ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 4-1: Zeroth-Order Methods with Random Directions of Gradient Estimations

Jia (Kevin) Liu

Associate Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Autumn 2024

Outline

In this lecture:

- Overview of Zeroth-Order Methods and Their Applications
- Representative Techniques for Random Directions of Gradient Estimations
- Convergence Results

Overview of Zeroth-Order Methods

- Zeroth-order (gradient free) method: Use only function values
  - Reinforcement learning [Malik et al., AISTATS'20]
  - Blackbox adversarial attacks on DNN [Papernot et al., CCS'17]
  - Or problems with structure making gradients difficult or infeasible to obtain

- Two major classes of zeroth-order methods
  - Methods that do not have any connections to gradient
    - Random search algorithm [Schumer and Steiglitz, TAC'68]
    - Nelder-Mead algorithm [Nelder and Mead, Comp J. '65]
    - Model-based methods [Conn et al., SIAM'09]
    - Stochastic three points methods (STP) [Bergou et al., SIAM J. Opt. '20]
    - STP with momentum [Gorbunov et al., ICLR'20]
  - Methods that rely on gradient estimations
    - More modern approach, the focus of this course

Reinforcement Learning

Model-based RL

MDP: $P(s_{t+1}, s_t, a_t)$ $\quad r(s_t, a_t)$

Policy iteration $\pi_\theta(s,a)$

Value iteration $V(s)$

Dynamic Prog. &

Bellman Optimality Principle.

Opt. Ctrl: HJB

CT, DT, Linear/Nonlinear.

$\underline{x}(t+1) = f(\underline{x}(t), \underline{u}(t)) + n$

$\dot{\underline{x}} = f(\underline{x}, \underline{u}) + n.$

Actor-Critic

$Q(s,a) = V(s) + A(s,a)$

DQN

Deep Policy Netw.

Deep RL

Model-free RL

Gradient-free

TD (temporal diff).

Bellman: $V(s) = \mathbb{E}[r_t + \gamma V(s_{t+1})]$

TD(0): $r_t + V(s_{t+1}) - V(s_t)$

⋮

TD(n): — — — —

TD(λ): $R_\Sigma^\lambda - V(s_t)$

TD learning on Q-fn: $Q(s_t, a_t)$

Off-Policy (Q-learning)

exp. vs. exploit

On-Policy (SARSA).

Gradient-based Approach.

$\theta^{new} = \theta^{old} + \delta \nabla_\theta R_{\Sigma,\theta}$

Surrogate Obj: TRPO, PPO ...

Ex: Discrete-time Linear-Quadratic Regulator (LQR)

$(\underline{s}_t, \underline{a}_t)$: $\quad \underline{s}_t \in \mathbb{R}^m, \ \underline{a}_t \in \mathbb{R}^k$

$$\begin{cases} c_t = \underline{s}_t^T \underline{\underline{Q}}\, \underline{s}_t + \underline{a}_t^T \underline{\underline{R}}\, \underline{a}_t \\ \Sigma^{-\frac{1}{2}}\underline{s}_{t+1} = \Sigma^{-\frac{1}{2}}\underline{\underline{A}}\, \underline{s}_t + \Sigma^{-\frac{1}{2}}\underline{\underline{B}}\, \underline{a}_t + \Sigma^{-\frac{1}{2}}\underline{v} \end{cases}$$

$\underline{s}'_{t+1} \quad A' \quad B' \quad v'$

$(\underline{\underline{Q}}, \underline{\underline{R}} \succ 0)$ — $m\times m$, $k \times k$.

