Zeroth-Order Methods for Constrained Nonconvex Nonsmooth Stochastic Optimization

Zhuanghua Liu, Cheng Chen, Luo Luo, Bryan Kian Hsiang Low

Problem Formulation

$$\min_{x \in \Omega} F(x) \triangleq \mathbb{E}[f(x;\xi)],$$

where $f(x;\xi)$ is probably nonconvex and nonsmooth,

and $\Omega \subseteq \mathbb{R}^d$ is convex and compact.

Standard assumptions: Lipschitz continuous gradient

$$\|\nabla f(x;\xi) - \nabla f(y;\xi)\| \leq L \|x - y\|$$

Problem Formulation

$$\min_{x\in\Omega} F(x) \triangleq \mathbb{E}[f(x;\xi)],$$

where $f(x;\xi)$ is probably nonconvex and nonsmooth,

and $\Omega \subseteq \mathbb{R}^d$ is convex and compact.

Our assumptions: Lipschitz continuous function

$$\|f(x;\xi) - f(y;\xi)\| \leq L\|x - y\|$$

Applications: deep RELU neural networks, regularized SVM

Most existing work only provide asymptotic convergence analysis

Challenges

RQ. *How to provide non-asymptotic convergence analysis for the constrained nonsmooth optimization problem?*

Classical convergence criteria for constrained smooth opt:

Gradient Mapping

$$\frac{1}{\gamma}\left(x - \psi(x, \nabla f(x), \gamma)\right)$$

$$\psi(x, g, \gamma) \triangleq \arg\min_{y \in \Omega}\left(\langle g, y\rangle + \frac{1}{2\gamma}\|y - x\|^2\right)$$

Frank-Wolfe Gap

$$\max_{u \in \Omega}\langle u - x,\ -\nabla f(x)\rangle$$

Challenges

RQ. *How to provide non-asymptotic convergence analysis for the constrained nonsmooth optimization problem?*

Classical convergence criteria for constrained smooth opt:

Gradient Mapping

$$\frac{1}{\gamma}\left(x - \psi(x, \nabla f(x), \gamma)\right)$$

$$\psi(x, g, \gamma) \triangleq \arg\min_{y \in \Omega}\left(\langle g, y\rangle + \frac{1}{2\gamma}\|y - x\|^2\right)$$

Frank-Wolfe Gap

$$\max_{u \in \Omega}\langle u - x, \ -\nabla f(x)\rangle$$

Our objective function may not be differentiable everywhere

$\Longrightarrow \nabla f(x)$ is not well-defined everywhere

Extension with Clarke Subdifferential

Rademacher's Theorem:

A Lipschitz function is differentiable almost everywhere.

Clarke Subdifferential:

$$\partial f(x) = \text{conv}\{s: x^k \to x, \nabla f(x^k) \text{ exists, and } \nabla f(x^k) \to s\}$$

Extension with Clarke Subdifferential

Clarke Subdifferential:

$$\partial f(x) = \text{conv}\{s: x^k \to x, \nabla f(x^k) \text{ exists, and } \nabla f(x^k) \to s\}$$

- (γ, ϵ)-Generalized Clarke stationary point (GCSP):

$$\min_{g \in \partial f(x)} \left\| \frac{1}{\gamma}\left(x - \psi(x, \mathrm{g}, \gamma)\right) \right\| \leq \epsilon$$

- ϵ-Clarke Frank-Wolfe stationary point (CFWSP):

$$\min_{g \in \partial f(x)} \max_{u \in \Omega} \langle u - \mathrm{x}, \ -\mathrm{g} \rangle \leq \epsilon$$

Impossibility Result

- (γ, ϵ)-Generalized Clarke stationary point (GCSP):

$$\min_{g \in \partial f(x)} \left\| \frac{1}{\gamma}\left(x - \psi(x, \mathrm{g}, \gamma)\right) \right\| \leq \epsilon$$

- ϵ-Clarke Frank-Wolfe stationary point (CFWSP):

$$\min_{g \in \partial f(x)} \max_{u \in \Omega} \langle u - \mathrm{x}, \ -\mathrm{g} \rangle \leq \epsilon$$

Theorem (Informal). Neither (γ, ϵ)-GCSP nor ϵ-CFWSP permits a finite-time convergence analysis for any deterministic or randomized algorithms to solve the constrained nonsmooth problem.

