Optimal Stochastic Strongly Convex Optimization with a Logarithmic Number of Projections

Jianhui Chen¹, Tianbao Yang², Qihang Lin², Lijun Zhang³, and Yi Chang⁴
July 18, 2016

Yahoo Research¹, The University of Iowa², Nanjing University³
Table of contents

1. Problem Settings

2. The proposed Epro-SGD and its Proximal Variant

3. Comparisons and Experiments
Problem Settings
We consider the constrained optimization problem

$$
\min_{x \in \mathbb{R}^d / \mathbb{R}^p \times q} f(x)
$$

s.t. \quad c(x) \leq 0,

where $c(x)$ is convex and $f(x)$ is $\beta$-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 $c(x)$

\[ c(x) \geq c(\hat{x}) + \nabla c(x)^T (x - \hat{x}) \]  

- $\beta$-Strongly Convex in $f(x)$

\[ f(x) \geq f(\hat{x}) + \nabla f(x)^T (x - \hat{x}) + \frac{\beta}{2} \|x - \hat{x}\|^2 \]  

which implies

\[ f(x) \geq f(x^\ast) + \frac{\beta}{2} \|x - x^\ast\|^2. \]
Examples from Machine Learning

- **Constrained Lasso**

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

  \[
  c(w) = \sum_{j=1}^{d} |w_j| - \lambda
  \]

- **Large Margin Nearest Neighbor Classification Formulation**

  \[
  f(A) = \frac{1}{n} \sum_{j=1}^{n} \ell(A, x_1^j, x_2^j, x_3^j)
  \]

  \[
  f(A) = \frac{1}{n} \sum_{j=1}^{n} \max(0, \|x_1^j - x_2^j\|_A^2 - \|x_1^j - x_3^j\|_A^2 + 1)
  \]

  \[
  c(A) = A - \epsilon I
  \]

- 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

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

where \( P_D[\hat{x}] \) is a projection operator defined as

\[ P_D[\hat{x}] = \arg \min_{x \in D} \| x - \hat{x} \|_2^2. \quad (5) \]

- Return the final solution as

\[ \hat{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t. \quad (6) \]
The computation in $P_D[\hat{x}] = \arg\min_{x \in D} \|x - \hat{x}\|_2^2$ may be expensive if $c(x)$ is complex.

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

\[
\min_{x \in \mathbb{R}^{d \times d}} \|x - \hat{x}\|_2^2 \\
\text{s.t. } 0 \preceq x \preceq \epsilon I
\]  

has the complexity of order $O(d^3)$. 

The proposed Epro-SGD and its Proximal Variant
Proposed Epro-SGD Approach

The standard SGD solves

$$\min_x f(x)$$
$$\text{s.t. } c(x) \leq 0$$

(8)

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

$$f(x) + \lambda [c(x)]_+.\quad (9)$$

• $[s]_+$ is a hinge operator defined as $[s]_+ = s$ if $s \geq 0$, and $[s]_+ = 0$ otherwise.

• $\lambda$ 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 $\tilde{x}_T = P_D[\hat{x}_T]$

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

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. \hspace{2em} **for** \( t = 1, \ldots, T_k \)
4. \hspace{4em} Compute a stochastic gradient \( g(x_t^k) \)
5. \hspace{4em} Compute \( x_{t+1}^k = x_t^k - \eta_k (g(x_t^k) + \lambda \partial [c(x_t^k)]_+) \)
6. \hspace{2em} **endfor**
7. \hspace{2em} Compute \( \bar{x}_T^k = P_D[\bar{x}_T^k] \), where \( \bar{x}_T^k = \sum_{t=1}^{T_k} x_t^k / T_k \)
8. \hspace{2em} Update \( x_1^{k+1} = \bar{x}_T^k \), \( T_{k+1} = 2T_k \), \( \eta_{k+1} = \eta_k / 2 \)
9. \hspace{2em} 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
$$\min_{c(x)=0, v \in \partial c(x), v \neq 0} \|v\|_2 \geq \rho.$$  

Remarks on A3

• For any $\hat{x}$, let $\tilde{x} = \arg \min_{c(x) \leq 0} \|x - \hat{x}\|_2$.
$$\|\hat{x} - \tilde{x}\|_2 \leq \frac{1}{\rho} [c(\hat{x})]_+, \quad \rho > 0. \quad (10)$$

• 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∼A3, we derive

- Expected convergence bounds
- High-probability convergence bounds

all with optimal rates for strongly convex optimization.
Under Assumptions A1~A3 and given that \( f(x) \) is \( \beta \)-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

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

the solution \( x_{1^\dagger + 1} \) enjoys a convergence rate of

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

and \( c(x_{1^\dagger + 1}) \leq 0 \).
Under Assumptions A1∼A3 and given $\|x_t - x_*\|_2 \leq D$ for all $t$. If we let

$$
\mu = \rho/(\rho - G_1/\lambda), \quad G^2 = G^2_1 + \lambda^2 G^2_2,
$$

$$
C = (8G^2_1/\beta + 2G_1 D) \ln(m/\epsilon) + 2G_1 D,
$$

and set

$$
T_1 \geq \max \left(3C\beta/ (\mu G^2), 9\right), \quad \eta_1 = \mu/(3\beta),
$$

the total number of epochs $k^\dagger$ is given by

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

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

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

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

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

- 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.
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

\[ h(x) + \lambda [c(x)]_+ + k(x). \]  \hspace{1cm} (13)
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 $\tilde{x}_T = P_D[\hat{x}_T]$

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

Proximal Variant of Epro-SGD
Expected Convergence Bound

Under Assumptions A1∼A3 and given that \( \hat{f}(x) \) is \( \beta \)-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

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

and \( c(x_1^{k^\dagger + 1}) \leq 0. \)
Comparisons and Experiments
## Comparison with Competing Algorithms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Convergence Rate</th>
<th>Project Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard SGD (SGD)</td>
<td>$O(\log T / T)$</td>
<td>$O(T)$</td>
</tr>
<tr>
<td>One-Projection SGD (OneProj)</td>
<td>$O(\log T / T)$</td>
<td>1</td>
</tr>
<tr>
<td>logT-projection SGD (logT)</td>
<td>$O(1 / T)$</td>
<td>$O(\kappa \log T)$</td>
</tr>
<tr>
<td>Epro-SGD</td>
<td>$O(1 / T)$</td>
<td>$O(\log T)$</td>
</tr>
</tbody>
</table>

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

- Solve L1-norm constrained least squares optimization problem

\[
\min_w \frac{1}{2N} \sum_{i=1}^{N} (x_i^T w - y_i)^2 + \alpha \|w\|^2 \\
\text{s.t.} \quad \|w\|_1 \leq \beta.
\]

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

\[
\min_A \frac{c}{N} \sum_{j=1}^{N} \ell(A, x^j_1, x^j_2, x^j_3) + (1 - c) tr(AL) \\
+ \frac{\mu_1}{2} \|A\|_F^2 + \mu_2 \|A\|_1^{\text{off}} \\
s.t. \quad A \succeq \epsilon I, \tag{14}
\]

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