$(\underline{\underline{A}}, \underline{\underline{B}})$ — $m\times m$, $m \times k$

w.l.o.g. assume $\underline{v}$ a r. vec. $\underline{v} \sim D$ s.t. $\mathbb{E}[\underline{v}] = \underline{0}$ $\mathbb{E}[\underline{v}\underline{v}^T] = \underline{\underline{I}}$

$\mathbb{E}[\underline{v}\underline{v}^T] = \underline{\underline{\Sigma}}$ $\qquad \underline{\underline{\Sigma}}^{-\frac{1}{2}} \underline{v}\underline{v}^T \underline{\underline{\Sigma}}^{-\frac{1}{2}} = \underline{\underline{I}}$

By classical opt. ctrl theory: $\underline{a}_t = -\underline{\underline{K}}^* \underline{s}_t$, where $\underline{\underline{K}}^*$ can be found by the dt-Riccati eqn. (assuming $\underline{\underline{A}}, \underline{\underline{B}}, \underline{\underline{Q}}, \underline{\underline{R}}$)

If we don't know $\underline{A}, \underline{B}, \underline{R}, \underline{Q}$, we can search over lin. policies:

1. Rand initialization: $C_{init}(\underline{K}, \underline{s}_0) \leftarrow$ cost of executing a lin. policy $\underline{K}$ from $\underline{s}_0$.

$$C_{init}(\underline{K}, \underline{s}_0) = \sum_{t=0}^{\infty} (\underline{s}_t^T \underline{Q}\, \underline{s}_t + \underline{a}_t^T \underline{R}\, \underline{a}_t + \cancel{w_t}) \gamma^t$$

ZO - opt.

Goal: $\min C_{init,\gamma}(\underline{K}) = \mathbb{E}_{s_0 \sim D_0} \left[C_{init,\gamma}(\underline{K}, \underline{s}_0) \right]$

$\underline{K}$ "contrible".

we don't know $\underline{A}, \underline{B}, \underline{Q}, \underline{R}$. Can only observe a noisy fn evaluation of $C_{init}(\underline{K}, s_0)$.

spectral radis.

A policy $\underline{K}$ is said to be ctrlble for $(\underline{A}, \underline{B})$ if $\rho(\underline{A} - \underline{K}\underline{B}) < 1$

$\{\underline{K} : \rho(\underline{A} - \underline{K}\underline{B}) < 1\}$

Note: 1°. LQR is locally Lipschitz

$$|C_{init,\gamma}(\underline{K}', \underline{s}_0) - C_{init,\gamma}(\underline{K}, \underline{s}_0)| \le \lambda \, \|\underline{K}' - \underline{K}\|_F$$

2°. LQR has locally Lipschitz cont. grad.

$$\|\nabla C_{init,\gamma}(\underline{K}') - \nabla C_{init,\gamma}(\underline{K})\|_F \le \phi \, \|\underline{K}' - \underline{K}\|_F.$$

3°. $C_{init,\gamma}(\underline{K})$ is nonconvex: $\{K : \rho(\underline{A} - \underline{K}\underline{B}) < 1\}$ is nonconvex.

4°. LQR is PL.

Basic Idea of (Deterministic) Gradient Estimation

- Gradient estimation with finite-difference directional derivative estimation:

$$\text{(Forward version): } \mathbf{g}(\mathbf{x}) = \sum_{i=1}^{d} \frac{f(\mathbf{x}+\mu\mathbf{e}_i) - f(\mathbf{x})}{\mu}\mathbf{e}_i,$$

$$\text{(Centered version): } \mathbf{g}(\mathbf{x}) = \sum_{i=1}^{d} \frac{f(\mathbf{x}+\mu\mathbf{e}_i) - f(\mathbf{x}-\mu\mathbf{e}_i)}{2\mu}\mathbf{e}_i,$$

  where $\mathbf{e}_i$ is the i-th natural basis vector of $\mathbb{R}^n$ and μ is the sampling radius

  cont. diff. FO: L-Lip. grad.

