Non-convex Finite-Sum Optimization Via SCSG Methods

Lihua Lei , Cheng Ju , Jianbo Chen & Michael I. Jordan

Problem Setup

Unconstrained Finite-sum Objective:

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

Assumptions

A1 $\nabla f_i(x)$ is uniformly bounded by $\mathcal{H}^*$ for all $i \in [n]$ and $x \in \mathbb{R}^d$;

A2 $\nabla f_i(x)$ is L-Lipschitz for all $i \in [n]$.

A3 (Polyak-Lojasiewicz condition, optional):

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - \min f(x)).$$

Goal: Find an ϵ-approximate first-order stationary point y such that

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

SCSG

Outer-Loop Update

Inputs: Initial value $\tilde{x}_0$, number of epochs T, step-sizes $\{\eta_j\}_{j=1}^T$, block sizes $\{B_j\}_{j=1}^T$, mini-batch sizes $\{b_j\}_{j=1}^T$.

Procedure:

```
1: for t = 1, 2, ..., T do
2:   x̃_j ← SCSGepoch(x̃_{j−1}; B_j, b_j, η_j)
3: end for
```

Output: Sample $\tilde{x}_T^*$ from $(\tilde{x}_j)_{j=1}^T$ with $P(\tilde{x}_T^* = \tilde{x}_j) \propto \eta_j B_j / b_j$

Inner-Loop/Within-Epoch Update (no mini-batch, i.e. $b_j \equiv 1$)

SVRGepoch

1. Input: x_0, η
2. $\mathcal{I} \leftarrow [n]$
3. $g \leftarrow \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} f_i'(x_0)$
4. Generate $N \leftarrow n$
5. **for** $k = 1, 2, \cdots, N$ **do**
6. Randomly pick $i \in [n]$
7. $\nu \leftarrow f_i'(x) - f_i'(x_0) + g$
8. $x \leftarrow x - \eta\nu$
9. **end for**
10. Output: x_N

SCSGepoch

1. Input: x_0, η, B
2. Randomly pick $\mathcal{I}$ with size B
3. $g \leftarrow \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} f_i'(x_0)$
4. Gen. $N \sim$ Geo with $\mathbb{E}N = B$
5. **for** $k = 1, 2, \cdots, N$ **do**
6. Randomly pick $i \in [n]$
7. $\nu \leftarrow f_i'(x) - f_i'(x_0) + g$
8. $x \leftarrow x - \eta\nu$
9. **end for**
10. Output: x_N

Theoretical Results

Theorem 1. *Let $\eta_j L = \gamma(B_j/b_j)^{-\frac{2}{3}}$. Suppose $\gamma \leq \frac{1}{6}$ and $B_j \geq 9$ for all j, then under Assumption* **A1**,

$$\mathbb{E}\|\nabla f(\tilde{x}_j)\|^2 \leq \frac{5L}{\gamma}\cdot\left(\frac{b_j}{B_j}\right)^{\frac{1}{3}}\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{6I(B_j < n)}{B_j}\cdot\mathcal{H}^*.$$

Magic of the geometric distribution

$$N \sim \mathrm{Geo}(\gamma) \Longrightarrow \mathbb{E}(W_N - W_{N+1}) = \frac{1-\gamma}{\gamma}(W_1 - \mathbb{E}W_N), \quad \forall W_1, W_2, \ldots$$

Main challenge in the analysis: ν is no longer unbiased:

$$\mathbb{E}\nu = \nabla f(x) + e, \quad e = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\nabla f_i(x) - \nabla f(x).$$

Parameter settings analyzed in the paper:

η_j	B_j	b_j	Type of Objectives	Computation Cost
$\frac{1}{2LB^{2/3}}$	$O\left(\frac{1}{\epsilon}\wedge n\right)$	1	Smooth	$O\left(\frac{1}{\epsilon^{5/3}}\wedge\frac{n^{2/3}}{\epsilon}\right)$
$\frac{1}{2LB_j^{2/3}}$	$j^{\frac{3}{2}}\wedge n$	1	Smooth	$\tilde{O}\left(\frac{1}{\epsilon^{5/3}}\wedge\frac{n^{2/3}}{\epsilon}\right)$
$\frac{1}{2LB_j^{2/3}}$	$O\left(\frac{1}{\mu\epsilon}\wedge n\right)$	1	Polyak-Lojasiewicz	$\tilde{O}\left((\frac{1}{\mu\epsilon}\wedge n) + \frac{1}{\mu}(\frac{1}{\mu\epsilon}\wedge n)^{2/3}\right)$

