Randomized Coordinate Descent and Acceleration

1 Introduction

In this lecture, we consider optimization problem with the model parameter $w \in \mathbb{R}^d$. Here w can be decomposed into p components $w = [w_1, \ldots, w_p]$, where each w_j is a d_j dimensional vector, with $\sum_{j=1}^p d_j = d$.

We consider the following form of optimization problem:

$$\phi(w) = f(w) + g(w), \tag{1}$$

where

$$f(w) = \psi\left(\sum_{j=1}^p A_j w_j\right), \quad g(w) = \sum_{j=1}^p g(w_j).$$

We assume that $f(\cdot)$ is L_i-smooth with respect tor w_j, and $g(\cdot)$ is convex but may not be smooth.

Note that if $\psi(\cdot)$ is L smooth and $\|A_i\|_2$ is the spectral norm of A_i, then $L_i \leq \|A_i\|_2^2 \cdot L$.

Example 1 *Consider the Lasso problem with $w \in \mathbb{R}^d$ and $p = d$:*

$$\frac{1}{2n}\left\|\sum_{j=1}^d [x]_j w_j - y\right\|_2^2 + \sum_{j=1}^d \mu |w_j|,$$

where $y \in \mathbb{R}^n$ is the target vector for the training data. The feature vector $[x]_j \in \mathbb{R}^n$ is the j-th column of the data matrix, and w_j is the coefficient for the j-th feature.

For this problem, we have $\psi(u) = 0.5n^{-1}\|u - y\|_2^2$, which is n^{-1}-smooth, and $A_j = [x]_j$, and $g(w_j) = \mu|w_j|$. The smoothness parameter $L_i \leq \|[x]_j\|_2^2/n$.

Example 2 *Consider the dual formulation of the regularized loss minimization problem:*

$$\phi_D(\alpha) = \frac{1}{n}\sum_{i=1}^n -f_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^n X_i \alpha_i\right),$$

where each $\alpha_i \in \mathbb{R}^k$. Here $-\phi_D(\alpha)$ can be written as

$$\tilde{\psi}\left(\sum_{j=1}^p A_j \tilde{w}_j\right) + \sum_{j=1}^p \tilde{g}_j(\tilde{w}_j).$$

Here $\tilde{w}_j = \alpha_j$, $p = n$, $d = nk$, $\tilde{\psi}(u) = \lambda g^*(u)$, $A_j = (\lambda n)^{-1} X_j$, $\tilde{g}_j(\tilde{w}_j) = n^{-1} f_j^*(-\alpha_j)$ for $j = 1, \ldots, p$.

Example 3 *If A_j is a $d \times d_j$ matrix with identity matrix in the j-th block of size $d_j \times d_j$, then $\sum_{j=1}^p A_j w_j = [w_1, \ldots, w_p] \in \mathbb{R}^d$. We have a general situation where we have general $f(w) = \psi([w_1, \ldots, w_p])$, and the variable w is decomposed into p-components $w_1, \ldots, w_p$.*

2 Randomized Coordinate Descent

In randomized coordinate descent algorithm for solving (1), we randomly select a variable i from 1 to p, and minimize the objective with respect to w_i using proximal gradient. That is, we select i, and optimize with respect $w_i + \Delta w_i$:

$$\psi\left(\sum_{j=1}^{p} A_j w_j + A_i \Delta w_i\right) + \sum_{j=1}^{p} g_j\left(w_j + \Delta w_i \delta_i^j\right).$$

Given $\eta_i \leq 1/L_i$, we use an upper bound of $f(\cdot)$ as follows:

$$\psi\left(\sum_{j=1}^{p} A_j w_j\right) + \nabla\psi\left(\sum_{j=1}^{p} A_j w_j\right)^\top (A_i \Delta w_i) + \frac{1}{2\eta_i}\|\Delta w_i\|_2^2 + g_i\left(w_i + \Delta w_i\right).$$

Let $\|A_i\|_2$ be the spectral norm of A_i. Let

$$u = \sum_{j=1}^{p} A_j w_j,$$

then we can optimize

$$\begin{aligned}
\Delta w_i =& \arg\min_{\Delta w}\left[(A_i^\top \nabla f(u))^\top \Delta w + \frac{1}{2\eta}\|A_i\|_2^2 \|\Delta w\|_2^2 + g_i(w_i + \Delta w)\right] \\
=& \arg\min_{\Delta w}\left[\frac{1}{2\eta_i}\|\Delta w + \eta_i A_i^\top \nabla f(u)\|_2^2 + g_i(w_i + \Delta w)\right] \\
=& \mathrm{prox}_{\eta_i \cdot g_i}(w_i - \eta_i A_i^\top \nabla f(u)) - w_i,
\end{aligned}$$

and

$$\mathrm{prox}_{\eta_i g_i}(w) = \arg\min_{z \in \mathbb{R}^{d_i}}\left[\frac{1}{2}\|z - w\|_2^2 + \eta_i g_i(z)\right].$$

This leads to Algorithm 1, which is the primal counterpart of the proximal SDCA.