- For the gradient estimation above, it can be shown that for $f \in C_L^{1,1}$ (i.e., continuously differentiable with Lipschitz-continuous gradient)

  Nesterov notation

$$\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 \le \mu L \sqrt{d}$$

- Natural idea: Replace actual gradient with gradient estimation in any first-order optimization scheme (deterministic ZO methods)
  - Pro: Use Lipschitz-like bound above to characterize convergence performance
  - Con: Suffer from problem dimensionality for large d ($O(d)$ ZO-oracle calls)

Randomized Gradient Estimation

- Two-point random gradient estimator

$$\hat{\nabla} f(\mathbf{x}) = (d/\mu)[f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x})]\mathbf{u},$$

where $\mathbf{u}$ is an i.i.d. random direction

- $(q+1)$-point random gradient estimator

$$\hat{\nabla} f(\mathbf{x}) = (d/(\mu q)) \sum_{i=1}^{q} [f(\mathbf{x}+\mu\mathbf{u}_i) - f(\mathbf{x})]\mathbf{u}_i,$$

which is also referred to as average random gradient estimator

- Benefits:
  - Make every iteration simpler
  - Easy convergence proof
  - For problems limited to only several (or even one) ZO oracle queries

Formalization of Stochastic Zeroth-Order Methods

- Consider the problem of the following form:

$$\min_{\mathbf{x} \in Q \subseteq \mathbb{R}^d} f(\mathbf{x})$$

- A stochastic ZO method generates $\{\mathbf{x}_k\}$ as follows:

$$\mathbf{x}_{k+1} = \mathcal{A}\left(\hat{f}, \mathbf{X}, P, \{\mathbf{x}_i\}_{i=0}^k, \{\mathbf{u}_i\}_{i=0}^k\right)$$

  - $\hat{f}$: ZO-oracle (could be noisy, i.e., $\hat{f}$ is not necessarily equal to f; e.g., $\hat{f}(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$ or $\hat{f}(\mathbf{x}, \mathbf{u}) = f(\mathbf{x}) + \epsilon(\mathbf{x}, \mathbf{u})$ with $\mathbb{E}_{\mathbf{u}}[\hat{f}(\mathbf{x}, \mathbf{u})] = f(\mathbf{x})$)
  - $\{\mathbf{x}_i\}_{i=0}^k$: history of $\mathbf{x}$-variables
  - $\{\mathbf{u}_i\}_{i=0}^k$: random sampling directions
  - P: parameters (dimension d of $\mathbf{x}$, L-Lipschitz constant, etc.)

- This lecture: Focus on non-convex objective function

Random Directions Gradient Estimations

- Consider the following ZO scheme using gradient approximation:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}(\mathbf{x}_k, \mathbf{u}_k),$$

where $\mathbf{g}(\mathbf{x}_k, \mathbf{u}_k)$ follows the two-point random gradient estimator:

$$\mathbf{g}(\mathbf{x}_k, \mathbf{u}_k) = \frac{\hat{f}(\mathbf{x}_k + \mu \mathbf{u}_k) - \hat{f}(\mathbf{x}_k)}{\mu} \mathbf{u}_k$$

- It makes sense to use centrally symmetric distributions for $\mathbf{u}_k$:
  - Uniformly distributed over unit Euclidean sphere [Flaxman et al. SODA'05], [Gorbunov et al. SIOPT'18], [Dvurechensky et al., E. J. OR'21]:

$$\mathbf{u}_k \sim \mathcal{U}\{S^{d-1}\}, \text{ where } S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$$

  - Gaussian smoothing [Nesterov and Spokoiny, Math Prog.'06]:

$$\mathbf{u}_k \sim \mathcal{N}(0, \mathbf{I}_d)$$

Gaussian Smoothing [Nesterov and Spokoiny, FCM'17]

- Gaussian smoothing approximation:

$$f_\mu(\mathbf{x}) = \frac{1}{\kappa}\int_{\mathbb{R}^d} f(\mathbf{x}+\mu\mathbf{u})e^{-\frac{1}{2}\|\mathbf{u}\|_2^2}d\mathbf{u},$$

where $\kappa = \int_{\mathbb{R}^d} e^{-\frac{1}{2}\|\mathbf{u}\|_2^2}d\mathbf{u} = (2\pi)^{d/2}$.

