Accelerated Zeroth-Order and First-Order Momentum Methods from Mini to Minimax Optimization

Feihu Huang[1], Shangqian Gao[1], Jian Pei[2], Heng Huang[1]

1. University of Pittsburgh, USA
2. Simon Fraser University, Canada

NeurIPS | 2022

Thirty-sixth Conference on Neural Information Processing Systems

JMLR

Outline

- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Outline

- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Background

- **Zeroth-order methods** are a class of powerful optimization tools to solve many complex machine learning problems, whose explicit gradients are difficult or even infeasible to access.

- **Minimax optimization** can effectively solve the problems with a hierarchical structure, so it is widely used to many machine learning applications such as adversarial training and robust federated learning.

Background

- For example, the robust federated learning could be defined as:

$$\min_{w\in\Omega}\max_{p\in\Pi}\left\{\sum_{i=1}^{n} p_i \mathbb{E}_{\xi\sim\mathcal{D}_i}[f_i(w;\xi)] - \lambda\psi(p)\right\},$$

where $p_i \in (0,1)$ denotes the proportion of i-th device in the entire model, and $f_i(w;\xi)$ is the loss function on i-th device, and $\lambda > 0$ is a tuning parameter, and $\psi(p)$ is a (strongly) convex regularization. Here $\Pi = \{p \in \mathbb{R}^n : \sum_{i=1}^n p_i = 1,\ p_i \geq 0\}$ is a n-dimensional simplex, and $\Omega \subseteq \mathbb{R}^d$ is a nonempty convex set.

Outline

- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Zeroth-Order Gradient Estimators

$$\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)],$$

$$\hat{\nabla} f(x;\xi) = \frac{f(x+\mu u;\xi) - f(x;\xi)}{\mu/d} u,$$

where $u \in \mathbb{R}^d$ is a vector generated from the uniform distribution over the unit sphere, and μ is a smoothing parameter. Let $f_\mu(x;\xi) = \mathbb{E}_{u \sim U_B}[f(x+\mu u;\xi)]$ be a smooth approximation of $f(x;\xi)$, where U_B is the uniform distribution over the d-dimensional unit Euclidean ball

Accelerated Zeroth- & First-Order Momentum Methods

Momentum Techniques:

1) Apply on the variables

2) Apply on the gradients

Accelerated Zeroth- & First-Order Momentum Methods

$$\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)],$$

Algorithm 1 Acc-ZOM Algorithm for Mini Optimization

Input: T, parameters $\{\gamma, k, m, c\}$ and initial input $x_1 \in \mathcal{X}$;
initialize: Draw a sample ξ_1, and sample a vector $u \in \mathbb{R}^d$ from uniform distribution over unit sphere, then compute $v_1 = \hat{\nabla} f(x_1;\xi_1)$, where the zeroth-order gradient is estimated from (3);
for $t = 1, 2, \ldots, T$ **do**
    Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
    if $\mathcal{X} = \mathbb{R}^d$ **then**
        Update $x_{t+1} = x_t - \gamma \eta_t v_t$;
    else
        Update $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)$, and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
    end if
    Compute $\alpha_{t+1} = c\eta_t^2$;
    Draw a sample ξ_{t+1}, and sample a vector $u \in \mathbb{R}^d$ from uniform distribution over unit sphere, then compute $\boxed{v_{t+1} = \hat{\nabla} f(x_{t+1};\xi_{t+1}) + (1-\alpha_{t+1})\big[v_t - \hat{\nabla} f(x_t;\xi_{t+1})\big],}$ where the zeroth-order gradients are estimated from (3);
end for
Output: (for theoretical) x_ζ chosen uniformly random from $\{x_t\}_{t=1}^T$.
Output: (for practical) x_T.

