1 Notations and settings

Let $\theta \in \mathbb{R}^d$ and $F(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$. Let $\vec{\epsilon} \sim \mathcal{N}(0, \mathbb{I}_d)$ be a random vector.

Definition 1. *We define antithetic ES gradient estimator as*

$$F^{AT(i)} = \frac{1}{2\sigma}(F(\theta + \sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)} - F(\theta - \sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)})$$

We define forward finite difference ES gradient estimator as

$$F^{FD(i)} = \frac{1}{\sigma}(F(\theta + \sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)} - F(\theta)\vec{\epsilon}^{(i)})$$

2 Condition for choosing gradient estimator

Assumption 1. *Assume reward function $F(x)$ is quadratic, i.e.*

$$F(\theta + \sigma\vec{\epsilon}) = F(\theta) + \sigma\nabla F(\theta)^\top\vec{\epsilon} + \frac{\sigma^2}{2}\vec{\epsilon}^\top\nabla^2 F(\theta)\vec{\epsilon}$$

Theorem 1. *MSE for antithetic ES gradient estimator is*

$$\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{AT,ort}} F_\sigma(\theta)\right) = \frac{1}{N}\mathbb{E}\left[\left\|(\nabla F(\theta)^\top\vec{\epsilon})\vec{\epsilon}\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2$$

MSE for forward finite difference ES gradient estimator is

$$\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{FD,ort}} F_\sigma(\theta)\right) = \frac{1}{N}\mathbb{E}\left[\left\|(\nabla F(\theta)^\top\vec{\epsilon} + \frac{\sigma}{2}\vec{\epsilon}^\top\nabla^2 F(\theta)\vec{\epsilon})\vec{\epsilon}\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2$$

The proof is relegated to the appendix.

Since we know exact distribution of ϵ. We can expand Theorem 1 result further.

Lemma 1. *Assume $N \leq d$, we have*

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{AT,ort}} F_\sigma(\theta)\right) =& \frac{1}{N}||\nabla F(\theta)||_2^2 \\
\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{FD,ort}} F_\sigma(\theta)\right) =& \frac{1}{N}||\nabla F(\theta)||_2^2 + \frac{(N+4)\sigma^4}{4N}||\nabla^2 F(\theta)||_F^2 + \frac{5\sigma^4}{2N}\left(\sum_i \nabla^2 F(\theta)_{ii}^2\right)
\end{aligned}$$

Corollary 1. *Given at most m calls of blackbox $F(\cdot)$, antithetic ES gradient estimator has smaller MSE than forward if and only if*

$$\frac{2m}{m+1}\mathbb{E}\left[\left\|(\nabla F(\theta)^\top\vec{\epsilon})\vec{\epsilon}\right\|_2^2\right] \leq \mathbb{E}\left[\left\|(\nabla F(\theta)^\top\vec{\epsilon} + \frac{\sigma}{2}\vec{\epsilon}^\top\nabla^2 F(\theta)\vec{\epsilon})\vec{\epsilon}\right\|_2^2\right]$$

Proof. (proof of corollary 1) Notice antithetic calls blackbox $F(\cdot)$ twice for each evaluation of $F^{AT(i)}$. The rest follows from Theorem 1 □

Remark 1. *Assume we can call blackbox $F(\cdot)$ at most m times. When $F(\cdot)$ is linear, hessian is zero matrix, and thus FD has smaller MSE. When $F(\theta)$ has zero gradient at θ, we note antithetic MSE equals zero, and thus AT has smaller MSE.*

3 Hessian approximation

We approximate the true hessian by Hessian of gaussian smoothing. The gaussian smoothing of $F(x)$ is defined as

$$F_\sigma(\theta) = \frac{1}{\kappa}\int F(\theta + \sigma\vec{\epsilon})e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}d\vec{\epsilon} \quad \text{, where } \kappa = (2\pi)^{d/2}.$$

The gradient is

$$\nabla F_\sigma(\theta) = \frac{1}{\sigma\kappa}\int F(\theta + \sigma\vec{\epsilon})e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}\vec{\epsilon}d\vec{\epsilon} \quad \text{can be approximated by } \frac{1}{N\sigma}\sum_{i=1}^{N} F(\theta + \sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)}$$

