Optimization for Deep Learning

Lecture 5: Zeroth-order Optimization

Kun Yuan

Peking University

Main contents in this lecture

- Motivation and Application
- Gradient estimation with finite difference
- Gradient estimation with linear interpolation
- Gradient estimation with sphere smoothing

Zeroth-order Optimization

- Consider the following unconstrained problem

$$\min_{x \in \mathbb{R}^d} \quad f(x) \tag{1}$$

- We cannot access $\nabla f(x)$ in many scenarios because
  - The closed-form of $f(x)$ is unknown
  - Computing $\nabla f(x)$ is very expensive
- Can we solve problem (1) without gradient information?

Application I: Black-box adversarial attack

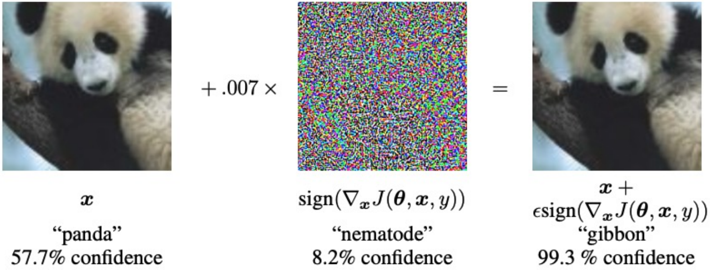

Figure: A demonstration of the adversarial example [Goodfellow et.al., 2015].

Application I: Black-box adversarial attack

- An adversarial example is a perturbation η to maximize misclassification
- Given an input pair (ξ, y), we can attack the neural network by

$$\max_{\eta : \|\eta\| \leq \epsilon} L\Big(h(x^\star, \xi + \eta), y\Big)$$

  where $x^\star$ is the trained DNN model.

- $h(\cdot)$ and $L(\cdot)$ are unknown in black-box scenario; gradient is not accessible
- Black-box model is more common in adversarial attack. White-box model is more common in robust learning.

Application II: Memory-efficient LLM fine-tuning

- Training large language models (LLM) consumes significant memory

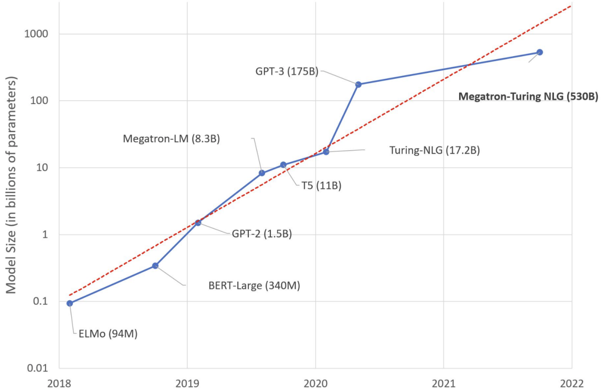

Figure: Large language models (LLM) show a new Moor's law[1].

[1] This plot is from `https://huggingface.co/blog/large-language-models`

Application II: Memory-efficient LLM fine-tuning

- Computing gradient needs forward-propagation and backward-propagation
- Backward-propagation takes much more memory
  - cache activations during the forward pass
  - cache gradients during the backward pass
  - cache gradient history in momemtum update
- Using zeroth-order optimizers can save great memory; A 30B LLM can be trained on a single A100 (Malladi et al., 2023).
- Using first-order optimizers can only train a 2.7B LLM with one A100

Estimate gradient with finite difference

- Let $f(x) : \mathbb{R}^d \to \mathbb{R}$, we have

$$\frac{\partial f(x)}{\partial x_i} = \lim_{\tau \to 0} \frac{f(x + \tau e_i) - f(x)}{\tau}, \quad i = 1, 2, \ldots, d.$$

  where e_i is the i-th column of the identity matrix I.

- The gradient $\nabla f(x) \in \mathbb{R}^d$ is defined as

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_d} \right)^\top = \sum_{i=1}^{d} \frac{\partial f(x)}{\partial x_i} e_i$$

Estimate gradient with finite difference

- We can also estimate gradient with

$$g(x) = \sum_{i=1}^{d} \frac{f(x + \tau e_i) - f(x)}{\tau} e_i$$

which is called **forward difference**; needs $(d+1)$ queries on $f(x)$.

- It is natural to estimate gradient with finite difference

$$g(x) = \sum_{i=1}^{d} \frac{f(x + \tau e_i) - f(x - \tau e_i)}{\tau} e_i$$

which is called **central difference**; needs $2d$ queries on $f(x)$.

