mu v_i)}{2\mu}v_i + \frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}_m(\theta + \mu v_i) - \rho(\theta + \mu v_i)}{2\mu}v_i + \frac{d}{n}\sum_{i=1}^{n}\frac{\rho(\theta - \mu v_i) - \hat{\rho}_m(\theta - \mu v_i)}{2\mu}v_i \\
&= \widehat{\nabla}_{\mu,n}\rho(\theta) + \frac{d}{n}\sum_{i=1}^{n}\frac{\hat{\rho}_m(\theta + \mu v_i) - \rho(\theta + \mu v_i)}{2\mu}v_i + \frac{d}{n}\sum_{i=1}^{n}\frac{\rho(\theta - \mu v_i) - \hat{\rho}_m(\theta - \mu v_i)}{2\mu}v_i. \qquad (37)
\end{aligned}$$

From (37) and Lemma 10, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta) - \nabla\rho(\theta)\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta) - \nabla\rho_g(\theta)\right\|^2\right] + \frac{d^2 C_1}{\mu^2 mn} \\
&\overset{(a)}{\leq} 4\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta) - \mathbb{E}\left[\widehat{\nabla}_{\mu,n}\rho(\theta) \mid \theta\right]\right\|^2\right] + 4\mathbb{E}\left[\|\nabla\rho_\mu(\theta) - \nabla\rho(\theta)\|^2\right] + \frac{d^2 C_1}{\mu^2 mn} \\
&\overset{(b)}{\leq} 4\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho_g(\theta)\right\|^2\right] + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2 C_1}{\mu^2 mn} \\
&\overset{(c)}{\leq} \frac{4d^2 L_\rho^2}{n} + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2 C_1}{\mu^2 mn},
\end{aligned}$$

where (a) follows from Lemma 7, (b) follows from Lemma 8 and since $\mathbb{E}[\|X - E[X \mid Y]\|^2] \leq \mathbb{E}[\|X\|^2]$, and (c) follows from Lemma 9. □

Lemma 12. $\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta)\right\|^2\right] \leq \frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}$.

Proof. Using (37) and Lemma 10, we obtain

$$\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta)\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\rho(\theta)\right\|^2\right] + \frac{d^2C_1}{\mu^2 mn} \leq \frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn},$$

where the last inequality follows from Lemma 9. □

A.3 PROOF OF PROPOSITION 1

Using the fundamental theorem of calculus, we obtain

$$\begin{aligned}
&\rho(\theta_k) - \rho(\theta_{k+1}) = \langle \nabla\rho(\theta_k), \theta_k - \theta_{k+1}\rangle + \int_0^1 \langle \nabla\rho(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})) - \nabla\rho(\theta_k), \theta_k - \theta_{k+1}\rangle \, d\tau \\
&\leq \langle \nabla\rho(\theta_k), \theta_k - \theta_{k+1}\rangle + \int_0^1 \|\nabla\rho(\theta_{k+1} + \tau(\theta_k - \theta_{k+1})) - \nabla\rho(\theta_k)\| \, \|\theta_k - \theta_{k+1}\| \, d\tau \\
&\overset{(a)}{\leq} \langle \nabla\rho(\theta_k), \theta_k - \theta_{k+1}\rangle + L_{\rho'} \|\theta_k - \theta_{k+1}\|^2 \int_0^1 (1-\tau) d\tau \\
&= \langle \nabla\rho(\theta_k), \theta_k - \theta_{k+1}\rangle + \frac{L_{\rho'}}{2} \|\theta_k - \theta_{k+1}\|^2 \\
&= \alpha \left\langle \nabla\rho(\theta_k), -\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\rangle + \frac{L_{\rho'}}{2}\alpha^2 \left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2 \\
&= \alpha \left\langle \nabla\rho(\theta_k), \nabla\rho(\theta_k) - \widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\rangle - \alpha\|\nabla\rho(\theta_k)\|^2 + \frac{L_{\rho'}}{2}\alpha^2 \left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2 \\
&\overset{(b)}{\leq} \frac{\alpha}{2}\|\nabla\rho(\theta_k)\|^2 + \frac{\alpha}{2}\left\|\nabla\rho(\theta_k) - \widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2 - \alpha\|\nabla\rho(\theta_k)\|^2 + \frac{L_{\rho'}}{2}\alpha^2 \left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2 \\
&= \frac{\alpha}{2}\left\|\nabla\rho(\theta_k) - \widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2 - \frac{\alpha}{2}\|\nabla\rho(\theta_k)\|^2 + \frac{L_{\rho'}}{2}\alpha^2 \left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2. \qquad (38)
\end{aligned}$$

In the above the step (a) follows since $\rho(\cdot)$ is smooth and the step (b) follows from $2\langle a, b\rangle \leq \|a\|^2 + \|b\|^2$. Rearranging and taking expectations on both sides of (38), we obtain

$$\begin{aligned}
&\alpha\mathbb{E}\left[\|\nabla\rho(\theta_k)\|^2\right] \\
&\leq 2\mathbb{E}\left[\rho(\theta_{k+1}) - \rho(\theta_k)\right] + L_{\rho'}\alpha^2\mathbb{E}\left[\left\|\widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2\right] + \alpha\mathbb{E}\left[\left\|\nabla\rho(\theta_k) - \widehat{\nabla}_{\mu,n}\hat{\rho}_m(\theta_k)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\rho(\theta_{k+1}) - \rho(\theta_k)\right] + L_{\rho'}\alpha^2\left(\frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}\right) + \alpha\left(\frac{4d^2L_\rho^2}{n} + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2C_1}{\mu^2 mn}\right) \qquad (39)
\end{aligned}$$

where the last inequality follows from lemmas 11-12.

