Zeroth-Order Stochastic Variance Reduction for Nonconvex Optimization

Bing-Shu Wu Rui-Teng Lin

Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Introduction

Zeorth-Order Stochastic Variance Reduction (ZO-SVRG) [Liu et al. 2018] is essentially the Variance Reduced(VR) version of ZO-SGD, where Zeroth-Order(ZO) implies that the algorithms uses **only function values** to estimate the descent direction for optimization.

Preliminaries

Consider a nonconvex finite-sum problem of the form

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

where $\{f_i(\mathbf{x})\}_{i=1}^n$ are n individual nonconvex cost functions.

Assumptions

Assumption A1 Functions $\{f_i\}$ have Lipschitz continuous gradients (L-smooth), i.e.,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \le L\|\mathbf{x}-\mathbf{y}\|_2$$

for any $\mathbf{x}$ and $\mathbf{y}$, $i \in [n]$ and some $L < \infty$.
For ease of notation, $[n]$ represents the integer set $\{1, 2, ..., n\}$.
Assumption A2 The variance of stochastic gradients is bounded as

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \le \sigma^2$$

ZO-SVRG Algorithm

We know that the variance-reduced gradient estimation in SVRG [Johnson and Zhang 2013] is obtained by gradient blending:

$$\mathbf{v}_k^s \leftarrow \nabla f_{\mathcal{I}_k}(\mathbf{x}_k^s) - \nabla f_{\mathcal{I}_k}(\mathbf{x}_0^s) + \mathbf{g}_s$$

where $\mathbf{g}_s$ is the complete gradient, and $\nabla f_{\mathcal{I}}$ is the batch gradient:

$$\mathbf{g} = \nabla f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}), \quad \nabla f_{\mathcal{I}}(\mathbf{x}) = \frac{1}{b}\sum_{i\in\mathcal{I}}\nabla f_i(\mathbf{x})$$

Similarly, we perform the tricks in zeroth order, replacing the gradients ∇f with zeroth order gradient estimator $\hat{\nabla} f$

$$\hat{\mathbf{g}} = \hat{\nabla} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\nabla} f_i(\mathbf{x}), \quad \hat{\nabla} f_{\mathcal{I}}(\mathbf{x}) = \frac{1}{b}\sum_{i\in\mathcal{I}}\hat{\nabla} f_i(\mathbf{x})$$

$$\hat{\mathbf{v}}_k^s \leftarrow \hat{\nabla} f_{\mathcal{I}_k}(\mathbf{x}_k^s) - \hat{\nabla} f_{\mathcal{I}_k}(\mathbf{x}_0^s) + \hat{\mathbf{g}}_s$$

Where the gradient estimators $\hat{\nabla} f_i$ in the is one of the estimators in the following block.

Zeroth-Order Gradient Estimators

Random Gradient Estimator:

$$\hat{\nabla} f_i(\mathbf{x}) = \frac{d}{\mu}[f_i(\mathbf{x}+\mu\mathbf{u}_i) - f_i(\mathbf{x})]\,\mathbf{u}_i, \text{ for } i \in [n]$$

Average Random Gradient Estimator:

$$\hat{\nabla} f_i(\mathbf{x}) = \frac{d}{\mu q}\sum_{j=1}^{q}[f_i(\mathbf{x}+\mu\mathbf{u}_{i,j}) - f_i(\mathbf{x})]\,\mathbf{u}_{i,j}, \text{ for } i \in [n]$$

Coordinate-wise Gradient Estimator:

$$\hat{\nabla} f_i(\mathbf{x}) = \sum_{\ell=1}^{d}\frac{1}{2\mu_\ell}[f_i(\mathbf{x}+\mu_\ell\mathbf{e}_\ell) - f_i(\mathbf{x}-\mu_\ell\mathbf{e}_\ell)]\,\mathbf{e}_\ell, \text{ for } i \in [n]$$

Theorem 1: Convergence Rate of ZO-SVRG

With the parameter setting $\mu_\ell = \mu = \frac{1}{\sqrt{dT}}$, $\eta_k = \eta = \frac{\rho}{Ld}$, and proper inner loop length m, where $0 < \rho \le 1$ is a small universal constant, then we can have the following convergence rate

