Variants of zeroth-order gradient estimates with improved bounds

Previously, we assumed

$$f \in C^3 \quad (\text{3-times cont diff'ble})$$

& showed

$$E\|\hat{\nabla} f(x) - \nabla f(x)\| \le C_1 d^2$$

Lets fix the unified estimate

Recall

$$\hat{\nabla} f(x) = \left[\frac{f(x+\delta v) - f(x-\delta v)}{2\delta}\right] v + \left[\frac{\xi^+ - \xi^-}{2\delta}\right] v$$

Assume
$E[\xi^+ - \xi^- | v] = 0$
$E[vv^T] = I$

Variant 1 : f is convex + L-Smooth

$$\frac{\delta \nabla f(x)^T u}{2\delta} \overset{\text{convex}}{\leq} \frac{f(x+\delta u) - f(x)}{2\delta} \overset{\text{L-Smooth}}{\leq} \frac{\delta \nabla f(x)^T u + \frac{L}{2}\delta^2 \|u\|^2}{2\delta}$$

A similar inequality holds for $f(x-\delta u)$

$$\nabla f(x)^T u - \frac{L\delta\|u\|^2}{2} \leq \frac{f(x+\delta u) - f(x-\delta u)}{2\delta} \leq \nabla f(x)^T u + \frac{L\delta\|u\|^2}{2}$$

Let $\phi(x,\delta,u) = \frac{1}{\delta}\left[\frac{f(x+\delta u) - f(x-\delta u)}{2\delta} - \nabla f(x)^T u\right]$

Then, $|\phi(x,\delta,u)| \leq \frac{L\|u\|^2}{2}$

$$\mathbb{E}\hat{\nabla} f(x) = \mathbb{E}\left[v \left(\frac{f(x+\delta u) - f(x-\delta u)}{2\delta} \right) \right]$$

$$= \mathbb{E}\left[v u^T \nabla f(x) + \delta \phi(x, \delta, u) v \right]$$

Using $\mathbb{E}[v u^T] = I$,

$$\mathbb{E}\|\hat{\nabla} f(x) - \nabla f(x)\| \le \delta \mathbb{E}\|\phi(\text{—}) v\|$$

$$\le \frac{\delta L}{2} \mathbb{E}\|v u^2\|$$

Assume $\mathbb{E}\|v u^2\| \le C_4$

Then,

$$\mathbb{E}\|\hat{\nabla} f(x) - \nabla f(x)\| \le \text{const} \times \delta$$

For the claim,

$$\mathbb{E}\|\hat{\nabla} f(x) - \mathbb{E}[\hat{\nabla} f(x) | x]\|^2 \le \frac{\text{Const}}{\delta^2},$$

the proof is as in the case $f \in C^3$.

Gaussian Smoothing

"Nesterov Spokoiny, 2017"

Goes back to "Katkovnik-Kulchitsky, 1972",

See also Kushner-Clark, 1978

Aka "Gaussian smoothed functional".

$$\hat{\nabla} f(x) = \Delta \left[\frac{\left(f(x+\delta\Delta)+\xi^{+}\right) - \left(f(x)+\xi\right)}{\delta} \right]$$

$\Delta \rightarrow$ vector of $N(0,1)$

$$\Delta \sim N(0, I_d)$$

$$E(\xi^{+}) = E(\xi) = 0$$

Then,

$$\| E \, \hat{\nabla} f(x) - \nabla f(x) \| \leq C_1 \delta$$

Pf:

"Smoothed functional"

$$f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int_{-\infty}^{\infty} f(x + \delta u) \exp\left(-\frac{\|u\|^2}{2}\right) du$$

Next lecture

Lecture - 12

Gaussian smoothing: $x \in \mathbb{R}^d$

$$f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int f(x+\delta u)\, e^{-\frac{1}{2}\|u\|^2}\, du$$

why? $$= \frac{1}{(2\pi)^{d/2}\,\delta^d} \int f(y)\, e^{-\frac{1}{2\delta^2}\|y-x\|^2}\, dy$$