Comparison With Other First-Order Methods

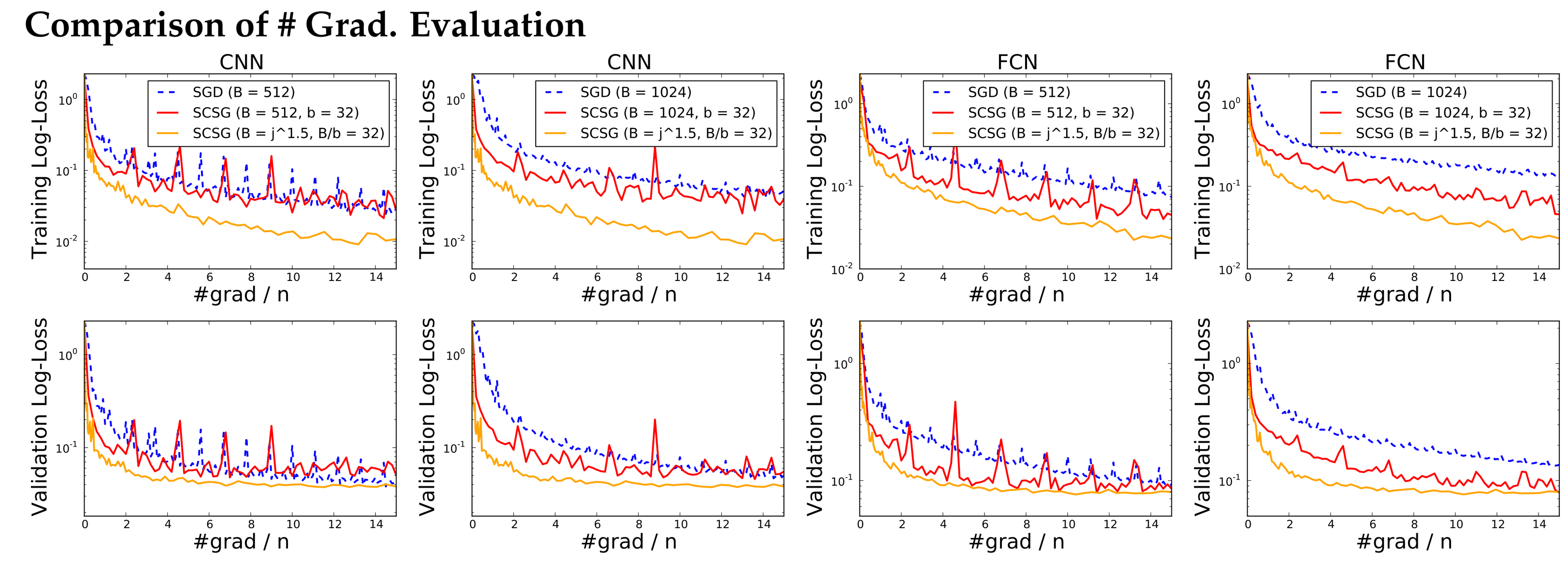
	Smooth			P-L
	General	$\epsilon\sim n^{-1/2}$	$\epsilon\sim n^{-1}$	General
Gradient Methods				
GD	$O\left(\frac{n}{\epsilon}\right)$	$O\left(n^{3/2}\right)$	$O\left(n^2\right)$	$\tilde{O}\left(\frac{n}{\mu}\right)$
Best available	$\tilde{O}\left(\frac{n}{\epsilon^{5/6}}\right)$	$\tilde{O}\left(n^{17/12}\right)$	$\tilde{O}\left(n^{11/6}\right)$	-
Stochastic Gradient Methods				
SGD	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(n\right)$	$O\left(n^2\right)$	$O\left(\frac{1}{\mu^2\epsilon}\right)$
Best available	$O\left(n+\frac{n^{2/3}}{\epsilon}\right)$	$O\left(n^{7/6}\right)$	$O\left(n^{5/3}\right)$	$\tilde{O}\left(n+\frac{n^{2/3}}{\mu}\right)$
SCSG	$\tilde{O}\left(\frac{1}{\epsilon^{5/3}}\wedge\frac{n^{2/3}}{\epsilon}\right)$	$\tilde{O}\left(n^{5/6}\right)$	$\tilde{O}\left(n^{5/3}\right)$	...

