Optimal Stochastic Strongly Convex Optimization with a Logarithmic Number of Projections

Jianhui Chen[1], Tianbao Yang[2], Qihang Lin[2], Lijun Zhang[3], and Yi Chang[4]

July 18, 2016

Yahoo Research[1], The University of Iowa[2], Nanjing University[3]

Table of contents

Problem Settings

Strongly Convex Optimization

We consider the constrained optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^d / \mathbb{R}^{p \times q}} \quad & f(x) \\ \text{s.t.} \quad & c(x) \leq 0, \end{aligned}$$

where $c(x)$ is convex and $f(x)$ is β-strongly convex.

- A stochastic access model for $f(\cdot)$, i.e., $E[g(x)] \in \partial f(x)$
- A full access to the (sub)gradient of $c(\cdot)$

Convex and Strongly Convex

- Convex in $c(x)$

$$c(x) \geq c(\widehat{x}) + \nabla c(x)^T (x - \widehat{x}) \tag{1}$$

- β-Strongly Convex in $f(x)$

$$f(x) \geq f(\widehat{x}) + \nabla f(x)^T (x - \widehat{x}) + \frac{\beta}{2} \|x - \widehat{x}\|^2, \tag{2}$$

which implies

$$f(x) \geq f(x_*) + \frac{\beta}{2} \|x - x_*\|^2. \tag{3}$$

Examples from Machine Learning

- Constrained Lasso

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(a_i, b_i, w\right) = \frac{1}{n}\sum_{i=1}^{n}\left(a_i^T w - b_i\right)^2$$

$$c(w) = \sum_{j=1}^{d}|w_i| - \lambda$$

- Large Margin Nearest Neighbor Classification Formulation

$$\begin{aligned} f(A) &= \frac{1}{n}\sum_{j=1}^{n}\ell(A, x_1^j, x_2^j, x_3^j) \\ &= \frac{1}{n}\sum_{j=1}^{n}\max(0, \|x_1^j - x_2^j\|_A^2 - \|x_1^j - x_3^j\|_A^2 + 1) \\ c(A) &= A - \epsilon I \end{aligned}$$

- Adding a L_2 regularization term, i.e., $\|w\|^2$ or $\|A\|_F^2$, to attain strong convexity.

Standard SGD for Solving Eq. (1)

- Iterate the following step

$$x_{t+1} = P_{\{c(x)\leq 0\}}\left[x_t - \eta_t g(x_t)\right], \tag{4}$$

where $P_D[\widehat{x}]$ is a projection operator defined as

$$P_D[\widehat{x}] = \arg\min_{x\in D} \|x - \widehat{x}\|_2^2. \tag{5}$$

- Return the final solution as

$$\widehat{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t. \tag{6}$$

Limitations in SGD

The computation in $P_D[\widehat{x}] = \arg\min_{x \in D} \|x - \widehat{x}\|_2^2$ may be expensive if $c(x)$ is complex.

- Popular types of D as $\{x \in R^{d \times d} : 0 \preceq x \preceq \epsilon I\}$ and $\{x \in R^d : Ax \leq b\}$
- A projection onto a PSD cone

$$\begin{aligned} \min_{x \in R^{d \times d}} \quad & \|x - \widehat{x}\|_2^2 \\ \text{s.t.} \quad & 0 \preceq x \preceq \epsilon I \end{aligned} \tag{7}$$

has the complexity of order $\mathcal{O}(d^3)$.

The proposed Epro-SGD and its Proximal Variant

Proposed Epro-SGD Approach

The standard SGD solves

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & c(x) \leq 0 \end{aligned} \tag{8}$$

Our proposed Epro-SGD (Epoch-Projection SGD) considers to minimize an augmented function

$$f(x) + \lambda [c(x)]_+. \tag{9}$$

- $[s]_+$ is a hinge operator defined as $[s]_+ = s$ if $s \geq 0$, and $[s]_+ = 0$ otherwise.
- λ is a prescribed parameter (our analysis shows it has to satisfy $\lambda > G_1/\rho$).

Proposed Epro-SGD Approach

Key ideas in Epro-SGD

- In the inner loop, iteratively optimize $f(x) + \lambda[c(x)]_+$, i.e,,

$$x_{t+1} = x_t - \eta \{g(x_t) + \lambda \partial [c(x_t)]_+\}$$

- In the outer loop, compute the projection $\widetilde{x}_T = P_D[\widehat{x}_T]$

$$\widetilde{x}_T = \arg\min_{x \in D} \|x - \widehat{x}\|_2^2, \ \widehat{x} = \frac{1}{T} \sum_{i=1}^{T} x_i.$$

Main Advantage

- A projection is computed after one epoch (one inner loop).
- The optimal convergence rate can be obtained.