$y = x + \delta u$ Jacobian matrix $= \begin{bmatrix} \delta & & 0 \\ & \ddots & \\ 0 & & \delta \end{bmatrix}$

$|\text{Jacobian}| = \delta^d$

$\frac{dy}{\delta^d} = du$

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}\,\delta^{d+2}} \int f(y)\, e^{-\frac{1}{2\delta^2}\|y-x\|^2}\,(y-x)\, dy$$

$$= \frac{1}{(2\pi)^{d/2}\,\delta} \int f(x+\delta u)\, e^{-\frac{1}{2}\|u\|^2}\, u\, du$$

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x+\delta u) - f(x)}{\delta}\, e^{-\frac{1}{2}\|u\|^2}\, u\, du$$

used $\int u\, e^{-\frac{1}{2}\|u\|^2}\, du = 0$

Also, $\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x+\delta u) - f(x-\delta u)}{2\delta}\, e^{-\frac{1}{2}\|u\|^2} u \, du$

Now,

$\| \nabla f_\delta(x) - \nabla f(x) \|$

$\leq \frac{1}{(2\pi)^{d/2}\delta} \int | f(x+\delta u) - f(x) - \delta \langle \nabla f(x), u \rangle | \, \|u\| \, e^{-\frac{1}{2}\|u\|^2} du$

used $\frac{1}{(2\pi)^{d/2}} \int \langle \nabla f(x), u \rangle \, e^{-\frac{1}{2}\|u\|^2} u \, du = \nabla f(x)$

Since $\frac{1}{(2\pi)^{d/2}} \int u u^T e^{-\frac{1}{2}\|u\|^2} du = I$

$\leq \frac{1}{(2\pi)^{d/2}} \times \frac{\delta L}{2} \int \|u\|^3 e^{-\frac{1}{2}\|u\|^2} du$

using $|f(y) - f(x) - \langle \nabla f(x), y-x \rangle| \leq \frac{1}{2} L \|x-y\|^2$

$$\leq \frac{\delta L}{2} (d+3)^{3/2}$$

using $\frac{1}{(2\pi)^{d/2}} \int \|u\|^3 e^{-\frac{1}{2}\|u\|^2} \, du = (d+3)^{3/2}$

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2} \delta^{d+2}} \int f(y) \, e^{-\frac{1}{2\delta^2}\|y-x\|} (y-x) \, dy$$

$$= \frac{1}{(2\pi)^{d/2} \delta} \int f(x+\delta u) \, e^{-\frac{1}{2}\|u\|^2} u \, du$$

$$\nabla f_\delta(x) = \frac{1}{\delta c} \int f(x+\delta u) \, e^{-1/2\|u\|^2} u \, du$$

$u = -v$, $du = -dv$

$$\nabla f_\delta(x) = \frac{1}{\delta c} \int f(x-\delta v) \, e^{-1/2\|v\|^2} v \, dv$$

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x+\delta u) - f(x)}{\delta} e^{-\frac{1}{2}\|u\|^2} u \, du$$

How?

$$= \frac{1}{(2\pi)^{d/2}} \int \frac{f(x) - f(x-\delta u)}{\delta} e^{-\frac{1}{2}\|u\|^2} u \, du$$

$x + \delta u = v$, $\qquad x = v - \delta u$

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(v) - f(v-\delta u)}{\delta} e^{-\frac{1}{2}\|u\|^2} u \, du$$

A Calculation:

$$\frac{1}{(2\pi)^{d/2}} \int_{-\infty}^{\infty} \nabla f(x)^T u \exp\left(-\frac{\|u\|^2}{2}\right) u \, du$$

$$= \sum_{i=1}^{d} \nabla_i f(x) \frac{1}{(2\pi)^{d/2}} \int_{-\infty}^{\infty} u_i \exp\left(-\frac{\|u\|^2}{2}\right) u \, du$$

$$= \sum_{i=1}^{d} \nabla_i f(x) \frac{1}{(2\pi)^{d/2}}$$

$$\times \int_{-\infty}^{\infty} (u_1 u_i, u_2 u_i, \ldots u_{i-1} u_i, u_i^2, u_{i+1} u_i, \ldots u_i u_d)$$

