Zeroth-Order Stochastic Projected Gradient Descent for Nonconvex Optimization

Sijia Liu†, Xingguo Li‡, Pin-Yu Chen†, Jarvis Haupt‡, Lisa Amini†

†MIT-IBM Watson AI Lab, IBM Research ‡University of Minnesota

Corresondence: Sijia.Liu@ibm.com

Zeroth-Order (Gradient-Free) Optimization

- Zeroth-order (gradient-free) optimization approximates the full gradient via a random gradient estimate.

Summary of Our Work

We investigate the convergence of ZO stochastic projected gradient descent (ZO-SPGD) for *constrained* convex/nonconvex optimization. Our work is motivated by the ZO proximal gradient algorithm proposed in [1]. However, the ZO gradient estimator considered in [1] is different from our work: we construct the gradient estimate through random direction samples drawn from a bounded uniform distribution rather than a Gaussian distribution in [1]. This analysis leads to different statistics of our random gradient estimator. We establish the following convergence results.

- ZO-SPGD has a $O(\frac{d}{bq\sqrt{T}} + \frac{1}{\sqrt{T}})$ convergence rate to minimize convex (but possibly *non-smooth*) loss functions.
- For constrained *nonconvex* optimization, ZO-SPGD yields a $O(\frac{1}{\sqrt{T}})$ convergence rate up to an *additional error correction* term of order $O(\frac{d+q}{bq})$.

Problem Statement

Consider a constrained finite-sum problem of the form

$$\underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}}\ f(\mathbf{x}) := \tfrac{1}{n}\textstyle\sum_{i=1}^n f_i(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable, $\mathcal{C} \in \mathbb{R}^d$ is a closed convex set, and $\{f_i(\mathbf{x})\}$ are n component functions (not necessarily convex).

We consider the problem setting in which **A1** and/or **A2** are satisfied.

A1: Functions $\{f_i\}$ are L_1-Lipschitz continuous for a finite positive constant L_1.

A2: Functions $\{f_i\}$ are differentiable and have L_2-Lipschitz continuous gradients, where L_2 is a finite positive constant.

A1 allows $\{f_i\}$ to be non-differentiable and implies that subgradients of $\{f_i\}$ are bounded. When **A2** holds, $\{f_i\}$ are restricted to differentiable functions and it implies that

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) \leq \nabla f_i(\mathbf{y})\mathbf{x} - \mathbf{y} + (L_2/2)\|\mathbf{x}-\mathbf{y}\|^2.$$

Random Gradient Estimation

Given an *arbitrary* function f (not necessarily in a finite-sum form), a two-point based average random gradient estimator $\hat{\nabla} f(\mathbf{x})$ is defined by

$$\hat{\nabla} f(\mathbf{x}) = \frac{d}{q}\sum_{j=1}^{q}\frac{f(\mathbf{x}+\mu\mathbf{u}_j) - f(\mathbf{x}-\mu\mathbf{u}_j)}{2\mu}\mathbf{u}_j, \tag{Avg-RandGradEst}$$

where d is the number of optimization variables, $\mu > 0$ is a smoothing parameter, and $\{\mathbf{u}_j\}$ are i.i.d. random directions drawn from a uniform distribution over a unit sphere.

Lemma 1: Statistics of random gradient estimate

Define $f_\mu = \mathbb{E}_{\mathbf{v}\in U_b}[f(\mathbf{x}+\mu\mathbf{v})]$, where U_b denotes a uniform distribution with respect to the unit Euclidean ball. Then Avg-RandGradEst yields the following results:

a) For any $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\hat{\nabla} f(\mathbf{x})\right] = \mathbb{E}_{\mathbf{u}}\left[(d/\mu)f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}\right] = \nabla f_\mu(\mathbf{x}), \tag{2}$$

where $\mathbf{u}$ is a vector picked uniformly at random from the Euclidean unit sphere. Moreover, under assumptions **A1**, the smoothing function f_μ is L_1-Lipschitz continuous. Under **A2**, f_μ has L_2-Lipschitz continuous gradient.