- Central difference is typically more accurate than forward difference.

Estimate error with finite difference

Lemma 1

Assume that $f(x)$ is L-smooth. Let $g(x)$ denote the forward finite difference approximation to the gradient $\nabla f(x)$, it holds for all $x \in \mathbb{R}^d$ that

$$\|g(x) - \nabla f(x)\| \leq \frac{\sqrt{d} L \tau}{2}.$$

Large d, L, and τ will reulst in less accuarate estimate.

Zeroth-order gradient descent (ZO-GD)

- With gradient estimated by forward difference, ZO-GD iterates as follows

$$g(x_k) = \sum_{i=1}^{d} \frac{f(x_k + \tau e_i) - f(x_k)}{\tau} e_i$$

$$x_{k+1} = x_k - \gamma g(x_k)$$

Theorem 1 (Non-convex convergence)

Assume $f(x)$ to be L-smooth. If we set $\gamma = 1/L$, ZO-GD converges as follows

$$\frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f(x^\star))}{K+1} + dL^2\tau^2.$$

- ZO-GD converges to a **neighborhood** $O(dL^2\tau^2)$ around the solution

Zeroth-order gradient descent (ZO-GD)

Theorem 2 (Strongly-convex convergence)

Assume $f(x)$ to be L-smooth and μ-strongly convex. If we set $\gamma = 1/L$, ZO-GD converges as follows

$$f(x_K) - f(x^\star) \leq (1 - \frac{\mu}{L})^K (f(x_0) - f(x^\star)) + \frac{dL^2\tau^2}{8\mu}$$

Estimate error with linear interpolation

- Is it possible that we estimate $\nabla f(x)$ with much fewer queries?

  When $\nabla f(x)$ is sparse, Yes!

- Given τ and $u \in \mathbb{R}^d$, Taylor's formular shows that

$$f(x+\tau u) \approx f(x) + \tau u^\top \nabla f(x)$$

  where the approximation becomes assumate when $\|\tau u\|$ is small

- Given $\{u_1, \cdots, u_m\}$ and τ, we have

$$\frac{1}{\tau}\begin{pmatrix} f(x+\tau u_1) - f(x) \\ \vdots \\ f(x+\tau u_m) - f(x) \end{pmatrix} = \begin{pmatrix} u_1^\top \\ \vdots \\ u_m^\top \end{pmatrix} g(x)$$

  where $g(x)$ is the gradient estimate

Estimate error with linear interpolation

- Given $\{u_1, \cdots, u_m\}$ and τ, we have

$$\underbrace{\frac{1}{\tau}\begin{pmatrix} f(x+\tau u_1) - f(x) \\ \vdots \\ f(x+\tau u_m) - f(x) \end{pmatrix}}_{\boldsymbol{b}} = \underbrace{\begin{pmatrix} u_1^\top \\ \vdots \\ u_m^\top \end{pmatrix}}_{\boldsymbol{Q}} \underbrace{g(x)}_{\boldsymbol{g}}$$

- Greadient estimate can be calculated by solving linear equaliion

$$\boldsymbol{Q}\boldsymbol{g} = \boldsymbol{b} \quad \text{where} \quad \boldsymbol{Q} \in \mathbb{R}^{m\times d}$$

- Generally speaking, $\boldsymbol{g}$ can be solved only when $m = d$ and $\boldsymbol{Q}$ is invertible

Estimate error with linear interpolation

- For example, if $\boldsymbol{Q} = I \in \mathbb{R}^{d\times d}$, linear interpolation reduces to finite difference

$$\boldsymbol{g} = \boldsymbol{b} = \frac{1}{\tau}\begin{pmatrix} f(x+\tau u_1) - f(x) \\ \vdots \\ f(x+\tau u_m) - f(x) \end{pmatrix} = \sum_{i=1}^{d} \frac{f(x+\tau e_i) - f(x)}{\tau} e_i$$

Lemma 2

If $f(x)$ is L-smooth and $g(x)$ is acheived via linear interpolation, it holds that

$$\|g(x) - \nabla f(x)\| \le \frac{\|\boldsymbol{Q}^{-1}\|\sqrt{d}Lt}{2}, \quad \forall x \in \mathbb{R}^d$$

It implies that an orthnormal $\boldsymbol{Q}$ is a best choice.