$$\times \exp\left(-\frac{\|u\|^2}{2}\right) du$$

$$= \sum_{i=1}^{d} \nabla_i f(x) \frac{1}{(2\pi)^{d/2}} \int u_i^2 \exp\left(-\frac{\|u\|^2}{2}\right) du$$

$$\left(\left(\text{we } \underset{i \neq j}{\mathcal{L}}\right) \int u_i u_j \exp\left(-\frac{\|u\|^2}{2}\right) du = 0 \right)$$

$$= \sum_{i=1}^{d} \nabla_i f(x) \frac{1}{(2\pi)^{d/2}} \left(\prod_{j \neq i} \int \exp\left(-\frac{u_j^2}{2}\right) du_j \right)$$

$$\times \int u_i^2 \exp\left(-\frac{u_i^2}{2}\right) du_i$$

$$= \nabla f(x)$$

Main point:

$$\nabla f_\delta(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x+\delta u) - f(x)}{\delta} e^{-\frac{1}{2}\|u\|^2} u \, du$$

$$= E\left(\left(\frac{f(x+\delta u) - f(x)}{\delta} \right) u \right)$$

An estimate of $\nabla f_\delta(x)$ is $u_1 \sim \mathcal{N}(0, I_d)$

$$\hat{g} = \frac{(f(x+\delta u_1) - f(x))}{\delta} u_1$$

$$E \hat{g} = \nabla f_\delta(x)$$

$$\| \nabla f_\delta - \nabla f \| \leq (\text{Const}) \times \delta$$

Lecture-13

Improved variance bound:

Setting: $f(x) = E_{\xi}[F(x,\xi)]$

objective — noise — Sample performance

Assume: "F is L-smooth", i.e.,

For any ξ,

$$\|\nabla F(x_1,\xi) - \nabla F(x_2,\xi)\| \le L\|x_1 - x_2\| \quad \text{a.s.}$$

Suppose we can fix "ξ" (noise) across two function evaluations:

ξ same across evals.

$$\hat{\nabla} f(x) = V\left(\frac{F(x+\delta U,\xi) - F(x-\delta U,\xi)}{2\delta}\right)$$

Similar to unified estimate

Claim: If F is L-smooth then f is L-smooth.

The converse is not true.

Try $\|\nabla E F(x_1,\xi) - \nabla E F(x_2,\xi)\| \longrightarrow$ proceed!

Bias/Variance guarantees:

See Sec 3.3.4 of the book

Assume f is convex & F is L-smooth.

Then, $\| E\, \hat{\nabla} f(x) - \nabla f(x) \| \le C_1 \delta$ and

$$E \, \| \hat{\nabla} f(x) - E\, \hat{\nabla} f(x) \|^2 \le C_2 + C_3 \delta^2$$

Proof. As in the proof of Proposition 3.2, for any convex function h with an L-Lipschitz gradient, for any $\delta > 0$, we have

$$\frac{\langle \nabla h(\theta), \delta u \rangle}{2\delta} \le \frac{h(\theta + \delta u) - h(\theta)}{2\delta} \le \frac{\langle \nabla h(\theta), \delta u \rangle + (L/2) \|\delta u\|^2}{2\delta}.$$

Using similar inequalities for $h(\theta - \delta u)$, we obtain

$$\langle \nabla h(\theta), u \rangle - \frac{L\delta \|u\|^2}{2} \le \frac{h(\theta + \delta u) - h(\theta - \delta u)}{2\delta} \le \langle \nabla h(\theta), u \rangle + \frac{L\delta \|u\|^2}{2}.$$

Letting $\phi(\theta, \delta, u) := \frac{1}{\delta}\left(\frac{h(\theta + \delta u) - h(\theta - \delta u)}{2\delta} - \langle \nabla h(\theta), u \rangle\right)$, we get

$$|\phi(\theta, \delta, u)| \le \frac{L}{2} \|u\|^2 .$$

Using $\mathbb{E}\left[VU^\top\right] = I$, we obtain

$$\mathbb{E}\left[V\left(\frac{h(\theta + \delta U) - h(\theta - \delta U)}{2\delta}\right)\right] = \mathbb{E}\left[VU^\top \nabla h(\theta) + \delta \phi(\theta, \delta, U) V\right]$$

$$= \nabla h(\theta) + \delta \phi(\theta, \delta),$$

where $\widehat{\phi}(\theta, \delta)$ satisfies $\left\| \widehat{\phi}(\theta, \delta) \right\| \leq \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

