

# Multivariate Statistical Analysis

## Lecture 04

Fudan University









luoluo@fudan.edu.cn

1 Zeroth-Order Optimization









2 Gaussian Smoothing

3 Complexity Analysis

# Optimization Problems: Your Feeling Before This Class

Settings	Smooth Convex	Nonsmooth Convex	Smooth Nonconvex	Nonsmooth Nonconvex
1st/2nd				
0th				

# Optimization Problems: Your Feeling After This Class

Settings	Smooth Convex	Nonsmooth Convex	Smooth Nonconvex	Nonsmooth Nonconvex
1st/2nd				
0th				

- 1 Zeroth-Order Optimization
- 2 Gaussian Smoothing
- 3 Complexity Analysis

In real applications, the explicit expression of gradient may be hard to achieve.

① Hyperparameter Tuning:

- It only returns the validation loss of the hyperparameter, and its gradient is unnecessary.

② Black-Box Attack to DNN:

- It only access to the input and the output of a targeted DNN.

# Zeroth-Order Optimization

We consider the problem

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

where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is continuous.

We focus on the scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{f(\mathbf{x}_t + \delta \mathbf{u}_t) - f(\mathbf{x}_t)}{\delta} \cdot \mathbf{u}_t$$

for some  $\eta_t > 0$  and  $\delta > 0$ , where  $\mathbf{u}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$ .

# Outline

1 Zeroth-Order Optimization

2 Gaussian Smoothing

3 Complexity Analysis



# Gaussian Smoothing

We define the Gaussian smoothing of  $f(\cdot)$  as

$$f_\delta(\mathbf{x}) = \mathbb{E}[f(\mathbf{x} + \delta \mathbf{u})] = \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} f(\mathbf{x} + \delta \mathbf{u}) \exp\left(-\frac{1}{2} \|\mathbf{u}\|_2^2\right) d\mathbf{u}$$

for some  $\delta > 0$ , where  $\mathbf{u} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$

The continuity of  $f(\cdot)$  means  $f_\delta(\cdot)$  is differentiable and it holds

$$\nabla f_\delta(\mathbf{x}) = \mathbb{E}\left[\frac{f(\mathbf{x} + \delta \mathbf{u}) - f(\mathbf{x})}{\delta} \cdot \mathbf{u}\right].$$

① If  $f(\cdot)$  is  $G$ -Lipschitz continuous, then

$$|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta G \sqrt{d}.$$

② If  $f(\cdot)$  is  $L$ -smooth, then

$$|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \frac{L\delta^2 d}{2} \quad \text{and} \quad \|\nabla f_\delta(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \leq \frac{L\delta(d+3)^{3/2}}{2}.$$

The properties of Gaussian smoothing:

- ① If  $f(\cdot)$  is  $G$ -Lipschitz continuous, then  $f_\delta(\cdot)$  is  $G$ -Lipschitz continuous and  $G\sqrt{d}/\delta$ -smooth.
- ② If  $f(\cdot)$  is  $L$ -smooth, then  $f_\delta(\cdot)$  is  $L$ -smooth.
- ③ If  $f(\cdot)$  is convex, then  $f_\delta(\cdot)$  is convex and  $f_\delta(\cdot) \geq f(\cdot)$ .

# Zeroth-Order Optimization

We study the scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t),$$

where

$$\mathbf{g}_\delta(\mathbf{x}; \mathbf{u}) = \frac{f(\mathbf{x} + \delta \mathbf{u}) - f(\mathbf{x})}{\delta} \cdot \mathbf{u}.$$

- ① If  $f(\cdot)$  is  $G$ -Lipschitz continuous, then

$$\mathbb{E} \|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})\|_2^2 \leq G^2(d+4)^2.$$

- ② If  $f(\cdot)$  is  $L$ -smooth, then

$$\mathbb{E} \|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})\|_2^2 \leq \frac{L^2 \delta^2 (d+6)^3}{2} + 2(d+4) \|\nabla f(\mathbf{x})\|_2^2.$$

# Outline

- 1 Zeroth-Order Optimization
- 2 Gaussian Smoothing
- 3 Complexity Analysis

# Zeroth-Order Optimization

## Theorem (Nonsmooth Convex)

Suppose  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex and  $G$ -Lipschitz. The iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t)$$

holds that

$$\begin{aligned} & \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t \mathbb{E}[(f(\mathbf{x}_t) - f(\mathbf{x}^*))] \\ & \leq \delta G \sqrt{d} + \frac{1}{2 \sum_{t=0}^{T-1} \eta_t} \left( \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + G^2 (d + 4)^2 \sum_{t=0}^{T-1} \eta_t^2 \right). \end{aligned}$$

# Zeroth-Order Optimization

## Theorem (Smooth Convex)

Suppose  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex and  $L$ -smooth. The iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t)$$

with  $\eta = 1/(4L(d+4))$  holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{4L(d+4) \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{T} + \frac{9L\delta^2(d+4)^2}{25}.$$

Additionally suppose  $f(\cdot)$  is  $\mu$ -strongly convex, then

$$\mathbb{E} \left[ \|\mathbf{x}_T - \mathbf{x}^*\|_2^2 - \Delta \right] \leq \left( 1 - \frac{\mu}{8L(d+4)} \right)^T \left( \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \Delta \right),$$

where  $\Delta = \frac{18\delta^2 L(d+4)^2}{25\mu}$ .

# Zeroth-Order Optimization

## Theorem (Smooth Nonconvex)

Suppose  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth. The iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_t)$$

with  $\eta = 1/(4L(d+4))$ ,

$$T = 16L(d+4)(f(\mathbf{x}_0) - f^*)\epsilon^{-2} \quad \text{and} \quad \delta = \frac{2\epsilon}{L} \sqrt{\frac{1}{(d+4)(d+16)}}$$

leads to

$$\mathbb{E} \|\nabla f(\mathbf{x}_{\text{out}})\|_2^2 \leq \epsilon^2$$

where  $\mathbf{x}_{\text{out}}$  is uniformly sampled from  $\{\mathbf{x}_0, \dots, \mathbf{x}_{T-1}\}$ .

# Zeroth-Order Optimization

The differentiability of  $\nabla f_\delta(\cdot)$  and the fact

$$\mathbb{E}[\mathbf{g}_\delta(\mathbf{x}; \mathbf{u})] = \nabla f_\delta(\mathbf{x})$$

means the mini-batch version scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \cdot \frac{1}{b} \sum_{i=1}^b \mathbf{g}_\delta(\mathbf{x}_t; \mathbf{u}_{t,i})$$

can reduce the iteration numbers.



# Zeroth-Order Optimization

The following lemma means we can also apply variance reduction on Gaussian smoothing.

## Lemma

For any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  and some  $\delta > 0$ , it holds that:

- ① If  $f(\cdot)$  is  $G$ -Lipschitz continuous, then

$$\mathbb{E} \|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u}) - \mathbf{g}_\delta(\mathbf{y}; \mathbf{u})\|_2^2 \leq \frac{2G^2 d \|\mathbf{x} - \mathbf{y}\|_2^2}{\delta}.$$

- ② If  $f(\cdot)$  is  $L$ -smooth continuous, then

$$\mathbb{E} \|\mathbf{g}_\delta(\mathbf{x}; \mathbf{u}) - \mathbf{g}_\delta(\mathbf{y}; \mathbf{u})\|_2^2 \leq \frac{3L^2 \delta^2 (d+6)^3}{2} + 3L^2 (d+4) \|\mathbf{x} - \mathbf{y}\|_2^2.$$