RandGradEst: $\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}})\|_2^2\right] \le O\left(\frac{d}{T} + \frac{\delta_n}{b}\right)$

Avg-RandGradEst: $\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}})\|_2^2\right] \le O\left(\frac{d}{T} + \frac{\delta_n}{b\min\{d,q\}}\right)$

CoordGradEst: $\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}})\|_2^2\right] \le O\left(\frac{d}{T}\right)$

Comparison of Convergence rate

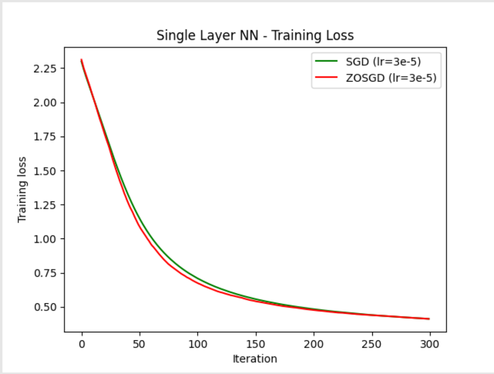
Algorithm	Gradient Estimator	Stepsize	Convergence Rate	Query Complexity
ZO-SGD	RandGradEst	$O(\min\{\frac{1}{d}, \frac{1}{\sqrt{dT}}\})$	$O(\sqrt{d/T})$	$O(bT)$
ZO-SVRC	CoordGradEst	$O(\frac{1}{n^\alpha}), \alpha \in (0,1)$	$O(\sqrt{d/T})$	$O(dnS + JbT)$
ZO-SVRG	RandGradEst	$O(\frac{1}{d})$	$O(\frac{d}{T} + \frac{1}{b})$	$O(nS + bT)$
ZO-SVRG-Avg	Avg-RandGradEst	$O(\frac{1}{d})$	$O(\frac{d}{T} + \frac{1}{b\min\{d,q\}})$	$O(q(nS + bT))$
ZO-SVRG-Coord	CoordGradEst	$O(\frac{1}{d})$	$O(\frac{d}{T})$	$O(d(nS + bT))$

Table 1. Comparison of Convergence rate with other ZO-methods

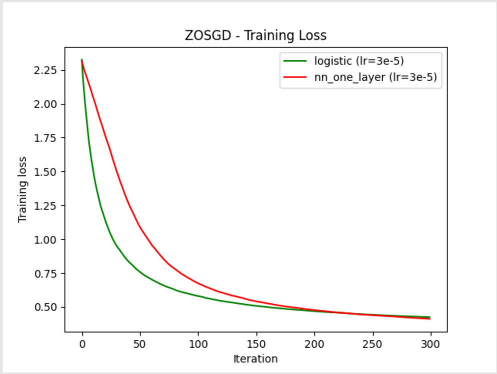
An Ablation Study: Validating Zeroth-Order Method Theory

In this experiment, we implemented a handcrafted zeroth-order SGD and trained it on the MNIST dataset to explore the advantages and limitations of zeroth-order methods.

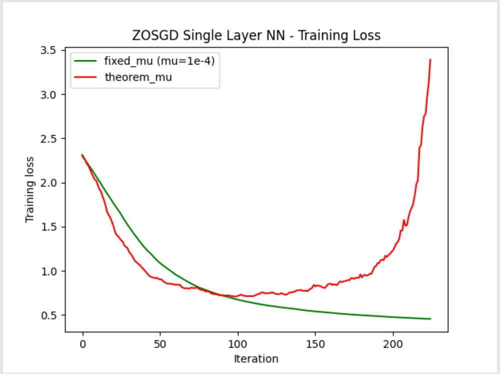
- Although the curves in (a) are nearly identical, ZOSGD requires slightly more time to train and needs a sufficiently small learning rate to ensure convergence. It is worth noting that both methods were trained on a CPU, whereas running SGD on a GPU would nearly double the training time.
- Since the convergence rate is indeed influenced by d, a large NN might not work.
- Using a fixed μ yields better results than the theoretically derived μ.