- Good properties:
  - Convexity preservation: If f is convex, so is f_μ
  - Differentiability
  - If $f \in C_{L_0}^{0,0}$ (or $f \in C_{L_1}^{1,1}$), the same holds for f_μ with $L_0(f_\mu) \leq L_0(f)$ (or $L_1(f_\mu) \leq L_1(f)$)
  - $|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \mu L_0\sqrt{d}$ if $f \in C_{L_0}^{0,0}$

Gaussian Smoothing [Nesterov and Spokoiny, FCM'17]

- Consider the following algorithm:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}(\mathbf{x}_k, \mathbf{u}_k), \text{ and } \mathbf{u}_k \sim \mathcal{N}(0, \mathbf{I}_d).$$

- For nonconvex $f \in C^{1,1}_{L_1}$, we have (let $U = \{\mathbf{u}_k\}_{k=0}^{K-1}$):

$$\|\nabla f(\mathbf{x})\| \leq \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_U[\|\nabla f_\mu(\mathbf{x}_k)\|_2^2] \leq 8\underbrace{(d+4)}_{O(d)} L_1 \left[\frac{f_\mu(\mathbf{x}_0) - f^*}{K} + \overbrace{\frac{3\mu^2(d+4)}{32}}^{=O(d)} L_1\right]$$

$O(\frac{1}{K})$

to a ball.

- Using the facts that $\|f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 \leq \frac{\mu L_1}{2}(d+3)^{\frac{3}{2}}$ and $\|\nabla f(\mathbf{x})\|_2^2 \leq 2\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 + 2\|\nabla f_\mu(\mathbf{x})\|_2^2$, we obtain:

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_U[\|\nabla f(\mathbf{x}_k)\|_2^2] \leq 2\frac{\mu^2 L_1^2}{4}\underbrace{(d+3)^3}_{O(d^3)}$$

$$+ 16(d+4)L_1\left[\frac{f_\mu(\mathbf{x}_0) - f^*}{K} + \frac{3\mu^2(d+4)}{32}L_1\right]$$

Gaussian Smoothing [Nesterov and Spokoiny, FCM'17]

- Choosing $\mu = O(\epsilon/[d^3 L_1])$ ensures $\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_U[\|\nabla f(\mathbf{x}_k)\|_2^2] \le \epsilon^2$, which implies the following sample complexity:

$$K = O(d\epsilon^{-2}). \quad \approx \text{GD}.$$

- For $f \in C_{L_0}^{0,0}$, we have (let $S_K = \sum_{k=0}^{K-1} s_k$):

$$\frac{1}{S_K}\sum_{k=0}^{K-1} s_k \mathbb{E}_U[\|\nabla f_\mu(\mathbf{x}_k)\|_2^2] \le \frac{1}{S_K}\left[(f_\mu(\mathbf{x}_0) - f^*) + \frac{1}{\mu} d^{\frac{1}{2}} (d+4)^2 L_0^3 \sum_{k=0}^{K-1} s_k^2\right]$$

- Consider a bounded domain Q with $\mathrm{diam}(Q) \le R$. To ensure $\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_U[\|\nabla f_\mu(\mathbf{x}_k)\|_2^2] \le \epsilon^2$ and $|f_\mu(\mathbf{x}) - f(\mathbf{x})| \le \delta$, we have the following sample complexity:

$$K = O\left(\frac{d(d+4)^2 L_0^5 R}{\epsilon^4 \delta}\right). \quad O\left(\frac{d^3}{\varepsilon^4}\right)$$

- If $s_k \to 0$ and $\mu \to 0$, convergence of $\mathbb{E}_U[\|\nabla f(\mathbf{x}_k)\|_2]$ can also be proved.