Applying the above expression to $F(\cdot, \psi)$ and using (3.23), we have

$$\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] = \nabla F(\theta, \psi) + \delta \widehat{\phi}(\theta, \delta) \text{ a.s.},$$

where $\widehat{\phi}(\theta, \delta)$ satisfies $\left\| \widehat{\phi}(\theta, \delta) \right\| \leq \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

A3.4 together with dominated convergence theorem leads to $E[\nabla F(\theta, \psi)] = \nabla f(\theta)$. Using this fact, we obtain

$$\begin{aligned}
&\left\| \mathbb{E}\left[\widehat{\nabla} f(\theta)\right] - \nabla f(\theta) \right\| \\
&= \left\| \mathbb{E}\left[V \left(\frac{f(\theta + \delta U) - f(\theta - \delta U)}{2\delta} \right) - V U^\top \nabla f(\theta) \right] \right\| \\
&\leq \delta \left\| \mathbb{E}[V \phi(\theta, \delta, U)] \right\| \\
&\leq \frac{\delta L}{2} \mathbb{E}[\|V\| \, \|U\|^2],
\end{aligned}$$

and the claim for the bias follows by setting $C_1 = \frac{L}{2} \mathbb{E}[\|V\| \, \|U\|^2]$.

We now bound $\mathbb{E}\left[\left\| \widehat{\nabla} f(\theta) \right\|^2\right]$ as follows:

Why?

$$\begin{aligned}
\mathbb{E}\left\| \widehat{\nabla} f(\theta) \right\|^2 &= \mathbb{E}\left\| V \left(\delta \phi(\theta, \delta, U) + U^\top \nabla f(\theta) \right) \right\|^2 \\
&\leq \mathbb{E}\left[\left(\|V U^\intercal \nabla f(\theta)\| + \frac{\delta L}{2} \|V\| \, \|U\|^2 \right)^2 \right] \\
&\leq 2\mathbb{E}\left[\|V U^\intercal \nabla f(\theta)\|^2 \right] + \frac{\delta^2 L^2}{2} \mathbb{E}\left[\|V\|^2 \, \|U\|^4 \right],
\end{aligned}$$

and the claim for the variance follows by setting $C_2 = 2B_1^2 + \frac{L^2}{2} \mathbb{E}\left[\|V\|^2 \, \|U\|^4 \right]$ with $B_1 = \sup_\theta \|\nabla f(\theta)\|$. □

$$\mathbb{E}\left\|\widehat{\nabla} f(\theta)\right\|^2 \overset{\text{why?}}{=} \mathbb{E}\left\|V\left(\delta\phi(\theta,\delta,U) + U^\top \nabla f(\theta)\right)\right\|^2 \qquad \text{— } (**)$$

$$\le \mathbb{E}\left[\left(\|VU^\intercal \nabla f(\theta)\| + \frac{\delta L}{2}\|V\|\,\|U\|^2\right)^2\right]$$

$$\le 2\mathbb{E}\left[\|VU^\intercal \nabla f(\theta)\|^2\right] + \frac{\delta^2 L^2}{2}\mathbb{E}\left[\|V\|^2\,\|U\|^4\right],$$

$$\mathbb{E}\left[\widehat{\nabla} f(\theta)\right] = \nabla F(\theta,\psi) + \delta\widehat{\phi}(\theta,\delta)$$

$$\mathbb{E}\left[V\left(\frac{h(\theta+\delta U) - h(\theta-\delta U)}{2\delta}\right)\right] = \mathbb{E}\left[VU^\top \nabla h(\theta) + \delta\phi(\theta,\delta,U)V\right] \qquad \text{— } (*)$$

A similar claim with improved variance holds without assuming convexity.