Accelerated Zeroth- & First-Order Momentum Methods

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}'}[f(x,y;\xi)],$$

Algorithm 2 Acc-ZOMDA Algorithm for Minimax Optimization

Input: T, parameters $\{\gamma,\lambda,k,m,c_1,c_2\}$ and initial input $x_1\in\mathcal{X}$ and $y_1\in\mathcal{Y}$;
initialize: Draw a mini-batch samples $\mathcal{B}_1=\{\xi_i^1\}_{i=1}^b$, and draw vectors $\{\hat{u}_i\in\mathbb{R}^{d_1}\}_{i=1}^b$ and $\{\tilde{u}_i\in\mathbb{R}^{d_2}\}_{i=1}^b$ from uniform distribution over unit sphere, then compute $v_1=\hat{\nabla}_x f(x_1,y_1;\mathcal{B}_1)$ and $w_1=\hat{\nabla}_y f(x_1,y_1;\mathcal{B}_1)$, where the zeroth-order gradients are estimated from (4) and (5);
for $t=1,2,\dots,T$ **do**
    Compute $\eta_t=\frac{k}{(m+t)^{1/3}}$;
    if $\mathcal{X}=\mathbb{R}^{d_1}$ **then**
        Update $x_{t+1}=x_t-\gamma\eta_t v_t$;
    else
        Update $\tilde{x}_{t+1}=\mathcal{P}_{\mathcal{X}}(x_t-\gamma v_t)$ and $x_{t+1}=x_t+\eta_t(\tilde{x}_{t+1}-x_t)$;
    end if
    Update $\tilde{y}_{t+1}=\mathcal{P}_{\mathcal{Y}}(y_t+\lambda w_t)$ and $y_{t+1}=y_t+\eta_t(\tilde{y}_{t+1}-y_t)$;
    Compute $\alpha_{t+1}=c_1\eta_t^2$ and $\beta_{t+1}=c_2\eta_t^2$;
    Draw a mini-batch samples $\mathcal{B}_{t+1}=\{\xi_i^{t+1}\}_{i=1}^b$, and draw vectors $\{\hat{u}_i\in\mathbb{R}^{d_1}\}_{i=1}^b$ and $\{\tilde{u}_i\in\mathbb{R}^{d_2}\}_{i=1}^b$ from uniform distribution over unit sphere;
    Compute $v_{t+1}=\hat{\nabla}_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})+(1-\alpha_{t+1})\big[v_t-\hat{\nabla}_x f(x_t,y_t;\mathcal{B}_{t+1})\big]$ and $w_{t+1}=\hat{\nabla}_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})+(1-\beta_{t+1})\big[w_t-\hat{\nabla}_y f(x_t,y_t;\mathcal{B}_{t+1})\big]$, where the zeroth-order gradients are estimated from (4) and (5).
end for
Output: (for theoretical) x_ζ and y_ζ chosen uniformly random from $\{x_t,y_t\}_{t=1}^T$.
Output: (for practical) x_T and y_T.

Accelerated Zeroth- & First-Order Momentum Methods

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}'}[f(x,y;\xi)],$$

Algorithm 3 Acc-MDA Algorithm for Minimax Optimization

Input: T, parameters $\{\gamma, \lambda, k, m, c_1, c_2\}$ and initial input $x_1 \in \mathcal{X}$ and $y_1 \in \mathcal{Y}$;
initialize: Draw a mini-batch samples $\mathcal{B}_1 = \{\xi_i^1\}_{i=1}^b$, and then compute stochastic gradients $v_1 = \nabla_x f(x_1, y_1; \mathcal{B}_1)$ and $w_1 = \nabla_y f(x_1, y_1; \mathcal{B}_1)$;
for $t = 1, 2, \ldots, T$ **do**
  Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
  if $\mathcal{X} = \mathbb{R}^{d_1}$ **then**
    Update $x_{t+1} = x_t - \gamma\eta_t v_t$;
  else
    Update $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)$ and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
  end if
  Update $\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t)$ and $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$;
  Compute $\alpha_{t+1} = c_1\eta_t^2$ and $\beta_{t+1} = c_2\eta_t^2$;
  Draw a mini-batch samples $\mathcal{B}_{t+1} = \{\xi_i^{t+1}\}_{i=1}^b$, and then compute stochastic gradients $v_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1-\alpha_{t+1})\big[v_t - \nabla_x f(x_t, y_t; \mathcal{B}_{t+1})\big]$ and $w_{t+1} = \nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1-\beta_{t+1})\big[w_t - \nabla_y f(x_t, y_t; \mathcal{B}_{t+1})\big]$;
end for
Output: (for theoretical) x_ζ and y_ζ chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.
Output: (for practical) x_T and y_T.

Outline

- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Convergence Properties

Table 1: **Query complexity** comparison of the representative non-convex zeroth-order methods for finding an ϵ-stationary point of the **black-box** mini-optimization problem (1) and minimax-optimization problem (2), respectively. GauGE, UniGE and CooGE are abbreviations of Gaussian, Uniform and Coordinate-Wise smoothing gradient estimators, respectively. Here κ_y denotes the condition number for function $f(\cdot, y)$. Note that Appendix B provides a comparison of assumptions used in the zeroth-order methods, and Appendix C provides a detailed proof to obtain a correct query complexity of ZO-Min-Max algorithm (Liu et al., 2019b).