Extensions of Gaussian Smoothing to Noisy $\hat{f}$

Consider the following:

- Noisy $\hat{f}$: $|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \le \delta$ RL :
- Hölder continuous gradient (intermediate smoothness)

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \le L_\nu \|\mathbf{x} - \mathbf{y}\|_2^\nu, \text{ for some } \nu \in [0,1],$$

which implies the following generalized descent lemma:

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L_\nu}{1+\nu} \|\mathbf{y} - \mathbf{x}\|^{1+\nu}$$

- To ensure $\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_U[\|\nabla f(\mathbf{x}_k)\|_2^2] \le \epsilon^2$, we have the following sample complexity [Shibaev et al., Opt. Lett. '21]:

$\nu\uparrow$

$$K = O\left(\frac{d^{2+\frac{1-\nu}{2\nu}}}{\epsilon^{\frac{2}{\nu}}}\right) \text{ if } \delta = O\left(\frac{\epsilon^{\frac{3+\nu}{2\nu}}}{d^{\frac{3+7\nu}{4\nu}}}\right).$$

Extensions of Gaussian Smoothing to Noisy $\hat{f}$

- Special case of $\nu = 1$ (i.e., $f \in C^{1,1}_{L_1}$): Sample complexity is improved to

$$K = \underline{O(d\epsilon^{-2})}, \qquad \text{Similar GD.}$$

same as

which is ~~d times better than~~ [Nesterov and Spokoiny, FCM'17]

- If $|\hat{f}(\mathbf{x}) - f(\mathbf{x})| \le \epsilon_f$, where f is convex and 1-Lipschitz and $\epsilon_f \sim \max\{\epsilon^2/\sqrt{d}, \epsilon/d\}$, then [Risteski and Li, NeurIPS'16] showed that there exists an algorithm that finds ϵ-optimal solution (i.e., $\hat{f}(\mathbf{x}) - \hat{f}^* \le \epsilon$) with sample complexity Poly(d, ϵ^{-1}). Also, the dependence $\epsilon_f(\epsilon)$ is optimal

LB-matching

Randomized Stochastic Gradient-Free Methods

If $s_k = \frac{1}{\sqrt{T}}$, $\forall k$. $O(\frac{1}{\sqrt{T}})$.

Gaussian smoothing is extended to [Ghadimi and Lan, SIAM J. Opt. '13] [Ghadimi et al., Math Prog. '16] (unconstrained case, i.e., $Q = \mathbb{R}^d$):

- $\hat{f} = F(\mathbf{x}, \xi)$ such that $\mathbb{E}_\xi[F(\mathbf{x}, \xi)] = f(\mathbf{x})$, where ξ is a random variable whose distribution P is supported on $\Xi \subseteq \mathbb{R}^d$
- $F(\cdot, \xi)$ has L_1-Lipschitz continuous gradient
- Consider the following randomized stochastic gradient-free method (RSGF):

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k G(\mathbf{x}_k, \xi_k, \mathbf{u}_k),$$

$$G(\mathbf{x}_k, \xi_k, \mathbf{u}_k) = \frac{\hat{f}(\mathbf{x}_k + \mu \mathbf{u}_k, \xi_k) - \hat{f}(\mathbf{x}_k, \xi_k)}{\mu} \mathbf{u}_k$$

- It follows from $\mathbb{E}_\xi[F(\mathbf{x}, \xi)] = f(\mathbf{x})$ that $\mathbb{E}_{\xi, \mathbf{u}}[G(\mathbf{x}, \xi, \mathbf{u})] = \nabla f_\mu(\mathbf{x})$
- Similar to FO-SGD in [Ghadimi and Lan, SIAM J. Opt. '13], RSGF chooses $\mathbf{x}_R$ from $\{\mathbf{x}_k\}_{k=1}^K$ where R is a r.v. with p.m.f. P_R supported on $\{1, \ldots, K\}$

random termination index

Randomized Stochastic Gradient-Free Methods

- For $f \in C^{1,1}_{L_1}$, smoothing parameter μ, $D_f = (2(f(\mathbf{x}_1) - f^*)/L)^{\frac{1}{2}}$, and $\mathbb{E}_\xi[\|\nabla \hat{f}(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2] \leq \sigma^2$ and p.m.f. of R being:

$\mathbb{E}\frac{1}{T}\sum\|\nabla f(\cdot)\|^2$

$$P_R(k) = \frac{s_k - 2L(d+4)s_k^2}{\sum_{i=1}^K (s_i - 2L(d+4)s_i^2)},$$

it then holds that:

$$\frac{1}{L_1}\mathbb{E}[\|\nabla f(\mathbf{x}_R)\|_2^2] \leq \frac{1}{\sum_{k=1}^K [s_k - 2L_1(d+4)s_k^2]} \Bigg[D_f^2 + 2\mu^2(d+4) \times \left(1 + L_1(d+4)^2 \sum_{k=1}^K (\frac{s_k}{4} + Ls_k^2)\right) + 2(d+4)\sigma^2 \sum_{k=1}^K s_k^2 \Bigg],$$

where the expectation is taken w.r.t. R and $\{\xi_k\}$.

Randomized Stochastic Gradient-Free Methods

$O(d)$

- Choose constant step-size $s_k = \frac{1}{\sqrt{d+4}} \min\{\frac{1}{4L\sqrt{d+4}}, \frac{\tilde{D}}{\sigma\sqrt{K}}\}$ for some $\tilde{D} > 0$ (some estimation of D_f):

$$\frac{1}{L_1}\mathbb{E}[\|\nabla f(\mathbf{x}_R)\|_2^2] \le \underbrace{\frac{12(d+4)L_1 D_f^2}{K}}_{O(\frac{d}{K})} + \underbrace{\frac{2\sigma\sqrt{d+4}}{\sqrt{K}}}_{O(\sqrt{\frac{d}{K}})}\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)$$

- To ensure $\Pr\{\|\nabla f(\mathbf{x}_R)\|_2^2 \le \epsilon\} \ge 1 - \delta$ (i.e., (ϵ, δ)-solution), the zeroth-order oracle sample complexity is: w.h.p. result.

$$O\left(\frac{dL_1^2 D_f^2}{\delta\epsilon} + \frac{dL_1^2}{\delta^2}\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma^2}{\epsilon^2}\right)$$

$O(\delta^{-2})$. $\log(\frac{1}{\delta})$

Randomized Stochastic Gradient-Free Methods

Two-phase randomized stochastic gradient-free method (2-RSGF) [Ghadimi and Lan, SIAM J. Opt. '13]

- Run RSGF $S = \log(1/\delta)$ times as a subroutine producing a list $\{\bar{\mathbf{x}}_k\}_{k=1}^S$
- Output point $\bar{\mathbf{x}}^*$ is chosen in such a way that:

$$\|\mathbf{g}(\bar{\mathbf{x}}^*)\|_2 = \min_{s=1,\ldots,S} \|\mathbf{g}(\bar{\mathbf{x}}_s)\|_2, \text{ where } \mathbf{g}(\bar{\mathbf{x}}_s) = \frac{1}{T}\sum_{k=1}^{T} G_\mu(\bar{\mathbf{x}}_s, \xi_k, \mathbf{u}_k),$$

  where $G_\mu(\bar{\mathbf{x}}_s, \xi_k, \mathbf{u}_k)$ is defined as in RSGF
- The zeroth-order oracle sample complexity for achieving (ϵ, δ)-solution:

$$O\left(\frac{dL_1^2 D_f^2 \log(1/\delta)}{\epsilon} + dL_1^2\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)^2 \frac{\sigma^2}{\epsilon^2}\log(1/\delta) + \underbrace{\frac{d\log^2(1/\delta)}{\delta}\left(1 + \frac{\sigma^2}{\epsilon}\right)}_{O(d\epsilon^{-2})}\right)$$

- A more general problem $\min_{\mathbf{x}\in Q\subseteq\mathbb{R}^d} \Psi(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$ is also solved in [Ghadimi et al., Math Prog.'16]
  - $f \in C_L^{1,1}$: nonconvex; $h(\mathbf{x})$ is simple convex and possibly non-smooth