Estimate error with linear interpolation

- It seems that linear interpolation cannot save queries; still needs $O(d)$ queries
- However, if $\nabla f(x)$ is sparse, we need much less queries.
- Recovering sparese gradient reduces to compressive sensing

$$\min_{\boldsymbol{g} \in \mathbb{R}^d} \quad \frac{1}{2} \|\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{g}\|^2 + \lambda \|\boldsymbol{g}\|_1$$

where $m \ll d$.

ZO-GD with sparse linear interpolation

- If the gradient is sparse, ZO-GD with linear interpolation iterates as follows

$$
\begin{aligned}
\boldsymbol{b}^k &= \frac{1}{\tau}\begin{pmatrix} f(x^k+\tau u_1^k)-f(x^k) \\ \vdots \\ f(x^k+\tau u_m^k)-f(x^k) \end{pmatrix} \in \mathbb{R}^m \\
\boldsymbol{Q}^k &= \begin{pmatrix} u_1^\top \\ \vdots \\ u_m^\top \end{pmatrix} \\
\boldsymbol{g}^k &= \arg\min_{\boldsymbol{g}\in\mathbb{R}^d}\{\frac{1}{2}\|\boldsymbol{b}^k-\boldsymbol{Q}^k\boldsymbol{g}\|^2+\lambda\|\boldsymbol{g}\|_1\} \\
\boldsymbol{x}^{k+1} &= \boldsymbol{x}^k-\gamma\boldsymbol{g}^k
\end{aligned}
$$

- Its convergence, which is beyond our lecture, can be found in (Wang et al., 2018; Cai et al., 2022)

Estimate error with sphere smoothing

- Sphere smoothing can estimate gradient with $O(1)$ queries on $f(x)$
- We let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d | \|x\| = 1\}$ be a unit sphere
- Sample $u \sim \mathcal{U}(\mathbb{S}^{d-1})$, we construct the gradient as

$$g = \frac{d[f(x + \tau u) - f(x)]}{\tau} \cdot u$$

  where g is a random estimate due to u; needs 2 queries
- Is g a good estimate of $\nabla f(x)$?

Smoothing function

- We introduce a new function

$$\hat{f}_\tau(x) = \mathbb{E}_v f(x + \tau v)$$

where τ is a given constant while $v \sim \mathcal{U}(\mathbb{B}^d(0,1))$ is a random variable

- Apparently, $\hat{f}_\tau(x)$ is a smoothing version of $f(x)$

- The following lemma shows that $g(x)$ is an **unbiased estimate** of $\nabla \hat{f}_\tau(x)$

Lemma 3

If $u \in \mathbb{R}^d$ and $u \sim \mathcal{U}(\mathbb{S}^{d-1}(0,1))$, $v \in \mathbb{R}^d$ and $v \sim \mathcal{U}(\mathbb{B}^d(0,1))$, for $f \in \mathbb{C}^1$, a certain x, and a certain τ, it holds that

$$\mathbb{E}_u[g(x)] = \mathbb{E}_u\left[\frac{d[f(x+\tau u) - f(x)]}{\tau} \cdot u\right] = \nabla \mathbb{E}_v f(x + \tau v) = \nabla \hat{f}_\tau(x)$$

Estimate error bound

- The gradient of smoothing function is close to the true gradient

Lemma 4

Let $v \in \mathbb{R}^d$, $v \sim \mathcal{U}(\mathbb{B}^d(0,1))$ *and* $\hat{f}_\tau(x) = \mathbb{E}_v f(x+\tau v)$. *If* $f(x)$ *is* L*-smooth, it holds that*

$$\|\nabla f(x) - \nabla \hat{f}_\tau(x)\| \leq L\tau$$

- Since $g(x)$ is an unbiased estimate of $\nabla \hat{f}_\tau(x)$, $g(x)$ is close to $\nabla f(x)$

Lemma 5

Suppose f *is* L*-smooth, and let* $u \sim \mathcal{U}(\mathbb{S}^{d-1}(0,1))$. *It holds that*

$$\mathbb{E}_u\|g(x)\|^2 \leq 2d\|\nabla f(x)\|^2 + \frac{\tau^2 L^2 d^2}{2}$$

ZO-GD with sphere smoothing

- ZO-GD with sphere smoothing iterates as follows

$$g_k = \frac{d}{\tau}(f(x_k + \tau u_k) - f(x_k))u_k, \quad u_k \sim \mathcal{U}(\mathbb{S}^{d-1}(0,1))$$
$$x_{k+1} = x_k - \gamma g_k$$