Summing up (39) from $k = 0, \cdots, N-1$, we obtain

$$\alpha\sum_{k=0}^{N-1}\mathbb{E}\left[\|\nabla\rho(\theta_k)\|^2\right] \leq 2\mathbb{E}\left[\rho(\theta_N) - \rho(\theta_0)\right] + NL_{\rho'}\alpha^2\left(\frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}\right) + N\alpha\left(\frac{4d^2L_\rho^2}{n} + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2C_1}{\mu^2 mn}\right).$$

Since θ_R is chosen uniformly at random from the policy iterates $\{\theta_0, \cdots, \theta_{N-1}\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\rho(\theta_R)\|^2\right] &= \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|\nabla\rho(\theta_k)\|^2\right] \\
&\leq \frac{2\left(\rho^* - \rho(\theta_0)\right)}{N\alpha} + L_{\rho'}\alpha\left(\frac{2d^2L_\rho^2}{n} + \frac{d^2C_1}{\mu^2 mn}\right) + \frac{4d^2L_\rho^2}{n} + \mu^2 d^2 L_{\rho'}^2 + \frac{d^2C_1}{\mu^2 mn}.
\end{aligned}$$

□

B DRM

B.1 ESTIMATING DRM USING ORDER STATISTICS

The following lemma estimates the DRM in an on-policy RL setting.

Lemma 13. $\hat{\rho}_g^G(\theta) = \sum_{i=1}^{m} R_{(i)}^{\theta} \left(g\left(1 - \frac{i-1}{m}\right) - g\left(1 - \frac{i}{m}\right)\right)$.

Proof. Our proof follows the technique from Kim [2010]. We rewrite (18) as

$$G_{R^\theta}^m(x) = \begin{cases} 0, & \text{if } x < R_{(1)}^\theta \\ \frac{i}{m}, & \text{if } R_{(i)}^\theta \leq x < R_{(i+1)}^\theta, i \in \{1,\cdots,m-1\} \\ 1, & \text{if } x \geq R_{(m)}^\theta, \end{cases} \tag{40}$$

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 $R_{(j)}^\theta < 0 < R_{(j+1)}^\theta$, and obtain,

$$\begin{aligned}
\hat{\rho}_g^G(\theta) &= \int_{-M_r}^{0} (g(1 - G_{R^\theta}^m(x)) - 1)dx + \int_{0}^{M_r} g(1 - G_{R^\theta}^m(x))dx \\
&= \int_{-M_r}^{R_{(1)}^\theta} (g(1 - G_{R^\theta}^m(x)) - 1)dx + \sum_{i=2}^{j} \int_{R_{(i-1)}^\theta}^{R_{(i)}^\theta} (g(1 - G_{R^\theta}^m(x)) - 1)dx + \int_{R_{(j)}^\theta}^{0} (g(1 - G_{R^\theta}^m(x)) - 1)dx \\
&\quad + \int_{0}^{R_{(j+1)}^\theta} g(1 - G_{R^\theta}^m(x))dx + \sum_{i=j+1}^{m-1} \int_{R_{(i)}^\theta}^{R_{(i+1)}^\theta} g(1 - G_{R^\theta}^m(x))dx + \int_{R_{(m)}^\theta}^{M_r} g(1 - G_{R^\theta}^m(x))dx \\
&= \sum_{i=2}^{j} \int_{R_{(i-1)}^\theta}^{R_{(i)}^\theta} \left(g\left(1 - \frac{i-1}{m}\right) - 1\right) dx + \int_{R_{(j)}^\theta}^{0} \left(g\left(1 - \frac{j}{m}\right) - 1\right) dx + \int_{0}^{R_{(j+1)}^\theta} g\left(1 - \frac{j}{m}\right) dx \\
&\quad + \sum_{i=j+1}^{m-1} \int_{R_{(i)}^\theta}^{R_{(i+1)}^\theta} g\left(1 - \frac{i}{m}\right) dx \\
&= \sum_{i=2}^{j} \left(R_{(i)}^\theta - R_{(i-1)}^\theta\right) \left(g\left(1 - \frac{i-1}{m}\right) - 1\right) - R_{(j)}^\theta \left(g\left(1 - \frac{j}{m}\right) - 1\right) + R_{(j+1)}^\theta g\left(1 - \frac{j}{m}\right) \\
&\quad + \sum_{i=j+1}^{m-1} \left(R_{(i+1)}^\theta - R_{(i)}^\theta\right) g\left(1 - \frac{i}{m}\right) \\
&= \sum_{i=2}^{j} \left(R_{(i)}^\theta - R_{(i-1)}^\theta\right) g\left(1 - \frac{i-1}{m}\right) + R_{(1)}^\theta + \sum_{i=j}^{m-1} \left(R_{(i+1)}^\theta - R_{(i)}^\theta\right) g\left(1 - \frac{i}{m}\right) \\
&= \sum_{i=1}^{m} R_{(i)}^\theta g\left(1 - \frac{i-1}{m}\right) - \sum_{i=1}^{m-1} R_{(i)}^\theta g\left(1 - \frac{i}{m}\right) \\
&= \sum_{i=1}^{m} R_{(i)}^\theta \left(g\left(1 - \frac{i-1}{m}\right) - g\left(1 - \frac{i}{m}\right)\right).
\end{aligned}$$