The Hessian is

$$\nabla^2 F_\sigma(\theta) = \frac{1}{\sigma^2\kappa}\left(\int F(\theta + \sigma\vec{\epsilon})e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}\vec{\epsilon}\vec{\epsilon}^\top d\vec{\epsilon}\right) - \frac{F_\sigma(\theta)}{\sigma^2}\mathbb{I}_d$$

can be approximated by

$$\frac{1}{N\sigma^2}\left(\sum_{i=1}^{N} F(\theta + \sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)}\vec{\epsilon}^{(i)\top}\right) - \mathbb{I}_d\frac{1}{N\sigma^2}\left(\sum_{i=1}^{N} F(\theta + \sigma\vec{\epsilon}^{(i)})\right)$$

To bound the difference between true Hessian and gaussian smoothing approximate, we need to introduce smoothness assumptions.

Assumption 2 (Smoothness)**.** *Assume $F(\theta)$ is twice differentiable and*

$$||\nabla F(\theta) - \nabla F(\theta')||_2 \le L_1||\theta - \theta'||_2,$$
$$||\nabla^2 F(\theta) - \nabla^2 F(\theta')||_2 \le L_2||\theta - \theta'||_2$$

for some constants L_1, L_2.

Remark 2. *Above assumption is equivalent to $F(\theta)$ is twice differentiable and*

$$|F(\theta') - F(\theta) - \langle\nabla F(\theta), \theta' - \theta\rangle| \le \frac{L_1}{2}||\theta - \theta'||_2^2,$$
$$|F(\theta') - F(\theta) - \langle\nabla F(\theta), \theta' - \theta\rangle - \frac{1}{2}\langle\nabla^2 F(\theta)(\theta' - \theta), \theta' - \theta\rangle| \le \frac{L_2}{6}||\theta - \theta'||_2^3$$

where $||\theta - \theta'||_2^3 = -\langle\theta' - \theta, \theta' - \theta\rangle^{\frac{3}{2}}$.

Nesterov and Spokoiny (Nesterov and Spokoiny, 2017) showed an error bound on gaussian smoothing gradient estimate.

$$||\nabla F(\theta)||_2^2 \le 2||\nabla F_\sigma(\theta)||_2^2 + \frac{\sigma^2}{2}L_1^2(d+6)^3$$

Now we prove an error bound on gaussian smoothing Hessian estimate.

$$\begin{aligned}
&\frac{c_1}{2}||\nabla^2 F(\theta)||_F^2 \le \frac{1}{2}||\nabla^2 F(\theta)\int\frac{1}{\kappa}(\vec{\epsilon}\vec{\epsilon}^\top)(\vec{\epsilon}\vec{\epsilon}^\top)e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}d\epsilon||_F^2 \\
&\le \left\|\frac{1}{\kappa\sigma^2}\left(\int F(\theta + \sigma\vec{\epsilon})e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}\vec{\epsilon}\vec{\epsilon}^\top d\vec{\epsilon}\right) - \frac{F_\sigma(\theta)}{\sigma^2}\mathbb{I} - \frac{1}{\kappa\sigma^2}\left(\int\left[F(\theta + \sigma\vec{\epsilon}) - \langle\nabla F(\theta), \sigma\vec{\epsilon}\rangle - \frac{\sigma^2}{2}\langle\nabla^2 F(\theta)\vec{\epsilon}, \vec{\epsilon}\rangle\right]e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}\vec{\epsilon}\vec{\epsilon}^\top d\vec{\epsilon}\right)\right. \\
&\quad\left. + \frac{F_\sigma(\theta)}{\sigma^2}\mathbb{I} - \frac{F(\theta)}{\sigma^2}\int\frac{1}{\kappa}e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}\vec{\epsilon}\vec{\epsilon}^\top d\vec{\epsilon} + \frac{\nabla F(\theta)}{\sigma^2}\int\frac{1}{\kappa}\sigma\vec{\epsilon}\vec{\epsilon}\vec{\epsilon}^\top e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}d\vec{\epsilon}\right\|_F^2 \\
&\le \left\|\nabla^2 F_\sigma(\theta) - \frac{\sigma}{6}L_2 c_2\mathbb{I} + \frac{|F_\sigma(\theta) - F(\theta)|}{\sigma^2}\mathbb{I}\right\|_F^2
\end{aligned}$$

for some constants c_1, c_2 satisfying

$$c_1\mathbb{I} \le \int\frac{1}{\kappa}(\vec{\epsilon}\vec{\epsilon}^\top)(\vec{\epsilon}\vec{\epsilon}^\top)e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}d\epsilon \quad , c_2\mathbb{I} \ge \int\frac{1}{\kappa}||\vec{\epsilon}||_2^3(\vec{\epsilon}\vec{\epsilon}^\top)e^{-\frac{1}{2}||\vec{\epsilon}||_2^2}d\vec{\epsilon}$$