- According to Lemma 4, we have

$$\|\mathbb{E}_u[g_k] - \nabla f(x_k)\|^2 \le L^2\tau^2$$

- According to Lemma 5, we have

$$\mathbb{E}_u\|g_k\|^2 \le 2d\|\nabla f(x_k)\|^2 + \frac{\tau^2 L^2 d^2}{2}$$

- With the above inequalities, we next establish its convergence

Convergence in the non-convex scenario

Theorem 3 (Non-convex convergence)

Assume $f(x)$ to be L-smooth and $d \gg 1$. If we set $\gamma = 1/(4Ld)$, ZO-GD with sphere smoothing converges as follows

$$\frac{1}{K+1}\sum_{k=0}^{K}\|\nabla f(x_k)\|^2 \leq \frac{16Ld(f(x_0) - f(x^\star))}{K+1} + dL^2\tau^2.$$

ZO-GD with sphere smoothing converges to a **neighborhood** $O(dL^2\tau^2)$

If using decay τ, we can achieve exact convergence

Convergence in the non-convex scenario

- ZO-GD with sphere smoothing iterates as follows

$$g_k = \frac{d}{\tau_k}(f(x_k + \tau_k u_k) - f(x_k))u_k, \quad u_k \sim \mathcal{U}(\mathbb{S}^{d-1}(0,1))$$

$$x_{k+1} = x_k - \gamma g_k$$

Theorem 4 (Non-convex convergence)

Assume $f(x)$ to be L-smooth and $d \gg 1$. If we set $\gamma = 1/(4Ld)$ and $\sum_{k=0}^{K} \tau_k^2 \le R^2$, ZO-GD with sphere smoothing converges as follows

$$\frac{1}{K+1}\sum_{k=0}^{K} \|\nabla f(x_k)\|^2 \le \frac{16Ld(f(x_0) - f(x^\star))}{K+1} + \frac{dL^2R^2}{K+1}.$$

We leave the proof as an exercise.

Comparison with finite difference ZO-GD

- With Theorem 1, by choosing decaying τ_k such that $\sum_{k=0}^{K} \tau_k^2 \leq R^2$, the convergence rate of ZO-GD with forward difference is

$$\frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f(x^\star))}{K+1} + \frac{dL^2 R^2}{K+1}.$$

  We leave the proof as an exercise.

- Comparison with foward difference

Methods	**Rate**	**Queries per iteration**
Forward difference	$O(dL^2/K)$	$O(d)$
Sphere smoothing	$O(dL^2/K)$	$O(1)$

- Sphere smoothing saves queries without hurting convergence

Convergence in the strongly-convex scenario

Theorem 5 (Strongly-convex convergence)

Assume $f(x)$ *to be* L*-smooth,* μ*-strongly convex and* $d \gg 1$*. If we set* $\gamma = 1/(4Ld)$*, ZO-GD with sphere smoothing converges as follows*

$$f(x_K) - f(x^\star) \leq (1 - \frac{\mu}{8dL})^K (f(x_0) - f(x^\star)) + \frac{dL^2\tau^2}{8\mu}$$

We leave the proof as an exercise

Recall that ZO-GD with forward difference converges as

$$f(x_K) - f(x^\star) \leq (1 - \frac{\mu}{L})^K (f(x_0) - f(x^\star)) + \frac{dL^2\tau^2}{8\mu}$$

When $\tau = 1/\sqrt{K}$, both algorithm have the same dominant rate $O(1/K)$

Summary

- Zeroth-order optimization is useful when we cannot access gradient
- We introduced several methods to estimate gradient: finite difference, linear interpolation, random smoothing.
- We showed the convergence of zeroth-order gradient descent (ZO-GD)
- We theoretically prove that ZO-GD with sphere smoothing can save queries without hurting convergence compared to ZO-GD with forward difference.

References I

S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora, "Fine-tuning language models with just forward passes," *arXiv preprint arXiv:2305.17333*, 2023.

Y. Wang, S. Du, S. Balakrishnan, and A. Singh, "Stochastic zeroth-order optimization in high dimensions," in *International conference on artificial intelligence and statistics*. PMLR, 2018, pp. 1356–1365.

H. Cai, D. Mckenzie, W. Yin, and Z. Zhang, "Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 687–714, 2022.