□

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

Lemma 14. $\hat{\rho}_g^H(\theta) = R^b_{(1)} + \sum_{i=2}^{m} R^b_{(i)} g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i-1}\psi^\theta_{(k)}\right\}\right) - \sum_{i=1}^{m-1} R^b_{(i)} g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i}\psi^\theta_{(k)}\right\}\right)$.

Proof. We rewrite (22) as

$$H^m_{R^\theta}(x) = \begin{cases} 0, & \text{if } x < R^b_{(1)} \\ min\{1, \frac{1}{m}\sum_{j=1}^{i}\psi^\theta_{(j)}\}, & \text{if } R^b_{(i)} \le x < R^\theta_{(i+1)}, i \in \{1,\cdots,m-1\} \\ 1, & \text{if } x \ge R^b_{(m)}, \end{cases} \tag{41}$$

where $R^b_{(i)}$ is the i^{th} smallest order statistic from the samples $R^b_1, \cdots R^b_m$, and $\psi^\theta_{(i)}$ is the importance sampling ratio of $R^b_{(i)}$. We assume without loss of generality that $R^b_{(j)} < 0 < R^b_{(j+1)}$, and obtain,

$$\begin{aligned}
\hat{\rho}_g^H(\theta) &= \int_{-M_r}^{0} (g(1 - H^m_{R^\theta}(x)) - 1)dx + \int_{0}^{M_r} g(1 - H^m_{R^\theta}(x))dx \\
&= \int_{-M_r}^{R^b_{(1)}} (g(1 - H^m_{R^\theta}(x)) - 1)dx + \sum_{i=2}^{j} \int_{R^b_{(i-1)}}^{R^b_{(i)}} (g(1 - H^m_{R^\theta}(x)) - 1)dx + \int_{R^b_{(j)}}^{0} (g(1 - H^m_{R^\theta}(x)) - 1)dx \\
&\quad + \int_{0}^{R^b_{(j+1)}} g(1 - H^m_{R^\theta}(x))dx + \sum_{i=j+1}^{m-1} \int_{R^b_{(i)}}^{R^b_{(i+1)}} g(1 - H^m_{R^\theta}(x))dx + \int_{R^b_{(m)}}^{M_r} g(1 - H^m_{R^\theta}(x))dx \\
&= \sum_{i=2}^{j} \int_{R^b_{(i-1)}}^{R^b_{(i)}} \left(g\left(1 - min\left\{1, \frac{1}{m}\sum_{k=1}^{i-1}\psi^\theta_{(k)}\right\}\right) - 1\right)dx + \int_{R^b_{(j)}}^{0} \left(g\left(1 - min\left\{1, \frac{1}{m}\sum_{k=1}^{j}\psi^\theta_{(k)}\right\}\right) - 1\right)dx \\
&\quad + \int_{0}^{R^b_{(j+1)}} g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{j}\psi^\theta_{(k)}\right\}\right)dx + \sum_{i=j+1}^{m-1} \int_{R^b_{(i)}}^{R^b_{(i+1)}} g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i}\psi^\theta_{(k)}\right\}\right)dx \\
&= \sum_{i=2}^{j} \left(R^b_{(i)} - R^b_{(i-1)}\right)\left(g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i-1}\psi^\theta_{(k)}\right\}\right) - 1\right) - R^b_{(j)}\left(g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{j}\psi^\theta_{(k)}\right\}\right) - 1\right) \\
&\quad + R^b_{(j+1)} g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{j}\psi^\theta_{(k)}\right\}\right) + \sum_{i=j+1}^{m-1} \left(R^b_{(i+1)} - R^b_{(i)}\right) g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i}\psi^\theta_{(k)}\right\}\right) \\
&= \sum_{i=2}^{j} \left(R^b_{(i)} - R^b_{(i-1)}\right) g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i-1}\psi^\theta_{(k)}\right\}\right) + R^b_{(1)} \\
&\quad + \sum_{i=j}^{m-1} \left(R^b_{(i+1)} - R^b_{(i)}\right) g\left(1 - \min\left\{1, \frac{1}{m}\sum_{k=1}^{i}\psi^\theta_{(k)}\right\}\right) \\
&= 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).
\end{aligned}$$

□

B.2 THE ESTIMATION ERROR OF THE DRM

In the following lemma, we bound the estimation error of the DRM in an on-policy RL setting.