Problem	Algorithm	Reference	Estimator	Batch Size	Complexity
Mini	ZO-SGD	Ghadimi and Lan (2013)	GauGE	$O(1)$	$O(d\epsilon^{-4})$
	ZO-AdaMM	Chen et al. (2019)	UniGE	$O(\epsilon^{-2})$	$O(d^2\epsilon^{-4})$
	ZO-SVRG	Ji et al. (2019)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-10/3})$
	ZO-SPIDER-Coord	Ji et al. (2019)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-3})$
	SPIDER-SZO	Fang et al. (2018)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-3})$
	Acc-ZOM	Ours	UniGE	$O(1)$	$O(d^{3/4}\epsilon^{-3})$
Minimax	ZO-Min-Max	Liu et al. (2019b)	UniGE	$O((d_1+d_2)\kappa_y^2\epsilon^{-2})$	$O((d_1+d_2)\kappa_y^6\epsilon^{-6})$
	ZO-SGDA	Wang et al. (2020)	GauGE	$O((d_1+d_2)\epsilon^{-2})$	$O((d_1+d_2)\kappa_y^5\epsilon^{-4})$
	ZO-SGDMSA	Wang et al. (2020)	GauGE	$O((d_1+d_2)\epsilon^{-2})$	$\tilde{O}((d_1+d_2)\kappa_y^2\epsilon^{-4})$
	ZO-SREDA-Boost	Xu et al. (2020a)	CooGE	$O(\max(\kappa_y\epsilon^{-1}, d_1+d_2)\kappa_y\epsilon^{-1})$	$O((d_1+d_2)\kappa_y^3\epsilon^{-3})$
	Acc-ZOMDA	Ours	UniGE	$O(1)$	$\tilde{O}((d_1+d_2)^{3/4}\kappa_y^{4.5}\epsilon^{-3})$

$$\min_{x\in\mathcal{X}} f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(x;\xi)], \qquad \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}'}[f(x,y;\xi)],$$

Convergence Properties

Table 2: **Gradient complexity** comparison of the representative first-order methods for finding an ϵ-stationary point of the minimax problem (2). Here Y denotes the fact that there exists a convex constraint on variable, otherwise is N. Note that our theoretical results do not rely on any assumption on convex constraint sets $\mathcal{X}$ and $\mathcal{Y}$, so it can be easily extend to the unconstrained setting.

Algorithm	**Reference**	**Constraint on x, y**	**Loop(s)**	**Batch Size**	**Complexity**
PGSVRG	Rafique et al. (2018)	N, N	Double	$O(\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-4})$
SGDA	Lin et al. (2019)	N, Y	Single	$O(\kappa_y\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-4})$
SREDA	Luo et al. (2020)	N, Y	Double	$O(\kappa_y^2\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-3})$
SREDA-Boost	Xu et al. (2020a)	N, N	Double	$O(\kappa_y^2\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-3})$
Acc-MDA	Ours	Y (N), Y	Single	$O(1)$	$\tilde{O}(\kappa_y^{4.5}\epsilon^{-3})$
Acc-MDA	Ours	Y (N), Y	Single	$O(\kappa_y^{\nu}),\ \nu > 0$	$\tilde{O}(\kappa_y^{(4.5-\nu/2)}\epsilon^{-3})$

