

# Optimization Theory

## Lecture 15

Fudan University

luoluo@fudan.edu.cn

- 1 Stochastic Recursive Gradient Algorithm
- 2 Zeroth-Order Optimization

1 Stochastic Recursive Gradient Algorithm

2 Zeroth-Order Optimization

# Stochastic Recursive Gradient Algorithm (SARAH)

---

**Algorithm 1** Stochastic Variance Reduced Gradient

---

```
1: Input:  $\mathbf{x}_0, \eta, m, S$ 
2:  $\tilde{\mathbf{x}}^{(0)} = \mathbf{x}_0$ 
3: for  $s = 0, \dots, S - 1$ 
4:    $\mathbf{v}_0 = \nabla f(\tilde{\mathbf{x}}^{(s)})$ 
5:    $\mathbf{x}_0 = \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{(s)}$ 
6:   for  $t = 0, \dots, m - 1$ 
7:     draw  $i_t$  from  $\{1, \dots, n\}$  uniformly
8:      $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$ 
9:      $\mathbf{v}_{t+1} = \nabla f_{i_t}(\mathbf{x}_{t+1}) - \nabla f_{i_t}(\mathbf{x}_t) + \mathbf{v}_t$ 
10:  end for
11:   $\tilde{\mathbf{x}}^{(s+1)} = \mathbf{x}_t$  for randomly chosen  $t \in \{0, \dots, m - 1\}$ 
12: end for
13: Output:  $\tilde{\mathbf{x}}^{(S)}$ 
```

---

# Stochastic Recursive Gradient Algorithm (SARAH)

SARAH outputs  $\tilde{\mathbf{x}}^{(S)}$  satisfying  $\mathbb{E} \|\nabla f(\tilde{\mathbf{x}}^{(S)})\|_2 \leq \epsilon$  within

- ①  $\mathcal{O}((n + \kappa) \log(1/\epsilon))$  IFO complexity for strongly convex objective;
- ②  $\mathcal{O}((n + L/\epsilon^2) \log(1/\epsilon))$  IFO complexity for convex objective.

The more interesting result is in the nonconvex optimization:

- ① Cong Fang, Chris Junchi Li, Zhouchen Lin, Tong Zhang.  
SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *NeurIPS* 2018.

# SGD for Nonconvex Optimization

We consider the stochastic optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[F(\mathbf{x}; \xi)],$$

where  $f(\mathbf{x})$  is  $L$ -smooth and lower bounded, and each  $F(\mathbf{x}; \xi)$  is differentiable.

Suppose there exists  $\sigma > 0$  such that  $\mathbb{E} \|\nabla F(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|_2^2 \leq \sigma^2$  for any  $\mathbf{x} \in \mathbb{R}^d$ . We run SGD iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{1}{|\mathcal{S}_t|} \sum_{\xi \in \mathcal{S}_t} \nabla F(\mathbf{x}_t; \xi)$$

with  $\mathcal{S}_t = \{\xi_1, \dots, \xi_b\}$ , where  $\xi_i \stackrel{\text{i.i.d}}{\sim} \mathcal{D}$ .

It can find an  $\epsilon$ -stationary point of  $f(\cdot)$  within

$$\mathcal{O}(L\sigma^2\epsilon^{-4})$$

stochastic first-order oracle (SFO) complexity in expectation.

# SARAH/SPIDER for Nonconvex Optimization

We consider the  $L$ -average smooth function, i.e. there exists  $L > 0$  such that

$$\mathbb{E} \|\nabla F(\mathbf{x}; \xi) - \nabla F(\mathbf{y}; \xi)\|_2^2 \leq L^2 \|\mathbf{x} - \mathbf{y}\|_2^2$$

for any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ .

The algorithms with stochastic recursive gradient require

$$\mathcal{O}(\sigma^2 \epsilon^{-2} + L \sigma^2 \epsilon^{-3})$$

SFO complexity to find an  $\epsilon$ -stationary point.

# SARAH/SPIDER for Nonconvex Optimization

We consider the finite-sum problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}).$$

Under the  $L$ -average smooth assumption, the algorithms with stochastic recursive gradient require

$$\mathcal{O}(n + L\sqrt{n}\epsilon^{-2})$$

SFO complexity to find an  $\epsilon$ -stationary point.



---

**Algorithm 2** ProbAbilistic Gradient Estimator (PAGE)

---

```
1: Input:  $\eta$ ,  $T$ ,  $b_0$ ,  $b$  and  $p$ .
2:  $\mathcal{S}_0 = \{\xi_1, \dots, \xi_{b_0}\}$  with  $\xi_i \stackrel{\text{i.i.d}}{\sim} \mathcal{D}$ 
3:  $\mathbf{v}_0 = \frac{1}{b_0} \sum_{\xi \in \mathcal{S}_0} \nabla F(\mathbf{x}_0; \xi)$ 
4: for  $t = 0, 1, \dots, T$  do
5:    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$ 
6:   draw  $\zeta_t \sim \text{Bernoulli}(p)$ 
7:   if  $\zeta_t = 1$  then
8:      $\mathcal{S}_{t+1} = \{\xi_1, \dots, \xi_{b_0}\}$  where  $\xi_i \stackrel{\text{i.i.d}}{\sim} \mathcal{D}$ 
9:      $\mathbf{v}_{t+1} = \frac{1}{b_0} \sum_{\xi \in \mathcal{S}_{t+1}} \nabla F(\mathbf{x}_{t+1}; \xi)$ 
10:  else
11:     $\mathcal{S}_{t+1} = \{\xi_1, \dots, \xi_b\}$  where  $\xi_i \stackrel{\text{i.i.d}}{\sim} \mathcal{D}$ 
12:     $\mathbf{v}_{t+1} = \mathbf{v}_t + \frac{1}{b} \sum_{\xi \in \mathcal{S}_{t+1}} (\nabla F(\mathbf{x}_{t+1}; \xi) - \nabla F(\mathbf{x}_t; \xi))$ 
13:  end if
14: end for
15:  $\mathbf{x}_{\text{out}} = \mathbf{x}_t$  for randomly chosen  $t \in \{0, \dots, T-1\}$ 
```

---

1 Stochastic Recursive Gradient Algorithm

2 Zeroth-Order Optimization

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

# Gaussian Smoothing

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

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

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

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

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

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

❷ 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

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

# Zeroth-Order Optimization

## Theorem (Nonsmooth)

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)

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