b) Suppose that assumption **A1** holds, for any $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\hat{\nabla} f(\mathbf{x})\|^2\right] \leq \frac{(c_1 d + 4q)L_1^2}{4q}, \tag{3}$$

and under assumption **A2**, for any $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\hat{\nabla} f(\mathbf{x})\|^2\right] \leq 2\left(1+\frac{d}{q}\right)\|\nabla f(\mathbf{x})\|_2^2 + \left(1+\frac{1}{q}\right)\frac{\mu^2 L_2^2 d^2}{2}, \tag{4}$$

where the expectation is taken with respect to random direction vectors $\{\mathbf{u}_j\}$ in Avg-RandGradEst, and c_1 is a numerical constant in (3).

Lemma 1 uncovers important properties of Avg-RandGradEst.

- The use of *multiple* ($q > 1$) random direction vectors $\{\mathbf{u}_j\}$ does not reduce the bias of $\hat{\nabla} f$ (with respect to ∇f). That is because $\hat{\nabla} f$ is unbiased with respect to ∇f_μ, and the distance between ∇f and ∇f_μ is fixed: As $\mu \to 0$, we obtain $\nabla f_\mu(\mathbf{x}) \to \nabla f(\mathbf{x})$. However, if μ is too small, then the function difference could be dominated by the system noise and fails to represent the function differential.
- The variance of the random gradient estimator is reduced as q increases. In particular, a large q mitigates the dimension (d) dependency on the second-order moment of Avg-RandGradEst.

Algorithm

Input: Total number of iterations T, step sizes $\{\eta_k\}_{k=0}^{T-1}$, mini-batch size b, initial iterate $\mathbf{x}_0 \in \mathcal{C}$,
for $k = 0, 1, \ldots, T-1$ **do**
  choose a mini-batch $\mathcal{I}_k$ with b i.i.d. samples from $[n]$
  compute a gradient estimate $\hat{\mathbf{g}}_k = \frac{1}{b}\sum_{i\in\mathcal{I}_k}\hat{\nabla} f_i(\mathbf{x}_k)$
  project onto $\pi_{\mathcal{C}}$

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{C}}\left[\mathbf{x}_k - \eta_k\hat{\mathbf{g}}_k\right] \tag{5}$$

end for
output: $\mathbf{x}_R$ averaged/sampled from $\{\mathbf{x}_k\}_{k=0}^{T-1}$

Convergence Analysis: Convex Case

Theorem 1: Convergence rate of ZO-SPGD for convex optimization

Suppose that assumption **A1** holds and f in problem (1) is convex. Given $\eta_k = \eta$, $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$, and $\mathbf{x}_R = \frac{1}{T}\sum_{k=0}^{T-1}\mathbf{x}_k$ in Algorithm 1, then

$$\mathbb{E}[f(\mathbf{x}_R) - f(\mathbf{x}^*)] \leq \frac{R^2}{\eta T} + \frac{(c_1 d + 4q)L_1^2}{4bq}\eta + L_1^2\eta + 2L_1\mu.$$

In Theorem 1, let $\eta = \frac{1}{\sqrt{T}}$ and $\mu = \frac{1}{\sqrt{T}}$ (milder conditions than many other works), we obtain the convergence rate $O(\frac{d}{bq\sqrt{T}} + \frac{1}{\sqrt{T}})$. We can also conclude that the use of multiple minibatch samples (b) and random direction vectors (q) improves the convergence rate of ZO-SPGD.