RSGF Based on Uniform Sampling over Unit Sphere

- Consider the problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \mathbb{E}_\xi[F(\mathbf{x},\xi)] = \mathbb{E}_\xi[\hat{f}(\mathbf{x},\xi)]$
  - $f(\mathbf{x})$ is L-Lipschitz and μ-smooth
  - $|F(\mathbf{x},\xi)| \leq \Omega$ and F's variance is bounded by V_f
- Stochastic gradient estimation based on uniform sampling over unit sphere:

$$\mathbf{g}(\mathbf{x}_k,\xi_k,\mathbf{u}_k) = n\frac{\hat{f}(\mathbf{x}_k+\mu\mathbf{u}_k,\xi_k) - \hat{f}(\mathbf{x}_k-\mu\mathbf{u}_k,\xi_k)}{2\mu},$$

  where $\mathbf{u}_k \sim \mathcal{U}(S^{n-1})$. The update process is $\mathbf{x}_{k+1} = \mathbf{x}_k - s\mathbf{g}(\mathbf{x}_k,\xi_k,\mathbf{u}_k)$
- After K steps, we have [Sener and Koltun, ICML'20]:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] = O\left(\frac{n}{K^{1/2}} + \frac{n^{2/3}}{K^{1/3}}\right)$$

$O\left(\frac{1}{K^{1/3}}\right)$

RSGF Based on Uniform Sampling over Unit Sphere

- Consider the case for a given ξ, $F(\mathbf{x}, \xi) = g(r(\mathbf{x}, \theta^*), \Psi^*)$, where $g(\cdot, \Psi)$ and $r(\cdot, \theta)$ are parameterized function classes
  - $r(\cdot, \theta^*) : \mathbb{R}^n \to \mathbb{R}^d$, where $d \ll n$
  - $F(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}$ is actually defined on a d-dimensional manifold $\mathcal{M}$ for all ξ

- Thus, if one knows the manifold (i.e., θ^*) and g and r are smooth, we have from chain rule: $\nabla f(\mathbf{x}) = J(\mathbf{x}, \theta^*) \nabla_r g(r, \Psi)$, where $J(\mathbf{x}, \theta^*) = \frac{\partial r(\mathbf{x}, \theta^*)}{\partial \mathbf{x}}$. This leads to [Sener and Koltun, ICML'20]:

$$G(\mathbf{x}_k, \xi_k, \mathbf{u}_k) = d \frac{\hat{f}(\mathbf{x}_k + \mu J_q \mathbf{u}_k, \xi_k) - \hat{f}(\mathbf{x}_k - \mu J_q \mathbf{u}_k, \xi_k)}{2\mu} \mathbf{u}_k,$$

where J_q is the orthonomalized $J(\mathbf{x}_k, \theta^*)$ and $\mathbf{u}_k \sim \mathcal{U}(S^{d-1})$. It follows that

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_2^2] = O\left(\frac{n^{1/2}}{K} + \frac{n^{1/2} + d + dn^{1/2}}{K^{1/2}} + \frac{d^{2/3} + n^{1/2} d^{2/3}}{K^{1/3}} \right).$$

$O(\frac{1}{K^{1/3}})$

which is much better than the previous bound for $d \leq n^{1/2}$.

Which Gradient Estimation Works Better?

- Gradient estimations with random directions are worse than finite differences in terms of # of samples required to ensure the norm condition:

"SNR"

$$\underbrace{\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 \le \theta \|\nabla f(\mathbf{x})\|_2}, \text{ for some } \theta \in [0, 1)$$

- Gradient estimation methods are studied in [Berahas et al., FCM'21]: Compare the # of calls r (i.e., "batch size") to ensure norm condition

deterministic
"no randomness"

  - FFD (Forward Finite Differences): $\sum_{i=1}^{d} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{e}_i)-\hat{f}(\mathbf{x})}{\mu}\mathbf{e}_i$
  - CFD (Centered Finite Differences): $\sum_{i=1}^{d} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{e}_i)-\hat{f}(\mathbf{x}-\mu\mathbf{e}_i)}{2\mu}\mathbf{e}_i$
  - LI (Linear Interpolation): $\sum_{i=1}^{d} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{u}_i)-\hat{f}(\mathbf{x})}{\mu}\mathbf{u}_i$, $\mathbf{u}_i = [\mathbf{Q}]_i$ non-singular rotation of natural basis.

randomized

(r+1)-pt.