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}'}[f(x,y;\xi)],$$

Outline

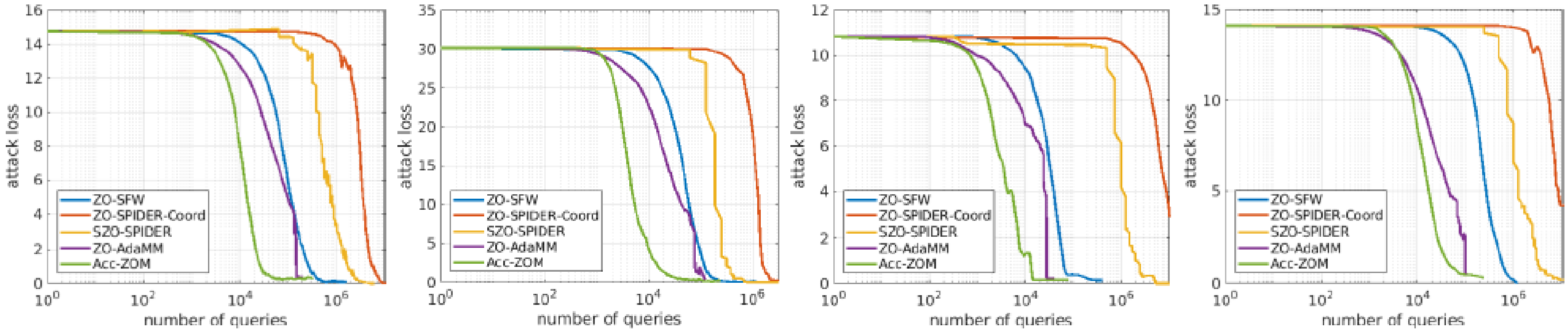
- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Experimental Results

(1) Black-Box Adversarial Attack to DNNs

$$\min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} \max\left(f_{b_i}(x + a_i) - \max_{j \neq b_i} f_j(x + a_i), 0\right), \quad \text{s.t. } \mathcal{X} = \{\|x\|_\infty \leq \varepsilon\}$$

Experimental Results



(a) MNIST (b) FashionMNIST (c) CIFAR-10 (d) SVHN

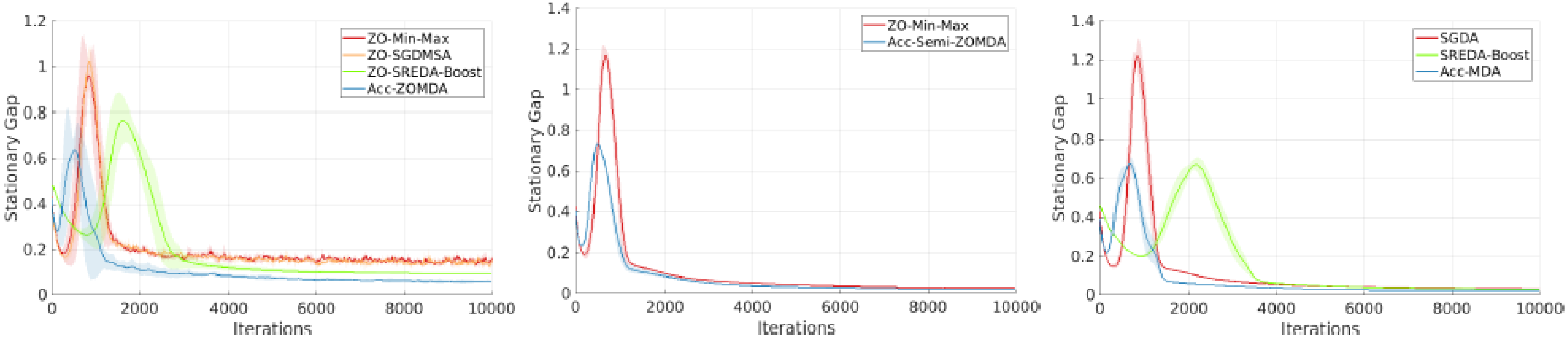
Figure 1: Experimental results of black-box adversarial attack on four datasets: MNIST, FashionMNIST, CIFAR-10 and SVHN.

Experimental Results

(2) Data Poisoning Attack to Logistic Regression

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x, y) = h(x, y; \mathcal{D}_p) + h(0, y; \mathcal{D}_t),$$
$$\text{s.t. } \mathcal{X} = \{\|x\|_\infty \leq \varepsilon\}, \ \mathcal{Y} = \{\|y\|_2^2 \leq \lambda_{\text{reg}}\}$$

Experimental Results



(a) Two-Side Black-Box Attack (b) One-Side Black-Box Attack (c) Transparent Attack

Figure 3: Stationary gap of different methods in two-side black-box scenario, one-side black-box scenario and transparent scenario.

Outline

- Background
- Accelerated Zeroth- & First-Order Momentum Methods
- Convergence Properties
- Experimental Results
- Conclusions

Conclusions

- 1) We proposed a class of accelerated zeroth-order and first-order momentum methods for both mini- and minimax-optimization.
- 2) We provided an effective convergence analysis framework for our methods, and proved that our **zeroth-order** methods obtain a low **query complexity** without any large batches. Meanwhile, our **first-order** method obtain a low **gradient complexity** without any large batches

Thanks!

Q&A