Main Algorithms

1: **Initialization:** $x_1^1 \in D$ and $k = 1$
2: **while** $\sum_{i=1}^{k} T_i \leq T$
3:     **for** $t = 1, \ldots, T_k$
4:         Compute a stochastic gradient $g(x_t^k)$
5:         Compute $x_{t+1}^k = x_t^k - \eta_k(g(x_t^k) + \lambda\partial[c(x_t^k)]_+)$
6:     **endfor**
7:     Compute $\widetilde{x}_T^k = P_D[\widehat{x}_T^k]$, where $\widehat{x}_T^k = \sum_{t=1}^{T_k} x_t^k / T_k$
8:     Update $x_1^{k+1} = \widetilde{x}_T^k,\ T_{k+1} = 2T_k, \eta_{k+1} = \eta_k/2$
9:     Set $k = k + 1$
10: **endwhile**

- Line 3 - 6: inner loop
- Line 2 - 10: outer loop

Convergence Analysis

Assumptions

A1. The stochastic subgradient $g(x)$ is uniformly bounded by G_1, i.e., $\|g(x)\|_2 \leq G_1$.

A2. The subgradient $\partial c(x)$ is uniformly bounded by G_2, i.e., $\|\partial c(x)\|_2 \leq G_2$.

A3. There exists a positive value $\rho > 0$ such that

$$\left[\min_{c(x)=0,\, v\in\partial c(x),\, v\neq 0} \|v\|_2\right] \geq \rho.$$

Remarks on A3

- For any $\widehat{x}$, let $\widetilde{x} = \arg\min_{c(x)\leq 0} \|x - \widehat{x}\|_2^2$.

$$\|\widehat{x} - \widetilde{x}\|_2 \leq \frac{1}{\rho}[c(\widehat{x})]_+, \quad \rho > 0. \tag{10}$$

- Eq. (10) ensures that the projection of a point onto a feasible domain does not deviate too much from this intermediate point.

Convergence Analysis

Under Assumptions A1$\sim$A3, we derive

- Expected convergence bounds
- High-probability convergence bounds

all with optimal rates for strongly convex optimization.

Expected Convergence Bound

Under Assumptions A1∼A3 and given that $f(x)$ is β-strongly convex, if we let $\mu = \rho/(\rho - G_1/\lambda)$, $G^2 = G_1^2 + \lambda^2 G_2^2$, and set $T_1 = 8, \eta_1 = \mu/(2\beta)$, the total number of epochs $k^\dagger$ is given by

$$k^\dagger = \left\lceil \log_2\left(\frac{T}{8} + 1\right)\right\rceil \leq \log_2\left(\frac{T}{4}\right), \tag{11}$$

the solution $x_1^{k^\dagger+1}$ enjoys a convergence rate of

$$E[f(x_1^{k^\dagger+1})] - f(x_*) \leq \frac{32\mu^2 G^2}{\beta(T+8)}, \tag{12}$$

and $c(x_1^{k^\dagger+1}) \leq 0$.

High Probability Bound

Under Assumptions A1∼A3 and given $\|x_t - x_*\|_2 \leq D$ for all t. If we let $\mu = \rho/(\rho - G_1/\lambda)$, $G^2 = G_1^2 + \lambda^2 G_2^2$,
$C = \left(8G_1^2/\beta + 2G_1 D\right)\ln(m/\epsilon) + 2G_1 D$, and set
$T_1 \geq \max\left(3C\beta/\left(\mu G^2\right), 9\right)$, $\eta_1 = \mu/(3\beta)$, the total number of epochs $k^\dagger$ is given by

$$k^\dagger = \left\lfloor \log_2\left(\frac{T}{T_1} + 1\right)\right\rfloor \leq \log_2(T/4),$$

and the final solution $x_1^{k^\dagger+1}$ enjoys a convergence rate of

$$f(x_1^{k^\dagger+1}) - f(x_*) \leq \frac{4T_1\mu^2 G^2}{\beta(T+T_1)}$$

with a probability at least $1-\delta$, where $m = \lceil 2\log_2 T\rceil$.

Limitations in Epro-SGD

- The proposed Epro-SGD introduces an augmented objective function

$$f(x) + \lambda[c(x)]_+$$

and optimize it in the inner loop as

$$x_{t+1} = x_t - \eta \left\{ g(x_t) + \lambda \partial [c(x_t)]_+ \right\}.$$

- The desirable structure of the objective function, for example, $f(x) = \frac{1}{n} \sum_{i=1}^{n} (a_i^T x - b_i)^2 + \gamma \|x\|_1$, is not exploited.