Convergence Analysis: Nonconvex Case

For constrained non-convex problems, the convergence of an algorithm at point $\mathbf{x}_k$ can be measured by 'gradient mapping' [2, 1],

$$P_{\mathcal{C}}(\mathbf{x}_k, \nabla f(\mathbf{x}_k), \eta) = (1/\eta)\left[\mathbf{x}_k - \Pi_{\mathcal{C}}\left(\mathbf{x}_k - \eta\nabla f(\mathbf{x}_k)\right)\right]. \tag{6}$$

Interpretation: projected gradient, which offers a feasible update from the previous point $\mathbf{x}_k$,

$$\Pi_{\mathcal{C}}\left(\mathbf{x}_k - \eta\nabla f(\mathbf{x}_k)\right) = \mathbf{x}_k - \eta P_{\mathcal{C}}(\mathbf{x}_k, \nabla f(\mathbf{x}_k), \eta).$$

Our goal is to bound $\mathbb{E}[\|P_{\mathcal{C}}(\mathbf{x}_k, \nabla f(\mathbf{x}_k), \eta)\|_2^2]$.

Convergence Analysis: Nonconvex Case

Proposition 1: Relationship between the variance of a gradient estimator and the convergence rate

If assumption **A2** holds and $\eta_k \in (0, 1/L_2)$, then the outputs $\{\mathbf{x}_k\}_{k=0}^{T-1}$ of Algorithm 1 satisfies

$$\sum_{k=0}^{T-1}\left((2\eta_k - L_2\eta_k^2)\,\mathbb{E}\left[\|P_{\mathcal{C}}(\mathbf{x}_k, \hat{\mathbf{g}}_k, \eta_k)\|^2\right]\right) \leq \sum_{k=0}^{T-1}\left(2\eta_k\mathbb{E}\left[\|\hat{\mathbf{g}}_k - \mathbb{E}[\hat{\mathbf{g}}_k|\mathbf{x}_k]\|^2\right]\right) + 2\mu^2 L_2 + c_2,$$

where $\mathbb{E}$ is taken with respect to all the randomness (e.g., minibatch and random directions), $P_{\mathcal{C}}$ is the gradient mapping given by (6), and $c_2 = 2(f(\mathbf{x}_0) - f(\mathbf{x}^*))$.

Theorem 2: Convergence rate of ZO-SPGD for nonconvex optimization

Suppose that **A1**-**A2** hold, and $\eta_k \in (0, 2/L_2)$. By randomly selecting $\mathbf{x}_R$ from $\{\mathbf{x}_k\}_{k=0}^{T-1}$ with probability

$$P(R=k) = \frac{2\eta_k - L_2\eta_k^2}{\sum_{k=0}^{T-1}(2\eta_k - L_2\eta_k^2)},$$

the convergence rate of Algorithm 1 is given by

$$\mathbb{E}\left[\|P_{\mathcal{C}}(\mathbf{x}_R, \nabla f(\mathbf{g}_R), \eta_R)\|^2\right] \leq \frac{3(c_1 d + 4q)L_1^2(\sum_{k=0}^{T-1}\eta_k)}{2bq\sum_{k=0}^{T-1}(2\eta_k - L_2\eta_k^2)} + \frac{6\mu^2 L_2 + 3c_2}{\sum_{k=0}^{T-1}(2\eta_k - L_2\eta_k^2)} + \frac{3\mu^2 L_2^2 d^2}{4} + \frac{3(c_1 d + 4q)L_1^2}{4bq}. \tag{7}$$

Several insights can be drawn from Theorem 2.

- The first term in the convergence rate (7) is bounded by by $O(\frac{d+q}{bq})$ up to a constant factor.
- If we choose the constant stepsize $\eta = \frac{c_\eta}{\sqrt{T}} \in (0, 1/L_2)$ for some constant c_η and $\mu = \frac{1}{d^{1/2}(bq)^{1/2}}$, then Theorem 1 implies the convergence rate $O(\frac{1}{\sqrt{T}} + \frac{d+q}{bq})$.

References

[1] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[2] S. J. Reddi, S. Sra, B. Poczos, and A. J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *NIPS*, pages 1145–1153, 2016.

[3] S. Liu, J. Chen, P.-Y. Chen, and A. O. Hero. Zeroth-order online ADMM: Convergence analysis and applications. In *AISTATS*, volume 84, pages 288–297, April 2018.

[4] S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *NIPS*, 2018.