Notice

$$\frac{|F_\sigma(\theta) - F(\theta)|}{\sigma^2} \le \frac{d}{2}L_1.$$

4 Generalized gradient estimator

Definition 2. *We define generalized gradient estimator as*

$$\alpha_0 F(\theta)\vec{\epsilon} + \sum_{i=1}^{K} \alpha_i F(\theta + \beta_i \sigma \vec{\epsilon})\vec{\epsilon}$$

The goal is to construct gradient estimator that satisfies

Condition 1.

$$\sum_i \alpha_i = 0, \quad \sum_{i\geq 1} \alpha_i \beta_i = 1, \quad \sum_{i \geq 1} \alpha \beta_i^j = 0 \quad \textit{for } j \geq 2$$

That is, when we do Taylor expansion, only the first order term (gradient) should remain and others should zero out. In addition, the first order term must sum to one for gradient estimation to be unbiased.

Lemma 2. *(FD is optimal for $K = 1$) With $K = 1$, there exists a gradient estimator that satisfies condition 1 up to order 1, and such gradient estimator is unique up to scaling σ. In addition, with $K = 1$, no gradient estimator can satisfy condition 1 up to order 2.*

Proof. With simple algebra, one could verify $\alpha_0 = -\alpha_1, \alpha_1 = 1/\beta_1$. FD estimator is obtained by choosing $\beta_1 = 1$. □

Lemma 3. *(AT is optimal for $K = 2$ and fixed $\alpha_0 = 0$) With $K = 2$, there exists a gradient estimator that satisfies condition 1 (with $\alpha_0 = 0$ fixed) up to order 2, and such gradient estimator is unique up to scaling σ. In addition, with $K = 2$, no gradient estimator can satisfy condition 1 (with α_0 not fixed) up to order 3.*

Proof. With simple algebra, one could verify $\alpha_1 = 1/2\beta_1, \alpha_2 = -1/2\beta_1, \beta_2 = -\beta_1$. AT estimator is obtain by choosing $\beta_1 = 1$. □

Remark 3. *(Generalized AT has no improvement) With $K = 2$ and $\alpha_0 = 0$ not fixed, AT can be generalized (basically adjust α_0 to rescale things) to gradient estimators that satisfy condition 1 up to order 2. But the MSE would not change under the assumption that $F(\cdot)$ is quadratic, and the third order term would not zero out if $F(\cdot)$ is not quadratic.*

Remark 4. *(We should only choose between AT and FD) When we use generalized gradient estimator with $K = 3$, we need to show the third order derivative is large enough to justify the extra query of objective $F(\cdot)$. But the third order derivative is super expensive to compute (the Hessian is already quite expensive). Thus, in practice, we should only choose between AT and FD.*

5 Experiments

Every 20 iterations (or 50 iterations, depending on the cost of Hessian approximation), we compute The gradient approximation

$$\frac{1}{N\sigma} \sum_{i=1}^{N} F(\theta + \sigma \vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)}$$

and hessian approximation

$$\frac{1}{N\sigma^2} \left(\sum_{i=1}^{N} F(\theta + \sigma \vec{\epsilon}^{(i)}) \vec{\epsilon}^{(i)} \vec{\epsilon}^{(i)^\top} \right) - \mathbb{I}_d \frac{1}{N\sigma^2} \left(\sum_{i=1}^{N} F(\theta + \sigma \vec{\epsilon}^{(i)}) \right)$$

Then we compute MSE using Lemma 1

$$\begin{aligned}
\text{MSE}\left(\hat{\nabla}_N^{\text{AT,ort}} F_\sigma(\theta)\right) =& \frac{1}{N} ||\nabla F(\theta)||_2^2 \\
\text{MSE}\left(\hat{\nabla}_N^{\text{FD,ort}} F_\sigma(\theta)\right) =& \frac{1}{N} ||\nabla F(\theta)||_2^2 + \frac{(N+4)\sigma^4}{4N} ||\nabla^2 F(\theta)||_F^2 + \frac{5\sigma^4}{2N} \left(\sum_i \nabla^2 F(\theta)_{ii}^2 \right)
\end{aligned}$$

If

$$\frac{2N}{N+1}\,\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{AT,ort}}F_\sigma(\theta)\right) \geq \mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{FD,ort}}F_\sigma(\theta)\right)$$

we use FD. Else, we use AT.