Proof. (**Lemma 1**) Since $\forall x \in (-M_r, M_r), \left|\mathbb{1}\{R^\theta \leq x\}\right| \leq 1$ a.s., using Hoeffding's inequality, we obtain $\forall x \in (-M_r, M_r)$,

$$\mathbb{P}\left(|G^m_{R^\theta}(x) - F_{R^\theta}(x)| > \epsilon\right) \leq 2\exp\left(-\frac{m\epsilon^2}{2}\right), \text{ and}$$
$$\mathbb{E}\left[|G^m_{R^\theta}(x) - F_{R^\theta}(x)|^2\right] = \int_0^\infty \mathbb{P}\left(|G^m_{R^\theta}(x) - F_{R^\theta}(x)| > \sqrt{\epsilon}\right) d\epsilon \leq \int_0^\infty 2\exp\left(-\frac{m\epsilon}{2}\right) d\epsilon = \frac{4}{m}. \tag{42}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\left|\rho_g(\theta) - \hat{\rho}^G_g(\theta)\right|^2\right] &= \mathbb{E}\left[\left|\int_{-M_r}^{M_r} (g(1 - F_{R^\theta}(x)) - g(1 - G^m_{R^\theta}(x)))dx\right|^2\right] \\
&\overset{(a)}{\leq} 2M_r\mathbb{E}\left[\int_{-M_r}^{M_r} \left|(g(1 - F_{R^\theta}(x)) - g(1 - G^m_{R^\theta}(x)))\right|^2 dx\right] \\
&\overset{(b)}{\leq} 2M_r \int_{-M_r}^{M_r} \mathbb{E}\left[\left|(g(1 - F_{R^\theta}(x)) - g(1 - G^m_{R^\theta}(x)))\right|^2\right] dx \\
&\overset{(c)}{\leq} 2M_r M^2_{g'} \int_{-M_r}^{M_r} \mathbb{E}\left[|G^m_{R^\theta}(x) - F_{R^\theta}(x)|^2\right] dx \\
&\overset{(d)}{\leq} 2M_r M^2_{g'} \int_{-M_r}^{M_r} \frac{4}{m} dx = \frac{16M_r^2 M^2_{g'}}{m},
\end{aligned} \tag{43}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) follows from the Fubini's theorem, (c) follows from Lemma 15, and (d) follows from (42). □

In the following lemma, we bound the estimation error of the DRM in an off-policy RL setting.

Proof. (**Lemma 2**) We use parallel arguments to the proof of Lemma 1.

From (6), we obtain $\forall x \in (-M_r, M_r), \left|\mathbb{1}\{R^\theta \leq x\}\psi^\theta\right| \leq M_s$ a.s. From Hoeffding inequality, we obtain $\forall x \in (-M_r, M_r)$,

$$\mathbb{P}\left(\left|\hat{H}^m_{R^\theta}(x) - F_{R^\theta}(x)\right| > \epsilon\right) \leq 2\exp\left(-\frac{m\epsilon^2}{2M_s^2}\right). \tag{44}$$

From (22) and (23), we observe that $\mathbb{P}\left(\left|H^m_{R^\theta}(x) - F_{R^\theta}(x)\right| > \epsilon\right) \leq \mathbb{P}\left(\left|\hat{H}^m_{R^\theta}(x) - F_{R^\theta}(x)\right| > \epsilon\right)$. Hence, we obtain $\forall x \in (-M_r, M_r)$,

$$\mathbb{P}\left(|H^m_{R^\theta}(x) - F_{R^\theta}(x)| > \epsilon\right) \leq 2\exp\left(-\frac{m\epsilon^2}{2M_s^2}\right). \tag{45}$$

Using similar arguments as in (42) along with (45), we obtain $\forall x \in [-M_r, M_r]$,

$$\mathbb{E}\left[|H^m_{R^\theta}(x) - F_{R^\theta}(x)|^2\right] \leq \frac{4M_s^2}{m}, \forall x. \tag{46}$$

Using similar arguments as in (43) along with (46), we obtain

$$\mathbb{E}\left[\left|\rho_g(\theta) - \hat{\rho}^H_g(\theta)\right|^2\right] = \frac{16M_r^2 M^2_{g'} M_s^2}{m}.$$

□

B.3 LIPSCHITZ PROPERTIES OF THE DRM AND ITS GRADIENT

B.3.1 Results related to the distortion function

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

Lemma 15. $\forall t, t' \in (0,1), |g(t) - g(t')| \leq M_{g'} |t - t'|$, *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 (A9), we obtain $\left|g'(\tilde{t})\right| \leq M_{g'}, \forall \tilde{t} \in (0,1)$. Hence, $|g(t) - g(t')| \leq M_{g'} |t - t'| \ \forall t, t' \in (0,1)$.

Similarly, we have $g'(t) - g'(t') = g''(\tilde{t})(t - t')$, where $\tilde{t} \in (t, t')$. From (A9), we obtain $\left|g''(\tilde{t})\right| \leq M_{g''}, \forall \tilde{t} \in (0,1)$. Hence, $|g'(t) - g'(t')| \leq M_{g''} |t - t'| \ \forall t, t' \in (0,1)$. □

B.3.2 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 Vijayan and Prashanth [2023]. For the sake of completeness, we provide the detailed proof.