(a) Convergence Validation



(b) Different Network Architecture



(c) Fixed μ strategy

Experiment Result: Black Box Adversarial Attack

The figure below compares the performance and quality of ZO-SGD and ZO-SVRG-Avg in generating adversarial attacks against a well-trained deep neural network (DNN) model on the MNIST dataset. The attack loss function for the i-th image is defined as:

$$f_i(\mathbf{x}) = c\cdot\max\{F_{y_i}(0.5\cdot\tanh(\tanh^{-1}(2\mathbf{a}_i)+\mathbf{x})) - \max_{j\ne y_i}\{F_j(0.5\cdot\tanh(\tanh^{-1}(2\mathbf{a}_i)+\mathbf{x}))\}, 0\} + \left\|0.5\cdot\tanh(\tanh^{-1}(2\mathbf{a}_i)+\mathbf{x}) - \mathbf{a}_i\right\|_2^2.$$

Here, $(\mathbf{a}_i, y_i)$ represents the i-th natural image $\mathbf{a}_i \in [-0.5, 0.5]^d$ and its corresponding class label y_i. F_{y_i} denotes the model's score function for class y_i and c is some constant coefficient.

It is important to note that we trained a deep neural network model with 1.0 training accuracy and 0.99 validation accuracy (different from the one used in the original paper) as the target for the adversarial attack.

Image ID	4	6	19	24	27	33	42	48	49	56
Original										
ZOSGD Predicted as	9	9	7	9	9	2	9	9	9	5
ZOSVRG($q=1$) Predicted as	2	8	2	2	2	2	8	8	8	2
ZOSVRG($q=10$) Predicted as	9	9	9	9	9	2	9	9	9	9
ZOSVRG($q=20$) Predicted as	9	9	9	9	9	2	9	9	9	9
ZOSVRG($q=30$) Predicted as	5	8	7	9	5	5	7	9	9	5

Figure 2. Comparison of generated adversarial examples from a black-box DNN on MNIST: digit class "4".

Image Id	7	9	12	16	20	58	62	73	78	92
Original										
ZOSGD Predicted as	3	4	4	7	4	4	8	8	4	4
ZOSVRG($q=1$) Predicted as	8	8	4	4	4	4	5	8	8	4
ZOSVRG($q=10$) Predicted as	3	4	4	7	4	4	7	8	4	4
ZOSVRG($q=20$) Predicted as	4	4	4	7	4	4	7	8	4	4
ZOSVRG($q=30$) Predicted as	4	4	4	7	4	4	5	8	8	4

Figure 3. Comparison of generated adversarial examples from a black-box DNN on MNIST: digit class "9".

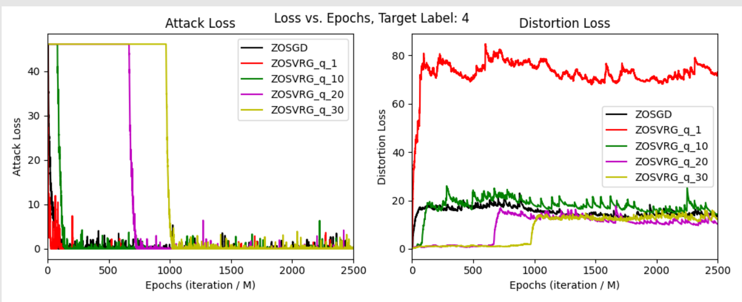


Figure 4. Loss curve of target 4

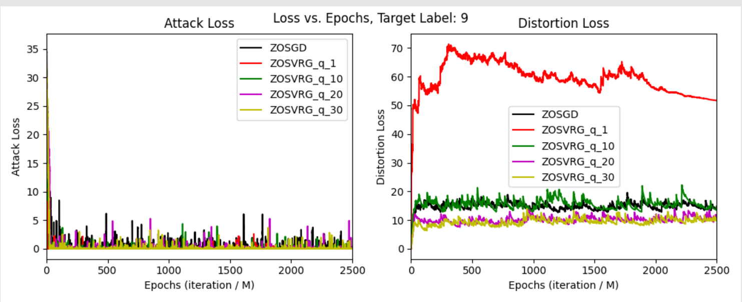


Figure 5. Loss curve of target 9

References

Liu, S., B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini (2018). *Zeroth-Order Stochastic Variance Reduction for Nonconvex Optimization*. arXiv: 1805.10367 [cs.LG]. URL: https://arxiv.org/abs/1805.10367.

Johnson, R. and T. Zhang (2013). "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.