A Theorem 1 proof

Proof. (proof of Theorem 1)

We derive MSE for antithetic ES gradient estimator

$$\begin{aligned}
&\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{AT,ort}}F_\sigma(\theta)\right)\\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}F^{AT(i)} - \nabla F_\sigma(\theta)\right\|_2^2\right]\\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}F^{AT(i)}\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2\\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|F^{AT(i)}\right\|_2^2\right] + \sum_{i\neq j}\mathbb{E}\left[\left\langle F^{AT(i)}, F^{AT(j)}\right\rangle\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2\\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|F^{AT(i)}\right\|_2^2\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by orthogonality}\\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{2\sigma}(F(\theta+\sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)} - F(\theta-\sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)})\right\|_2^2\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by definition of } F^{AT(i)}\\
&= \frac{1}{N}\mathbb{E}\left[\left\|\frac{1}{2\sigma}(F(\theta+\sigma\epsilon)\epsilon - F(\theta-\sigma\epsilon)\epsilon)\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{since } \vec{\epsilon}^{(i)} \text{ i.i.d.}\\
&= \frac{1}{N}\mathbb{E}\left[\left\|\frac{1}{2\sigma}\vec{\epsilon}\left(F(\theta)+\sigma\nabla F(\theta)^\top\vec{\epsilon}+\frac{\sigma^2}{2}\vec{\epsilon}^\top\nabla^2F(\theta)\vec{\epsilon} - F(\theta)+\sigma\nabla F(\theta)^\top\vec{\epsilon}-\frac{\sigma^2}{2}\vec{\epsilon}^\top\nabla^2F(\theta)\vec{\epsilon}\right)\right\|_2^2\right]\\
&\quad - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by Taylor expansion}\\
&= \frac{1}{N}\mathbb{E}\left[\left\|(\nabla F(\theta)^\top\vec{\epsilon})\vec{\epsilon}\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2
\end{aligned}$$

Similarly, we derive MSE for forward finite difference ES gradient estimator

$$
\begin{aligned}
&\mathrm{MSE}\left(\hat{\nabla}_N^{\mathrm{FD,ort}} F_\sigma(\theta)\right) \\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} F^{FD(i)} - \nabla F_\sigma(\theta)\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} F^{FD(i)}\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2 \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|F^{FD(i)}\right\|_2^2\right] + \sum_{i\neq j}\mathbb{E}\left[\left\langle F^{FD(i)}, F^{FD(j)}\right\rangle\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2 \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|F^{FD(i)}\right\|_2^2\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by orthogonality} \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{\sigma}(F(\theta+\sigma\vec{\epsilon}^{(i)})\vec{\epsilon}^{(i)} - F(\theta)\vec{\epsilon}^{(i)})\right\|_2^2\right]\right) - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by definition of } F^{AT(i)} \\
&= \frac{1}{N}\mathbb{E}\left[\left\|\frac{1}{\sigma}(F(\theta+\sigma\epsilon)\epsilon - F(\theta)\epsilon)\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{since } \vec{\epsilon}^{(i)} \text{ i.i.d.} \\
&= \frac{1}{N}\mathbb{E}\left[\left\|\frac{1}{\sigma}\vec{\epsilon}\left(F(\theta) + \sigma\nabla F(\theta)^\top\vec{\epsilon} + \frac{\sigma^2}{2}\vec{\epsilon}^\top\nabla^2 F(\theta)\vec{\epsilon} - F(\theta)\right)\right\|_2^2\right] \\
&\quad - \|\nabla F_\sigma(\theta)\|_2^2 \quad \text{by Taylor expansion} \\
&= \frac{1}{N}\mathbb{E}\left[\left\|\vec{\epsilon}(\nabla F(\theta)^\top\vec{\epsilon} + \frac{\sigma}{2}\vec{\epsilon}^\top\nabla^2 F(\theta)\vec{\epsilon})\right\|_2^2\right] - \|\nabla F_\sigma(\theta)\|_2^2
\end{aligned}
$$

□