- SCSG is the first algorithm that is provably better than SGD;
- SCSG is never worse than any stochastic gradient method in all regimes;
- SCSG is never worse than any gradient method in practical regimes;
- SCSG is the only algorithm that has sub-linear (to n) complexity in the practical regime $\epsilon \sim n^{-1/2}$.

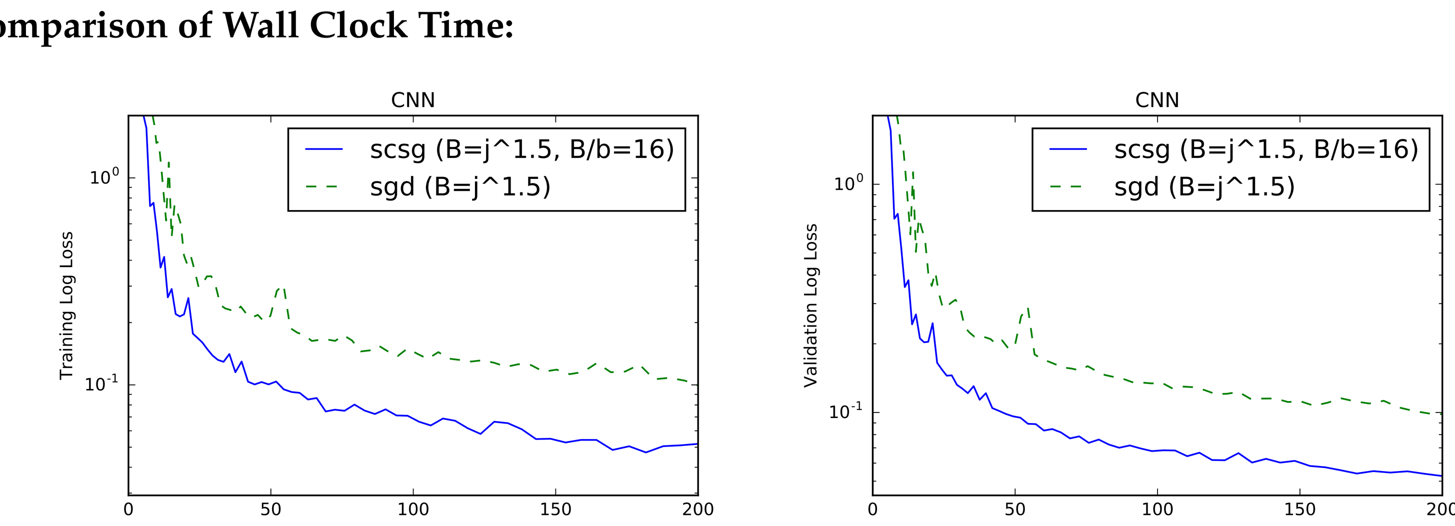
Experiments

- MNIST dataset (50000 training, 10000 testing);
- A three-layer fully-connected neural network (FCN for short);
- A standard convolutional neural network LeNet (CNN for short).

Comparison of # Grad. Evaluation



Comparison of Wall Clock Time:



SCSG for Stochastic Optimization

It is almost direct to extend SCSG to the general stochastic optimization where the objective is

$$f(x) = \mathbb{E}_{\xi\sim G}F(x;\xi).$$

Modify the algorithm by

- Line 2: Gen. $\xi_1^*, \ldots, \xi_B^* \overset{i.i.d.}{\sim} G$;
- Line 3: $g \leftarrow \frac{1}{B}\sum_{i=1}^{B}\nabla F(x_0;\xi_i^*)$;
- Line 6: Gen. $\xi_k \in [n]$;
- Line 7: $\nu \leftarrow F(x;\xi_k) - F(x_0;\xi_k) + g$.

Complexity results:

- Smooth Case (Assumptions **A1** - **A2**): $\tilde{O}\left(\frac{1}{\epsilon^{5/3}}\right)$;
- P-L Case (Assumptions **A1** - **A3**): $\tilde{O}\left(\frac{1}{\mu\epsilon} + \frac{1}{\mu^{5/3}\epsilon^{2/3}}\right)$.