Refined Approximate Stationarity

Goldstein δ-subdifferential. (Goldstein (1977))

$$\partial_\delta f(x) \triangleq \mathrm{conv}(\cup_{\|y-x\|\leq\delta} \partial f(y)),$$

where $\delta > 0$. If $\delta = 0$, Goldstein δ-subdifferential is reduced to Clarke subdifferential $\partial f(x)$.

Proposed approximate stationarity

- $(\gamma, \delta, \epsilon)$-Generalized Goldstein stationary point (GGSP):

$$\min_{g\in\partial_\delta f(x)} \left\| \frac{1}{\gamma}\left(x - \psi(x, \mathrm{g}, \gamma)\right) \right\| \leq \epsilon$$

- (δ, ϵ)- Goldstein Frank-Wolfe stationary point (GFWSP):

$$\min_{g\in\partial_\delta f(x)} \max_{u\in\Omega} \langle u - \mathrm{x}, \ -\mathrm{g}\rangle \leq \epsilon$$

Key Ingredients for ZO Stochastic Methods

Randomized Smoothing. (Lin et al. (2022))

Given a L-Lipschitz function $f(x)$ and a distribution Q, a smoothing function is

$$f_\delta(x) = \mathbb{E}_{u\sim Q}[f(x+\delta u)].$$

- $|f_\delta(x) - f(x)| \leq \delta L$;
- $f_\delta(x)$ is differentiable everywhere with $cL\sqrt{d}/\delta$-Lipschitz gradient;
- $\nabla f_\delta(x) \in \partial_\delta f(x)$

Gradient Estimators

The zeroth-order stochastic gradient estimator (Agarwal et al., 2010)

$$\hat{g}(x, w, \xi) = \frac{d}{2\delta}\big(f(x + \delta w; \xi) - f(x - \delta w; \xi)\big)w,$$

Satisfies $\mathbb{E}[\hat{g}(x, w, \xi)] = \nabla f_\delta(x).$

Gradient Estimators

The zeroth-order stochastic gradient estimator (Agarwal et al., 2010)

$$\hat{g}(x, w, \xi) = \frac{d}{2\delta}\big(f(x + \delta w; \xi) - f(x - \delta w; \xi)\big)w,$$

Satisfies $\mathbb{E}[\hat{g}(x, w, \xi)] = \nabla f_\delta(x) \in \partial_\delta f(x)$.

Gradient Estimators

The zeroth-order stochastic gradient estimator (Agarwal et al., 2010)

$$\hat{g}(x, w, \xi) = \frac{d}{2\delta}\big(f(x + \delta w; \xi) - f(x - \delta w; \xi)\big)w,$$

Satisfies $\mathbb{E}[\hat{g}(x, w, \xi)] = \nabla f_\delta(x) \in \partial_\delta f(x)$.

- Minibatch gradient estimator

$$v_t = \frac{1}{b}\sum_{i=1}^{b} \hat{g}\big(x_t, w_{i,t}, \xi_{i,t}\big).$$

Gradient Estimators

The zeroth-order stochastic gradient estimator (Agarwal et al., 2010)

$$\hat{g}(x, w, \xi) = \frac{d}{2\delta}\big(f(x + \delta w; \xi) - f(x - \delta w; \xi)\big)w,$$

Satisfies $\mathbb{E}[\hat{g}(x, w, \xi)] = \nabla f_\delta(x) \in \partial_\delta f(x)$.

- Minibatch gradient estimator

$$v_t = \frac{1}{b}\sum_{i=1}^{b} \hat{g}\big(x_t, w_{i,t}, \xi_{i,t}\big).$$

- Variance-reduced gradient estimator

$$v_t = \frac{1}{b}\sum_{i=1}^{b} \hat{g}\big(x_t, w_{i,t}, \xi_{i,t}\big) - \frac{1}{b}\sum_{i=1}^{b} \hat{g}\big(x_{t-1}, w_{i,t}, \xi_{i,t}\big) + v_{t-1}.$$