Lemma 16. $\forall x \in (-M_r, M_r)$,

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

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

Proof. Let Ω 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 ω, and let $\mathbb{P}_\theta(\omega) = \prod_{t=0}^{T(\omega)-1} \pi_\theta(A_t(\omega)|S_t(\omega)) p(S_{t+1}(\omega), S_t(\omega), A_t(\omega))$.

From $\frac{\nabla \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} = \sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega))$, we obtain

$$\begin{aligned}
\nabla F_{R^\theta}(x) &= \nabla \mathbb{E}\big[\mathbb{1}\{R^\theta \leq x\}\big] = \nabla \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \mathbb{P}_\theta(\omega) \\
&\overset{(a)}{=} \sum_{\omega \in \Omega} \nabla \left(\mathbb{1}\{R(\omega) \leq x\} \mathbb{P}_\theta(\omega)\right) \\
&\overset{(b)}{=} \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \nabla \mathbb{P}_\theta(\omega) \\
&= \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \frac{\nabla \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} \mathbb{P}_\theta(\omega) \\
&= \sum_{\omega \in \Omega} \mathbb{1}\{R(\omega) \leq x\} \sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega)) \mathbb{P}_\theta(\omega) \\
&= \mathbb{E}\left[\mathbb{1}\{R^\theta \leq x\} \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right].
\end{aligned}$$

In the above, the equality in (a) follows by an application of the dominated convergence theorem to interchange the differentiation and the expectation operation. The aforementioned application is allowed since (i) Ω 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) $\nabla \log \pi_\theta(A_t|S_t)$ is bounded from (A2). The equality in (b) follows, since for a given episode ω, the cumulative reward $R(\omega)$ does not depend on θ.

Similarly,

from $\frac{\nabla^2 \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} = \sum_{t=0}^{T(\omega)-1} \nabla^2 \log \pi_\theta(A_t(\omega)|S_t(\omega)) + \left[\sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega))\right]\left[\sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega))\right]^T$,

we obtain

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

□

Lemma 17. $\forall x \in (-M_r, M_r), \|\nabla F_{R^\theta}(x)\| \le M_e M_d$, *and* $\left\|\nabla^2 F_{R^\theta}(x)\right\| \le M_e M_h + M_e^2 M_d^2$.

Proof. From (A2) and (2), for any $x \in (-M_r, M_r)$, we have

$$\left\|\mathbb{1}\{R^\theta \le x\}\sum_{t=0}^{T-1}\nabla \log \pi_\theta(A_t \mid S_t)\right\| \le M_e M_d \text{ a.s,} \tag{47}$$

and

$$\left\|\mathbb{1}\{R^\theta \le x\}\left(\sum_{t=0}^{T-1}\nabla^2 \log \pi_\theta(A_t|S_t) + \left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]^T\right)\right\| \le M_e M_h + M_e^2 M_d^2 \text{ a.s.} \tag{48}$$

From Lemma 16, for any $x \in (-M_r, M_r)$, we have

$$\|\nabla F_{R^\theta}(x)\| \le \mathbb{E}\left[\left\|\mathbb{1}\{R^\theta \le x\}\sum_{t=0}^{T-1}\nabla \log \pi_\theta(A_t|S_t)\right\|\right] \le M_e M_d, \tag{49}$$

and

$$\begin{aligned}\left\|\nabla^2 F_{R^\theta}(x)\right\| &\le \mathbb{E}\left[\left\|\mathbb{1}\{R^\theta \le x\}\left(\sum_{t=0}^{T-1}\nabla^2 \log \pi_\theta(A_t|S_t) + \left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]^T\right)\right\|\right] \\ &\le M_e M_h + M_e^2 M_d^2,\end{aligned} \tag{50}$$

where these inequalities follow from (47), (48), and the assumption that the state and action spaces are finite. □

The following lemma establishes Lipschitzness of the CDF and its gradient.

Lemma 18. $\forall x \in (-M_r, M_r)$,

$$\begin{aligned}&|F_{R^{\theta_1}}(x) - F_{R^{\theta_2}}(x)| \le M_e M_d \|\theta_1 - \theta_2\|, \text{ and} \\ &\|\nabla F_{R^{\theta_1}}(x) - \nabla F_{R^{\theta_2}}(x)\| \le (M_e M_h + M_e^2 M_d^2)\|\theta_1 - \theta_2\|.\end{aligned}$$

Proof. The result follows by Lemma 17 and Lemma 1.2.2 in Nesterov [2004]. □

B.3.3 Gradient of the DRM

The following lemma derives an expression for the gradient of the DRM. This lemma is similar to Theorem 1 in Vijayan and Prashanth [2023]. For the sake of completeness, we provide detailed proof.

Lemma 19. $\nabla \rho_g(\theta) = -\int_{-M_r}^{M_r} g'(1 - F_{R^\theta}(x)) \nabla F_{R^\theta}(x) dx$.

