# Improved Zeroth-Order Variance Reduced Algorithms and Analysis for Nonconvex Optimization

**Kaiyi Ji**[1], Zhe Wang[1], Yi Zhou[2], Yingbin Liang[1]

[1]Ohio State University, [2]Duke University

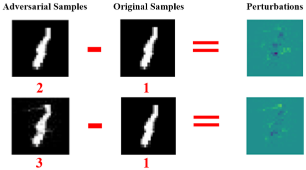
ICML 2019

# Zeroth-order (Gradient-free) Nonconvex Optimization

- Problem fomulation:

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

  - $f_i(\cdot)$: individual nonconvex loss function
  - Gradient of $f_i(\cdot)$ is unknown
  - Only the function value of $f_i(\cdot)$ is accessible
  - Examples:
    - Generation of black-box adversarial samples
    - Parameter optimization for black-box systems
    - Action exploration in reinforcement learning




Generating black-box adversarial samples

# Zeroth-order (Gradient-free) Nonconvex Optimization

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

- Standard assumptions on $f(\cdot)$:
  - $f(\cdot)$ is bounded below, i.e., $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$
  - $\nabla f_i(\cdot)$ is $L$-smooth, i.e.,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$$

  - (Online case) $\nabla f_i(\cdot)$ has bounded variance, i.e., there exists $\sigma > 0$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2$$

- Optimization goal: find an $\epsilon$-accurate stationary solution

$$\mathbb{E}\|\nabla f(x)\|^2 \leq \epsilon$$

# Existing Zeroth-Order SVRG

## ZO-SVRG ( Liu et al, 2018)

- Each outer-loop iteration estimates gradient by $\hat{\mathbf{g}}_s = \hat{\nabla}_{\text{rand}} f(\mathbf{x}_0^s, \mathbf{u}_0^s)$
- Each inner-loop iteration computes

$$\mathbf{v}_t^s = \frac{1}{|B|} \sum_{i \in B} \left( \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_t^s; \mathbf{u}_t^s) - \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_0^s; \mathbf{u}_0^s) \right) + \hat{\mathbf{g}}_s,$$

- Two-point gradient estimator: $\hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_t^s, \mathbf{u}_t^s) = \frac{d}{\beta}(f_i(\mathbf{x}_t^s + \beta \mathbf{u}_t^s) - f_i(\mathbf{x}_t^s))\mathbf{u}_t^s$
- $\mathbf{u}_t^s$: smoothing vector; $\beta$: smoothing parameter

# Existing Zeroth-Order SVRG

## ZO-SVRG ( Liu et al, 2018)

- Each outer-loop iteration estimates gradient by $\hat{\mathbf{g}}_s = \hat{\nabla}_{\text{rand}} f(\mathbf{x}_0^s, \mathbf{u}_0^s)$
- Each inner-loop iteration computes

$$\mathbf{v}_t^s = \frac{1}{|B|} \sum_{i \in B} \left( \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_t^s; \mathbf{u}_t^s) - \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_0^s; \mathbf{u}_0^s) \right) + \hat{\mathbf{g}}_s,$$

- Two-point gradient estimator: $\hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_t^s, \mathbf{u}_t^s) = \frac{d}{\beta}(f_i(\mathbf{x}_t^s + \beta \mathbf{u}_t^s) - f_i(\mathbf{x}_t^s))\mathbf{u}_t^s$
- $\mathbf{u}_t^s$: smoothing vector; $\beta$: smoothing parameter

| Algorithms | Convergence rate | # of function queries |
|---|---|---|
| ZO-SGD | $\mathcal{O}(\sqrt{d/T})$ | $\mathcal{O}(d\epsilon^{-2})$ |
| ZO-SVRG | $\mathcal{O}(d/T + 1/\lvert B\rvert)$ | $\mathcal{O}(d\epsilon^{-2} + n\epsilon^{-1})$ |