ZO Stochastic Algorithms

Projected gradient descent

$$x_{t+1} = \Pi_{\Omega}(x_t - \gamma v_t)$$

Require projection every step

Frank-Wolfe algorithm

$$u_t = \arg\max_{u \in \Omega} \langle u, -v_t \rangle$$

$$x_{t+1} = x_t + \gamma_t(u_t - x_t)$$

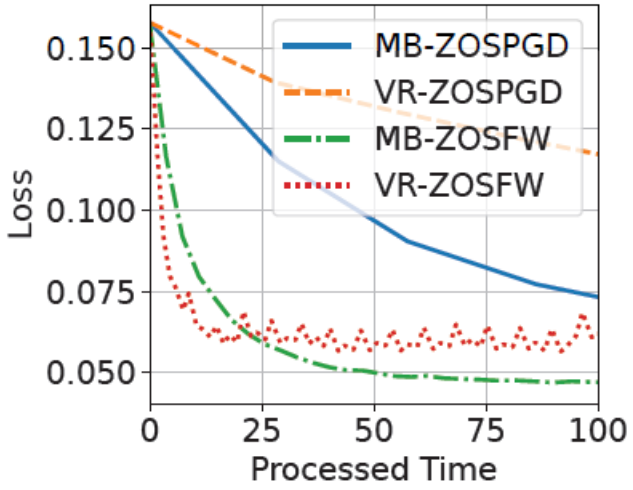
Require linear maximization oracle

v_t can be either minibatch or variance-reduced gradient estimator

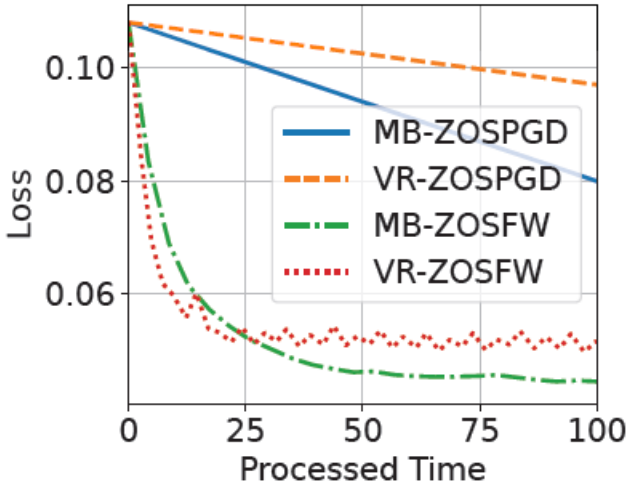
Comparison of Oracle Complexity

Methods	Convergence Criteria	ZO Oracle Calls
MB-ZOSPGD	$(\gamma, \delta, \epsilon) - \text{GGSP}$	$\mathcal{O}(d^{1.5}\delta^{-1}\epsilon^{-4})$
VR-ZOSPGD	$(\gamma, \delta, \epsilon) - \text{GGSP}$	$\mathcal{O}(d^{1.5}\delta^{-1}\epsilon^{-3})$
MB-ZOSFW	$(\delta, \epsilon) - \text{GFWSP}$	$\mathcal{O}(d^{1.5}\delta^{-1}\epsilon^{-4})$
VR-ZOSFW	$(\delta, \epsilon) - \text{GFWSP}$	$\mathcal{O}(d^{1.5}\delta^{-1}\epsilon^{-3})$

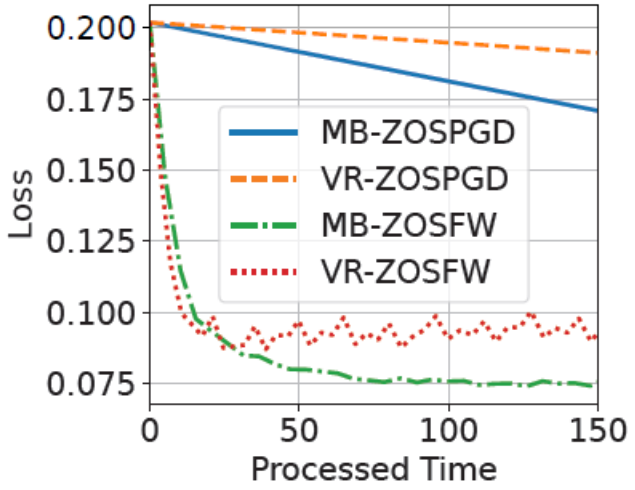
Experiments



(a) $d = 3000$



(b) $d = 4000$



(c) $d = 5000$

Robust low-rank matrix recovery problem

Conclusion

- We propose novel convergence criteria for constrained nonconvex nonsmooth optimization problem
- We propose zeroth-order stochastic algorithms and present the non-asymptotic convergence analysis w.r.t. the proposed convergence criteria

Conclusion

- We propose novel convergence criteria for constrained nonconvex nonsmooth optimization problem
- We propose zeroth-order stochastic algorithms and present the non-asymptotic convergence analysis w.r.t. the proposed convergence criteria

Future directions

- Lower bound
- Dimension-independent first-order methods