Proof. Notice that

$$\begin{aligned}
\nabla \rho_g(\theta) &= \nabla \int_{-M_r}^{0} \left(g(1 - F_{R^\theta}(x)) - 1\right) dx + \nabla \int_{0}^{M_r} g(1 - F_{R^\theta}(x)) dx \\
&\overset{(a)}{=} \int_{-M_r}^{0} \nabla \left(g(1 - F_{R^\theta}(x)) - 1\right) dx + \int_{0}^{M_r} \nabla g(1 - F_{R^\theta}(x)) dx \\
&= -\int_{-M_r}^{M_r} g'(1 - F_{R^\theta}(x)) \nabla F_{R^\theta}(x) dx.
\end{aligned}$$

In the above, (a) follows by an application of the dominated convergence theorem to interchange the differentiation and the integration 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 (A9), and $\nabla F_{R^\theta}(\cdot)$ is bounded from (49). The bounds on g' and ∇F_{R^θ} imply $\int_{-M_r}^{M_r} \|g'(1 - F_{R^\theta}(x)) \nabla F_{R^\theta}(x)\| \, dx \leq 2M_r M_{g'} M_e M_d$. □

B.3.4 Lipschitz properties of the DRM and its gradient

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

Proof. (**Lemma 3**)

$$\begin{aligned}
|\rho_g(\theta_1) - \rho_g(\theta_2)| &\leq \int_{-M_r}^{M_r} |g(1 - F_{R^{\theta_1}}(x)) - g(1 - F_{R^{\theta_2}}(x))| \, dx \\
&\overset{(a)}{\leq} M_{g'} \int_{-M_r}^{M_r} |F_{R^{\theta_1}}(x) - F_{R^{\theta_2}}(x)| \, dx \\
&\overset{(b)}{\leq} 2M_r M_{g'} M_e M_d \|\theta_1 - \theta_2\|,
\end{aligned}$$

where (a) follows from Lemma 15 and (b) follows from Lemma 18. The result follows since $L_\rho = 2M_r M_{g'} M_e M_d$.

From Lemma 19, we obtain

$$\begin{aligned}
&\|\nabla \rho_g(\theta_1) - \nabla \rho_g(\theta_2)\| \\
&\leq \int_{-M_r}^{M_r} \|g'(1 - F_{R^{\theta_1}}(x)) \nabla F_{R^{\theta_1}}(x) - g'(1 - F_{R^{\theta_2}}(x)) \nabla F_{R^{\theta_2}}(x)\| \, dx \\
&\leq \int_{-M_r}^{M_r} \|g'(1 - F_{R^{\theta_1}}(x)) \nabla F_{R^{\theta_1}}(x) - g'(1 - F_{R^{\theta_2}}(x)) \nabla F_{R^{\theta_2}}(x) + g'(1 - F_{R^{\theta_1}}(x)) \nabla F_{R^{\theta_2}}(x) \\
&\quad - g'(1 - F_{R^{\theta_2}}(x)) \nabla F_{R^{\theta_2}}(x)\| \, dx \\
&\leq \int_{-M_r}^{M_r} |g'(1 - F_{R^{\theta_1}}(x))| \, \|\nabla F_{R^{\theta_1}}(x) - \nabla F_{R^{\theta_2}}(x)\| + \|\nabla F_{R^{\theta_2}}(x)\| \, |g'(1 - F_{R^{\theta_1}}(x)) - g'(1 - F_{R^{\theta_2}}(x))| \, dx \\
&\overset{(a)}{\leq} \int_{-M_r}^{M_r} M_{g'} \|\nabla F_{R^{\theta_1}}(x) - \nabla F_{R^{\theta_2}}(x)\| + M_e M_d M_{g''} |F_{R^{\theta_1}}(x) - F_{R^{\theta_2}}(x)| \, dx \\
&\overset{(b)}{\leq} \int_{-M_r}^{M_r} M_{g'} (M_e M_h + M_e^2 M_d^2) \|\theta_1 - \theta_2\| + M_e^2 M_d^2 M_{g''} \|\theta_1 - \theta_2\| \, dx \\
&\leq 2M_r M_e \left(M_h M_{g'} + M_e M_d^2 (M_{g'} + M_{g''})\right) \|\theta_1 - \theta_2\|,
\end{aligned}$$

where (a) follows from (A9), and Lemmas 15, 17, and (b) follows from Lemma 18. The result follows since $L_{\rho'} = 2M_r M_e \left(M_h M_{g'} + M_e M_d^2 (M_{g'} + M_{g''})\right)$. □

C MEAN-VARIANCE RISK MEASURE

C.1 THE ESTIMATION ERROR OF THE MVRM

In the following lemma, we bound the estimation error of the MVRM in an on-policy RL setting.

Proof. (**Lemma 4**) From (27) and (31), we obtain

$$\begin{aligned}\mathbb{E}\left[\left|\hat{\rho}_\lambda^\pi(\theta)-\rho_\lambda(\theta)\right|^2\right] &\leq 2\mathbb{E}\left[\left|\hat{J}_m^\pi(\theta)-J(\theta)\right|^2\right]+2\lambda^2\mathbb{E}\left[\left|V(\theta)-\widehat{V}_m^\pi(\theta)\right|^2\right]\\ &\leq \frac{8M_r^2}{m}+\frac{32\lambda^2M_r^4}{m}=\frac{8M_r^2+32\lambda^2M_r^4}{m},\end{aligned} \tag{51}$$

where the last inequality follows from Theorem 2-3 [Mood et al., 1974, chapter V1] in conjunction with the fact $\left|R^\theta\right|\leq M_r$ and $m>2$. □

In the following lemma, we bound the estimation error of the MVRM in an off-policy RL setting.