Proximal Variant of Epro-SGD

- Propose a proximal variant to exploit the desirable structure.
- Denote the objective function by

$$f(x) = h(x) + k(x),$$

  where $k(x)$ embeds the structure of interest.
- The proposed Epro-SGD proximal variant introduces an augmented objective function as

$$h(x) + \lambda[c(x)]_+ + k(x). \tag{13}$$

Proximal Variant of Epro-SGD

Key ideas

- In the inner loop, iteratively optimize $h(x) + \lambda[c(x)]_+ + k(x)$, i.e,,

$$x_{t+1} = \arg\min_x \frac{1}{2} \|x - [x_t - \eta(g(x_t) + \lambda\partial[c(x_t)]_+)]\|_2^2 + \eta k(x).$$

- In the outer loop, compute the projection $\widetilde{x}_T = P_D[\widehat{x}_T]$

$$\widetilde{x}_T = \arg\min_{x \in D} \|x - \widehat{x}\|_2^2,\ \widehat{x} = \frac{1}{T}\sum_{i=1}^{T} x_i.$$

Expected Convergence Bound

Under Assumptions A1∼A3 and given that $\widehat{f}(x)$ is β-strongly convex, if we let $\mu = \rho/(\rho - G_1/\lambda)$ and $G = 3G_1 + 2\lambda G_2$, and set $T_1 = 16$, $\eta_1 = \mu/\beta$, then the total number of epochs $k^\dagger$ is given by

$$k^\dagger = \left\lfloor \log_2 \left(\frac{T}{17} + 1 \right) \right\rfloor \leq \log_2(T/8),$$

and the final solution $x_1^{k^\dagger+1}$ enjoys a convergence rate of

$$E[\widehat{f}(x_1^{k^\dagger+1})] - \widehat{f}(x_*) \leq \frac{68\mu^2 G^2}{\beta(T+17)},$$

and $c(x_1^{k^\dagger+1}) \leq 0$.

Comparisons and Experiments

Comparison with Competing Algorithms

Algorithms	Convergence Rate	Project Number
Standard SGD (SGD)	$\mathcal{O}(logT/T)$	$\mathcal{O}(T)$
One-Projection SGD (OneProj)	$\mathcal{O}(logT/T)$	1
logT-projection SGD (logT)	$\mathcal{O}(1/T)$	$\mathcal{O}(\kappa \log T)$
Epro-SGD	$O(1/T)$	$\mathcal{O}(\log T)$

- In SGD, OneProj, and Epro-SGD, η_t is set to $1/(\lambda t)$.
- In LogT, η_t is set to $1/(\sqrt{6}L)$ as suggested in the original paper.

Experiments

- Solve L1-norm constrained least squares optimization problem

$$\begin{aligned} \min_{w} \quad & \frac{1}{2N}\sum_{i=1}^{N}\left(x_i^T w - y_i\right)^2 + \alpha\|w\|^2 \\ s.t. \quad & \|w\|_1 \leq \beta. \end{aligned}$$

- Compare SGD, OneProj, logT, Epro-SGD, in terms of objective values, and the required computation time.

Experiments

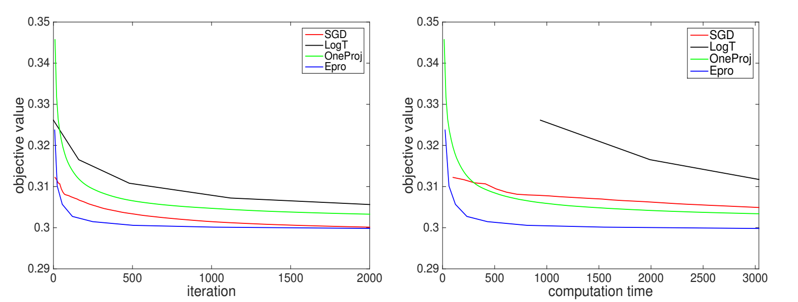

Figure 1: Empirical comparison of the four competing methods for solving the constrained Lasso. (1) Left plot: the change of the objective values with respect to the iteration number. (2) Right plot: the change of the objective values with respect to the computation time (in seconds).

LMNN

- Solve the large margin nearest neighbor (LMNN) classification formulation

$$
\begin{aligned}
\min_{A} \quad & \frac{c}{N}\sum_{j=1}^{N} \ell\left(A, x_1^j, x_2^j, x_3^j\right) + (1-c)tr(AL) \\
& + \frac{\mu_1}{2}\|A\|_F^2 + \mu_2\|A\|_1^{\text{off}} \\
s.t. \quad & A \succeq \epsilon I,
\end{aligned} \tag{14}
$$

- Compare SGD, OneProj, logT, Epro-SGD, in terms of objective values, and the required computation time.

Experiments

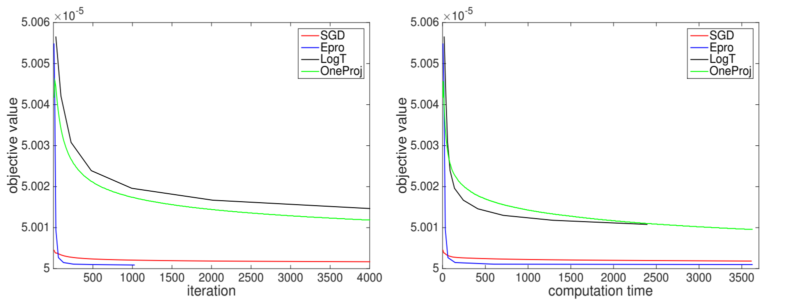

Figure 2: Empirical comparison of the four competing methods for solving LMNN. (1) Left plot: the change of the objective values with respect to the iteration number. (2) Right plot: the change of the objective value with respect to the computation time.

Thank you!