  - GSG (Gaussian Smoothed Gradients): $\frac{1}{r}\sum_{i=1}^{r} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{u}_i)-\hat{f}(\mathbf{x})}{\mu}\mathbf{u}_i$, $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{I}_d)$
  - cGSG (Centered GSG): $\frac{1}{r}\sum_{i=1}^{r} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{u}_i)-\hat{f}(\mathbf{x}-\mu\mathbf{u}_i)}{2\mu}\mathbf{u}_i$, $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{I}_d)$
  - SSG (Sphere Smoothed Gradients): $\frac{d}{r}\sum_{i=1}^{r} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{u}_i)-\hat{f}(\mathbf{x})}{\mu}\mathbf{u}_i$, $\mathbf{u}_i \sim \mathcal{U}(S^{d-1})$
  - cSSG (Centered SSG): $\frac{d}{r}\sum_{i=1}^{r} \frac{\hat{f}(\mathbf{x}+\mu\mathbf{u}_i)-\hat{f}(\mathbf{x}-\mu\mathbf{u}_i)}{2\mu}\mathbf{u}_i$, $\mathbf{u}_i \sim \mathcal{U}(S^{d-1})$

Which Gradient Estimation Works Better?

- Consider an unconstrained problem $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ [Berahas et al., FCM'21]:
  - Noisy ZO oracle: $\hat{f}(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$
  - Noise ϵ is bounded uniformly: $|\epsilon(\mathbf{x})| \leq \epsilon_f$ (noise not neccessarily random)
  - $f(\mathbf{x}) \in C_L^{1,1}$ or $f(\mathbf{x}) \in C_M^{2,2}$ (twice continuously differentiable with M-Lipschitz Hessian)

$\|\nabla f(\underline{x})\|$ is "large enough".

Method	Number of calls r	$\|\nabla f(\mathbf{x})\|_2$
FFD	d	$\frac{2\sqrt{dL\epsilon_f}}{\theta}$
CFD	d	$\frac{2\sqrt{d}\sqrt[3]{M\epsilon_f^2}}{\sqrt[3]{6}\theta}$
LI	d	$\frac{2\|Q^{-1}\|\sqrt{dL\epsilon_f}}{\theta}$
GSG	$\frac{12d}{\sigma\theta^2} + \frac{d+20}{16\delta}$	$\frac{6d\sqrt{L\epsilon_f}}{\theta}$
cGSG	$\frac{12d}{\sigma\theta^2} + \frac{d+30}{144\delta}$	$\frac{12\sqrt[3]{d^{7/2}M\epsilon_f^2}}{\theta}$
SSG	$[\frac{8d}{\theta^2} + \frac{8d}{3\theta} + \frac{11d+104}{24}] \log\frac{d+1}{\delta}$	$\frac{4d\sqrt{L\epsilon_f}}{\theta}$
cSSG	$[\frac{8d}{\theta^2} + \frac{8d}{3\theta} + \frac{9d+192}{27}] \log\frac{d+1}{\delta}$	$\frac{4\sqrt[3]{d^{7/2}M\epsilon_f^2}}{\theta}$

$O(d)$ (GSG, cGSG calls) $\tilde{O}(d)$ (SSG, cSSG calls)

$O(d)$ (GSG, SSG) $O(d^{\frac{7}{6}})$ (cGSG, cSSG)

- LI is essentially FFD with directions given as columns of a nonsingular matrix $\mathbf{Q}$
- For GSG, cGSG, SSG, and cSSG, results are w.p. $1-\delta$

Next Class

Variance-Reduced Zeroth-Order Methods