Proof. (**Lemma 5**) From (27) and (35), we obtain

$$\begin{aligned}\mathbb{E}\left[\left|\hat{\rho}_\lambda^b(\theta)-\rho_\lambda(\theta)\right|^2\right] &\leq 2\mathbb{E}\left[\left|\hat{J}_m^b(\theta)-J(\theta)\right|^2\right]+2\lambda^2\mathbb{E}\left[\left|V(\theta)-\widehat{V}_m^b(\theta)\right|^2\right]\\ &\leq \frac{8M_r^2M_s^2}{m}+\frac{32\lambda^2M_r^4M_s^4}{m}=\frac{8M_r^2M_s^2+32\lambda^2M_r^4M_s^4}{m},\end{aligned} \tag{52}$$

where the last inequality follows from Theorem 2-3 [Mood et al., 1974, chapter V1] in conjunction with the fact $\left|R^b\psi_\theta\right|\leq M_rM_s$, and $m>2$. □

C.2 LIPSCHITZ PROPERTIES OF THE MVRM AND ITS GRADIENT

Proof. (**Lemma 6**) Let Ω 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 ω, and let $\mathbb{P}_\theta(\omega)=\prod_{t=0}^{T(\omega)-1}\pi_\theta(A_t(\omega)|S_t(\omega))p(S_{t+1}(\omega),S_t(\omega),A_t(\omega))$.

From $\frac{\nabla\mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)}=\sum_{t=0}^{T(\omega)-1}\nabla\log\pi_\theta(A_t(\omega)|S_t(\omega))$, we obtain

$$\begin{aligned}\nabla J(\theta)=\nabla\mathbb{E}\left[R^\theta\right]&=\nabla\sum_{\omega\in\Omega}R(\omega)\mathbb{P}_\theta(\omega)\\ &\overset{(a)}{=}\sum_{\omega\in\Omega}\nabla\left(R(\omega)\mathbb{P}_\theta(\omega)\right)\\ &\overset{(b)}{=}\sum_{\omega\in\Omega}R(\omega)\nabla\mathbb{P}_\theta(\omega)\\ &=\sum_{\omega\in\Omega}R(\omega)\frac{\nabla\mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)}\mathbb{P}_\theta(\omega) \qquad (53)\\ &=\sum_{\omega\in\Omega}R(\omega)\sum_{t=0}^{T(\omega)-1}\nabla\log\pi_\theta(A_t(\omega)|S_t(\omega))\mathbb{P}_\theta(\omega)\\ &=\mathbb{E}\left[R^\theta\sum_{t=0}^{T-1}\nabla\log\pi_\theta(A_t|S_t)\right]. \qquad (54)\end{aligned}$$

In the above, (a) follows by an application of the dominated convergence theorem to interchange the differentiation and the expectation operation. The aforementioned application is allowed since (i) Ω 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) $\nabla \log \pi_\theta(A_t|S_t)$ is bounded from (A2). The step (b) follows, since for a given episode ω, the cumulative reward $R(\omega)$ does not depend on θ.

Similarly,

from $\frac{\nabla^2 \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} = \sum_{t=0}^{T(\omega)-1} \nabla^2 \log \pi_\theta(A_t(\omega)|S_t(\omega)) + \left[\sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega))\right]\left[\sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega))\right]^T$,

we obtain

$$\nabla^2 J(\theta) = \mathbb{E}\left[R^\theta \left(\sum_{t=0}^{T-1} \nabla^2 \log \pi_\theta(A_t|S_t) + \left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]^T\right)\right]. \tag{55}$$

Similarly,

$$\begin{aligned}
\nabla \mathbb{E}\left[\left(R^\theta\right)^2\right] &= \nabla \sum_{\omega \in \Omega} R(\omega)^2 \mathbb{P}_\theta(\omega) \\
&= \sum_{\omega \in \Omega} \nabla \left(R(\omega)^2 \mathbb{P}_\theta(\omega)\right) \qquad (56) \\
&= \sum_{\omega \in \Omega} R(\omega)^2 \nabla \mathbb{P}_\theta(\omega) \\
&= \sum_{\omega \in \Omega} R(\omega)^2 \frac{\nabla \mathbb{P}_\theta(\omega)}{\mathbb{P}_\theta(\omega)} \mathbb{P}_\theta(\omega) \qquad (57) \\
&= \sum_{\omega \in \Omega} R(\omega)^2 \sum_{t=0}^{T(\omega)-1} \nabla \log \pi_\theta(A_t(\omega)|S_t(\omega)) \mathbb{P}_\theta(\omega) \\
&= \mathbb{E}\left[\left(R^\theta\right)^2 \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right], \qquad (58)
\end{aligned}$$

and

$$\nabla^2 \mathbb{E}\left[\left(R^\theta\right)^2\right] = \mathbb{E}\left[\left(R^\theta\right)^2 \left(\sum_{t=0}^{T-1} \nabla^2 \log \pi_\theta(A_t|S_t) + \left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]^T\right)\right]. \tag{59}$$