- Issue: ZO-SVRG has worse query complexity than ZO-SGD

# ZO-SVRG-Coord-Rand vs ZO-SVRG

## ZO-SVRG-Coord-Rand (This paper)

- Each outer-loop iteration estimates gradient by $\hat{\mathbf{g}}_s = \hat{\nabla}_{\text{coord}} f_S(\mathbf{x}^k)$
  - As a comparison, ZO-SVRG uses $\hat{\mathbf{g}}_s = \hat{\nabla}_{\text{rand}} f(\mathbf{x}_0^s, \mathbf{u}_0^s)$
- Each inner-loop iteration computes

$$\mathbf{v}_t^s = \frac{1}{|B|} \sum_{i \in B} \Big( \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_t^s; \underbrace{\mathbf{u}_{i,t}^s}_{\text{ZO-SVRG: } \mathbf{u}_t^s}) - \hat{\nabla}_{\text{rand}} f_i(\mathbf{x}_0^s; \underbrace{\mathbf{u}_{i,t}^s}_{\text{ZO-SVRG: } \mathbf{u}_0^s}) \Big) + \hat{\mathbf{g}}_s,$$

- $\hat{\nabla}_{\text{coord}} f(\cdot)$: coordinate-wise gradient estimator

| Algorithms | Convergence rate | Function query complexity |
|---|---|---|
| ZO-SGD | $\mathcal{O}(\sqrt{d/T})$ | $\mathcal{O}(d\epsilon^{-2})$ |
| ZO-SVRG | $\mathcal{O}(d/T + 1/\lvert B\rvert)$ | $\mathcal{O}(d\epsilon^{-2} + n\epsilon^{-1})$ |
| ZO-SVRG-Coord-Rand | $\mathcal{O}(1/T)$ | $\mathcal{O}\left(\min\left\{d\epsilon^{-5/3},\ dn^{2/3}\epsilon^{-1}\right\}\right)$ |

# Sharp Analysis for ZO-SVRG-Coord (Liu et al, 2018)

## ZO-SVRG-Coord (Liu et al, 2018)

- Each outer-loop iteration estimates gradient by $\hat{\mathbf{g}}_s = \hat{\nabla}_{\text{coord}} f_S(\mathbf{x}^k)$
- Each inner-loop iteration computes

$$\mathbf{v}_t^s = \frac{1}{|B|} \sum_{i \in B} \left( \hat{\nabla}_{\text{coord}} f_i(\mathbf{x}_t^s; \mathbf{u}_{i,t}^s) - \hat{\nabla}_{\text{coord}} f_i(\mathbf{x}_0^s; \mathbf{u}_{i,t}^s) \right) + \hat{\mathbf{g}}_s,$$

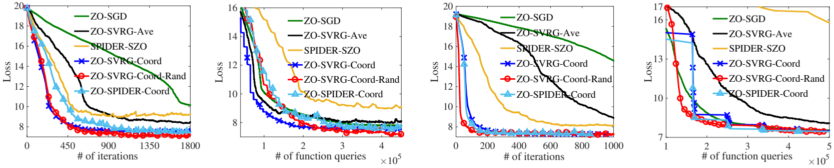
| Algorithms | Stepsize | Convergence rate | Function query complexity |
|---|---|---|---|
| ZO-SVRG-Coord | $\mathcal{O}(\frac{1}{d})$ | $\mathcal{O}(\frac{d}{T})$ | $\mathcal{O}\left(dn + \frac{d^2}{\epsilon} + \frac{dn}{\epsilon}\right)$ |
| ZO-SVRG-Coord (our analysis) | $\mathcal{O}(1)$ | $\mathcal{O}(\frac{1}{T})$ | $\mathcal{O}\left(\min\left\{\frac{d}{\epsilon^{5/3}}, \frac{dn^{2/3}}{\epsilon}\right\}\right)$ |

**Key idea:**

- Coordinate-wise gradient estimator → high accuracy → faster rate

# More Results

- Develop a faster zeroth-order SPIDER-type algorithm
- Develop improved zeroth-order algorithms for
  - nonconvex nonsmooth optimization
  - convex smooth optimization
  - Polyak-Łojasiewicz (PL) condition
- Experiments:




Generating black-box adversarial examples for DNNs

Thanks!