Less than a Single Pass: Stochastically Controlled Stochastic Gradient

Lihua Lei & Michael I. Jordan

Problem Setup

Composite Objectives

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

Assumptions

A1 f_i are smooth with L-Lipschitz gradients;

A2 f_i are strongly convex with modulus $\mu \geq 0$. ($\mu = 0$ is allowed.)

$f_i(x)$ can be deterministic/dependent/not identically distributed!

Measure of Intrinsic Difficulty

Question: how difficult is the problem?

Initialization	$\Delta_0 = L\|x_0 - x^*\|^2, D_0 = f(x_0) - f(x^*)$
Curvature	$\kappa = L/\mu$ (when $\mu > 0$)
Gradient Regularity	$G^2 = \max_i \sup_x \|\nabla f_i(x)\|^2$
Heterogeneity	$\mathcal{H}^* = \sup_x \frac{1}{nL}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2$

A toy example: $f_i(x) = (x - b_i)^2$ with

1. $b = (b_1, \ldots, b_n) \in \mathcal{B}_0 = \{b : b_1 = b_2 = \ldots = b_n \in [-1, 1]\}$;
2. $b = (b_1, \ldots, b_n) \in \mathcal{B}_1 = \{b : b_1, b_2, \ldots, b_n \in [-1, 1]\}$.

$\mathcal{B}_1$ is strictly harder than $\mathcal{B}_0$: heterogeneity matters!

$\mathcal{H}^*$ is too conservative: $\mathcal{H}^* = \infty$ in *various* realistic situations with unbounded domain! In our work, we define

$$\mathcal{H} = \inf_{x^* \in \arg\min f(x)} \frac{1}{nL}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2.$$

(i) $\mathcal{H} \leq \min\{\mathcal{H}^*, G^2/L\}$;

(ii) $\mathcal{H} = O_p(1)$ if $f_i(x)$ are i.i.d. (under standard conditions);

(iii) $\mathcal{H}$ can be efficiently estimated for *Generalized Linear Models*;

(iv) When $\mathcal{H} \ll n$ and ϵ is moderate, optimizing $f(x)$ with *less than a single pass* of data is possible!

SCSG

Outer-Loop Update

Inputs: Initial value $\tilde{x}_0$, block size B, stepsize η, number of epochs T, other parameter γ (non-strongly convex) or m (strongly convex)

Procedure:

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $\quad\tilde{x}_j \leftarrow$ SCSGepoch$(\tilde{x}_{j-1}; B, \eta, \gamma, m)$
3: **end for**

Output: (non-strongly convex) $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_t$; (strongly convex) $\tilde{x}_T$.

Inner-Loop/Within-Epoch Update

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 $N \sim U([n])$
5: **for** $k = 1, 2, \cdots, N$ **do**
6: $\quad$ Randomly pick $i \in [n]$
7: $\quad \nu \leftarrow f_i'(x) - f_i'(x_0) + g$
8: $\quad x \leftarrow x - \eta\nu$
9: **end for**
10: Output: x_N

SCSGepoch

1: Input: x_0, η, B, γ (or m)
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: Generate $N \sim \mathrm{Geo}(\gamma)/U([m])$
5: **for** $k = 1, 2, \cdots, N$ **do**
6: $\quad$ Randomly pick $i \in \mathcal{I}$
7: $\quad \nu \leftarrow f_i'(x) - f_i'(x_0) + g$
8: $\quad x \leftarrow x - \eta\nu$
9: **end for**
10: Output: x_N

Magic of The Geometric Distribution

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

Computation Cost

Parameter Defaults of SCSG

1. Non-strongly Convex: $\eta < \frac{1}{5L}$, $\quad B \geq n \wedge \frac{2\mathcal{H}}{\epsilon}$, $\quad \gamma = \frac{B-1}{B}$
2. Strongly Convex: $\quad \eta < \frac{1}{5(L+\mu)}$, $\quad B \geq n \wedge \frac{8\mathcal{H}}{\epsilon}$, $\quad m \geq \frac{1}{2L\mu\eta^2}$

	Non-strongly Convex	Strongly Convex
SGD	$O\left(\frac{\mathcal{H}^*}{\epsilon^2}\right)$	$O\left(\frac{\mathcal{H}^*\kappa}{\epsilon}\log\frac{1}{\epsilon}\right)$
AGD	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$	$O\left(n\sqrt{\kappa}\log\frac{1}{\epsilon}\right)$
SVRG	-	$O\left((n+\kappa)\log\frac{1}{\epsilon}\right)$
Katyusha	$O\left(n\log\frac{1}{\epsilon} + \sqrt{\frac{n}{\epsilon}}\right)$	$O\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon}\right)$
SCSG	$O\left(\frac{\mathcal{H}}{\epsilon^2} \wedge \frac{n}{\epsilon}\right)$	$O\left(\left[\frac{\mathcal{H}}{\epsilon} \wedge n + \kappa\right]\log\frac{1}{\epsilon}\right)$
SCSG+	$O\left(\frac{1}{\epsilon}\log\left(\frac{\mathcal{H}}{\epsilon}\wedge n\right) + \sqrt{\frac{\mathcal{H}\log n}{n\epsilon^3}}\right)$	$O\left(\kappa e^{\sqrt{\log\frac{1}{\epsilon}}} + \sqrt{\frac{\kappa}{\epsilon}}\right)$

Communication Cost

- When $B \leq$ memory limit $\ll n$, communication matters!
- Even sampling a data is costly (SVRG/its variants are inefficient);
- SCSG is efficient since SCSGepoch is implemented in memory.

	General Convex	Strongly Convex	Dimension
SCSG	$O\left(\left(n\wedge\frac{\mathcal{H}}{\epsilon}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\left(n\wedge\frac{\mathcal{H}}{\epsilon}\right)\log\frac{1}{\epsilon}\right)$	$O(d)$
CoCoA	-	$O\left(m^2\cdot\frac{n+\kappa}{n\wedge\frac{\mathcal{H}}{\epsilon}+\kappa}\log\frac{1}{\epsilon}\right)$	$O(d)$
DANE	-	$O\left(m\kappa\log\frac{1}{\epsilon}\right)$	$O(d^2)$
DiSCO	-	$O\left(m\sqrt{\kappa}\log\kappa\log\frac{1}{\epsilon}\right)$	$O(d^2)$

More Details on $\mathcal{H}$

Generalized Linear Models: $f_i(x) = \rho(y_i, a_i^T x)$:

$$L\mathcal{H} \leq \sup_{z,w}\rho_2^2(z,w)\cdot\frac{1}{n}\sum_i\|a_i\|^2, \quad L \leq \sup_{z,w}\rho_{22}(z,w)\cdot\max_i\|a_i\|^2.$$

1. Multi-class logistic regression: $L\mathcal{H} \leq \frac{2}{n}\sum_i\|a_i\|^2$, $L \leq \max_i\|a_i\|^2$;
2. Linear regression: $L\mathcal{H} \leq \frac{\|y\|^2}{n}\cdot\max_i\|a_i\|^2$, $L \leq \max_i\|a_i\|^2$.

Experiments

- Multi-class logistic regression on MNIST ($L = 292.82, \mathcal{H} = 0.6$);
- Default parameters: $B = 2220, \eta_0 = 0.0017$;
- A 2.6M memory with 8 accesses of the disk is sufficient ($\epsilon = 10^{-3}$).