From (53)-(55), we obtain

$$\|\nabla J(\theta)\| \leq \mathbb{E}\left[\left\|R^\theta \sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right\|\right] \leq M_r E\left[\left\|\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right\|\right] \leq M_r M_e M_d, \tag{60}$$

and

$$\begin{aligned}
\left\|\nabla^2 J(\theta)\right\| &\leq \mathbb{E}\left[\left\|R^\theta \left(\sum_{t=0}^{T-1} \nabla^2 \log \pi_\theta(A_t|S_t) + \left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right]^T\right)\right\|\right] \\
&\leq M_r \mathbb{E}\left[\left\|\sum_{t=0}^{T-1} \nabla^2 \log \pi_\theta(A_t|S_t)\right\| + \left\|\sum_{t=0}^{T-1} \nabla \log \pi_\theta(A_t|S_t)\right\|^2\right] \\
&\leq M_r \left(M_e M_h + M_e^2 M_d^2\right). \qquad (61)
\end{aligned}$$

Hence from (61) and Lemma 1.2.2 in Nesterov [2004], we obtain

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq M_r \left(M_e M_h + M_e^2 M_d^2\right) \|\theta_1 - \theta_2\| \tag{62}$$

Similarly, from (56)-(59), we obtain

$$\left\|\nabla\mathbb{E}\left[\left(R^{\theta}\right)^2\right]\right\| \leq M_r^2\mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\nabla\log\pi_\theta(A_t|S_t)\right\|\right] \leq M_r^2 M_e M_d, \tag{63}$$

and

$$\left\|\nabla^2\mathbb{E}\left[\left(R^{\theta}\right)^2\right]\right\| \leq M_r^2\mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\nabla^2\log\pi_\theta(A_t|S_t)\right\| + \left\|\sum_{t=0}^{T-1}\nabla\log\pi_\theta(A_t|S_t)\right\|^2\right] \leq M_r^2\left(M_e M_h + M_e^2 M_d^2\right). \tag{64}$$

Now,

$$\begin{aligned}
\left\|\nabla^2 V(\theta)\right\| &= \left\|\nabla^2\left(\mathbb{E}\left[\left(R^{\theta}\right)^2\right] - J(\theta)^2\right)\right\| \\
&= \left\|\nabla^2\mathbb{E}\left[\left(R^{\theta}\right)^2\right] - 2J(\theta)\nabla^2 J(\theta) - 2\nabla J(\theta)\nabla J(\theta)^\top\right\| \\
&\leq \left\|\nabla^2\mathbb{E}\left[\left(R^{\theta}\right)^2\right]\right\| + 2\left|J(\theta)\right|\left\|\nabla^2 J(\theta)\right\| + 2\left\|\nabla J(\theta)\right\|^2 \\
&\leq 3M_r^2 M_e M_h + 5M_r^2 M_e^2 M_d^2.
\end{aligned} \tag{65}$$

Hence, from (65) and Lemma 1.2.2 in Nesterov [2004], we obtain

$$\left\|\nabla V(\theta_1) - \nabla V(\theta_2)\right\| \leq \lambda\left(3M_r^2 M_e M_h + 5M_r^2 M_e^2 M_d^2\right)\left\|\theta_1 - \theta_2\right\| \tag{66}$$

Now,

$$\begin{aligned}
\left\|\nabla\rho_\lambda(\theta)\right\| &= \left\|\nabla J(\theta) - \lambda\nabla V(\theta)\right\| \\
&\leq \left\|\nabla J(\theta)\right\| + \lambda\left\|\nabla\mathbb{E}\left[\left(R^{\theta}\right)^2\right]\right\| + 2\lambda\left|J(\theta)\right|\left\|\nabla J(\theta)\right\| \\
&\leq M_r M_e M_d + 3\lambda M_r^2 M_e M_d.
\end{aligned} \tag{67}$$

Hence, from (67) and Lemma 1.2.2 in Nesterov [2004], we obtain

$$\left|\rho_\lambda(\theta_1) - \rho_\lambda(\theta_1)\right| \leq \left(M_r M_e M_d + 3\lambda M_r^2 M_e M_d\right)\left\|\theta_1 - \theta_2\right\|. \tag{68}$$

From (62) and (66), we obtain

$$\begin{aligned}
\left\|\nabla\rho_\lambda(\theta_1) - \nabla\rho_\lambda(\theta_1)\right\| &\leq \left\|\nabla J(\theta_1) - \nabla J(\theta_2)\right\| + \lambda\left\|\nabla V(\theta_2) - \nabla V(\theta_1)\right\| \\
&\leq \left(M_r M_e\left(M_h + M_e M_d^2\right) + \lambda M_r^2 M_e\left(3M_h + 5M_e M_d^2\right)\right)\left\|\theta_1 - \theta_2\right\|.
\end{aligned} \tag{69}$$

□

References

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

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

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

A. M. Mood, , F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw Hill, 3rd edition, 1974.

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

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

N. Vijayan and L.A. Prashanth. Policy gradient methods for distortion risk measures. *arXiv preprint arXiv:2107.04422*, 2023.