Acceleration of SVRG and Katyusha X by Inexact Preconditioning

Yanli Liu, Fei Feng, and Wotao Yin

University of California, Los Angeles

ICML 2019

Background

We focus on solving

$$\operatorname{minimize} F(x) = f(x) + \psi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x),$$

where $x \in \mathbb{R}^{\mathrm{d}}$, $f(x)$ is strongly convex and smooth, $\psi(x)$ is convex, and can be non-differentiable. n is large and $d = o(n)$.

Examples: Lasso, Logistic regression, PCA...

Common solvers: SVRG, Katyusha X (a Nesterov-accelerated SVRG), SAGA, SDCA,...

Challenge: As first-order methods, they suffer from ill-conditioning.

In this talk

In this work, we propose to accelerate SVRG and Katyusha X by simple yet effective preconditioning.

Acceleration is demonstrated both theoretically and numerically ($7\times$ runtime speedup on average).

iPreSVRG

SVRG:

$$w_{t+1} = \arg\min_{y \in \mathbb{R}^d} \{\psi(y) + \frac{1}{2\eta}\|y - w_t\|^2 + \langle \tilde{\nabla}_t, y \rangle\},$$

where $\tilde{\nabla}_t$ is a variance-reduced stochastic gradient of $f = \frac{1}{n}\sum f_i$.

Inexact Preconditioned SVRG (iPreSVRG):

$$w_{t+1} \approx \arg\min_{y \in \mathbb{R}^d} \{\psi(y) + \frac{1}{2\eta}\|y - w_t\|_M^2 + \langle \tilde{\nabla}_t, y \rangle\}$$

The preconditioner $M \succ 0$ approximates the Hessian of f.

The subproblem is solved highly inexactly by applying FISTA a fixed number of times.

This acceleration technique also applies to Katyusha X.

Choosing M for Lasso

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2n}\|Ax-b\|_2^2+\lambda_1\|x\|_1+\lambda_2\|x\|_2^2.$$

Two choices of M for Lasso:

1. When d is small, we choose
$$M_1=\frac{1}{n}A^TA,$$
this is the exact Hessian of the first part.
2. When d is large and A^TA is almost diagonally dominant, we choose
$$M_2=\frac{1}{n}\text{diag}(A^TA)+\alpha I,$$
where $\alpha>0$.

Lasso results

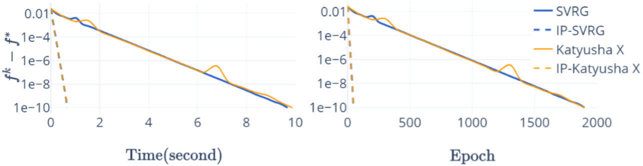

Figure 1: australian dataset[1], $d = 14$, $M = M_1$, 10× runtime speedup

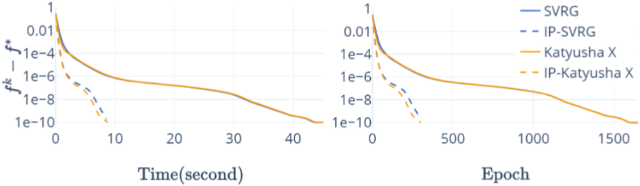

Figure 2: w1a.t dataset[1], $d = 300$, $M = M_2$, 5× runtime speedup

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Choosing M for Logistic

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}\ln(1+\exp(-b_i\cdot a_i^T x))+\lambda_1\|x\|_1+\lambda_2\|x\|_2^2.$$

Let $B = \text{diag}(b)A = \text{diag}(b)(a_1, a_2, ..., a_n)^T$.

Two choices of M for logistic regression:

1. When d is small, we choose
$$M_1 = \frac{1}{4n}B^T B,$$
this is approximately the Hessian of the first part.
2. When d is large and $B^T B$ is almost diagonally dominant, we choose
$$M_2 = \frac{1}{4n}\text{diag}(B^T B) + \alpha I,$$
where $\alpha > 0$.

Logistic results

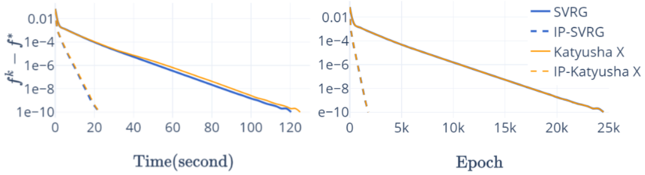

Figure 3: `australian` dataset, $d = 14$, $M = M_1$, 6× runtime speedup

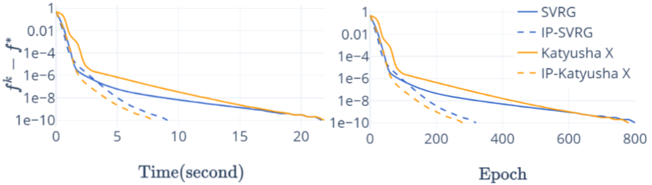

Figure 4: `w1a.t` dataset, $d = 300$, $M = M_2$, 4× runtime speedup

Theoretical Speedup

Theorem 1

Let $C_1(m,\varepsilon)$ and $C_1'(m,\varepsilon)$ be the gradient complexities of SVRG and iPreSVRG to reach $\varepsilon-$suboptimality, respectively. Here m is the epoch length.

1. *When $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f < n^2 d^{-2}$, we have*

$$\frac{\min_{m\geq 1} C_1'(m,\varepsilon)}{\min_{m\geq 1} C_1(m,\varepsilon)} \leq \mathcal{O}(\frac{n^{\frac{1}{2}}}{\kappa_f}).$$

2. *When $\kappa_f > n^{\frac{1}{2}}$ and $\kappa_f > n^2 d^{-2}$, we have*

$$\frac{\min_{m\geq 1} C_1'(m,\varepsilon)}{\min_{m\geq 1} C_1(m,\varepsilon)} \leq \mathcal{O}(\frac{d}{\sqrt{n\kappa_f}}).$$

iPreKatX has a similar speedup.

Conclusions

1. In this work, we apply inexact preconditioning on SVRG and Katyusha X.
2. With appropriate preconditioners and fast subproblem solvers, we obtain significant speedups in both theory and practice.

Poster: Today 6:30 PM – 9:00 PM, Pacific Ballroom #192

Code: `https://github.com/uclaopt/IPSVRG`