Algorithm 1: Randomized Proximal Coordinate Descent

Input: $\phi(\cdot)$, $\eta_i \leq 1/L_i (i = 1, \ldots, p)$, $w^{(0)}$
Output: $w^{(T)}$

Let $u^{(0)} = \sum_{j=1}^{p} A_j w_j^{(0)}$
for $t = 1, 2, \ldots, T$ **do**
- Randomly pick $i \sim [1, \ldots, p]$
- Let $w_i^{(t)} = \mathrm{prox}_{\eta_i g_i}(w_i^{(t)} - \eta_i A_i^\top \nabla f(u^{(t-1)}))$
- Let $w_j^{(t)} = w_j^{(t-1)}$ for $j \neq i$
- Let $u^{(t)} = u^{(t-1)} + A_i(w_i^{(t)} - w_i^{(t-1)})$

Return: $w^{(T)}$

Theorem 1 *In Algorithm 1, assume that $\eta \le 1/L$, then $\forall w = [w_1, \ldots, w_p] \in \mathbb{R}^d$:*

$$\frac{p-1}{T}\mathbf{E}\phi(w^{(T)}) + \frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\phi(w^{(t)}) \le \frac{p-1}{T}\phi(w^{(0)}) + \phi(w) + \frac{1}{T}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i^{(0)} - w_i\|_2^2.$$

Proof Let $\sum_j A_j w_j^{(t)} = \sum_j A_j w_j^{(t-1)} + A_i(w_i^{(t)} - w_i^{(t-1)})$. We have for all $w \in \mathbb{R}^d$:

$$\begin{aligned}
\phi(w^{(t)}) =& \left[\psi\left(u^{(t-1)} + A_i(w_i^{(t)} - w_i^{(t-1)})\right) + g(w^{(t)})\right] \\
\le& \psi\left(u^{(t-1)}\right) + \left(A_i^\top \nabla\psi\left(u^{(t-1)}\right)\right)^\top (w_i^{(t)} - w_i^{(t-1)}) + \frac{1}{2\eta_i}\|w_i^{(t)} - w_i^{(t-1)}\|_2^2 + g(w^{(t)}) \qquad (2) \\
\le& \psi\left(u^{(t-1)}\right) + \left(A_i^\top \nabla\psi\left(u^{(t-1)}\right)\right)^\top (w_i - w_i^{(t-1)}) + \frac{1}{2\eta_i}\|w_i - w_i^{(t-1)}\|_2^2 + g(w_i) \\
& + \sum_{j\ne i} g(w_j^{(t-1)}) - \frac{1}{2\eta_i}\|w_i - w_i^{(t)}\|_2^2.
\end{aligned}$$

The first inequality uses the fact that $f(w^{(t)})$ is η_i^{-1}-smoothness with respect to $w_i^{(t)}$. The second inequality uses the fact that $w_i^{(t)}$ is the minimizer of (2).

Take expectation with respect to i, we obtain

$$\begin{aligned}
\mathbf{E}_i\phi(w^{(t)}) \le& \psi\left(u^{(t-1)}\right) + \frac{1}{p}\nabla\psi\left(u^{(t-1)}\right)^\top\left(\sum_{i=1}^{p} A_i w_i - u^{(t-1)}\right) + \frac{1}{p}g(w) + \frac{p-1}{p}g(w^{(t-1)}) \\
& + \frac{1}{p}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i - w_i^{(t-1)}\|_2^2 - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i - w_i^{(t)}\|_2^2 \\
\le& \frac{p-1}{p}\phi(w^{(t-1)}) + \frac{1}{p}\phi(w) + \frac{1}{p}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i - w_i^{(t-1)}\|_2^2 - \frac{1}{p}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i - w_i^{(t)}\|_2^2,
\end{aligned}$$

where the second inequality uses

$$\frac{1}{p}\psi\left(u^{(t-1)}\right) + \frac{1}{p}\nabla\psi\left(u^{(t-1)}\right)^\top\left(\sum_{i=1}^{p} A_i w_i - u^{(t-1)}\right) \le \frac{1}{p}\psi\left(\sum_{i=1}^{p} A_i w_i\right) = \frac{1}{p}f(w).$$

By summing over $t = 1$ to $t = T$, we obtain

$$\mathbf{E}\phi(w^{(T)}) + \frac{1}{p}\sum_{t=1}^{T-1}\mathbf{E}\phi(w^{(t)}) \le \frac{p-1}{p}\phi(w^{(0)}) + \frac{T}{p}\phi(w) + \frac{1}{p}\sum_{i=1}^{p}\frac{1}{2\eta_i}\|w_i^{(0)} - w_i\|_2^2.$$

This proves the theorem. ∎

3 Acceleration

It is possible to derive accelerated coordinate descent methods. We present an accelerated method for SDCA in Algorithm 2, which applies to the dual formulation for regularized loss minimization problem:

$$\frac{1}{n}\sum_{i=1}^{n} -f^*(-\alpha_i) - g^*\left(\frac{1}{n}\sum_{i=1}^{n} X_i\alpha_i\right)$$

strongly convex problems, as described in [1]. In this method, the proximal mapping is defined as:

$$\mathrm{prox}_{\tau g}(w) = \arg\min_z \left[\frac{1}{2\tau}\|z-w\|_2^2 + g(z)\right].$$

Algorithm 2: Stochastic Primal-Dual Coordinate Method (SPDC)

Input: $\phi(\cdot)$, L, λ, $\alpha^{(0)}$, and R such that $\|X_i\|_2 \leq R$
Output: $\alpha^{(T)}$, $w^{(T)}$
Let $\tau = 1/(2R\sqrt{n\lambda L})$
Let $\sigma = \sqrt{n\lambda L}/(2R)$
Let $\theta = 1 - 1/(n + R\sqrt{nL/\lambda})$
Let $u^{(0)} = n^{-1}\sum_{i=1}^{n} X_i\alpha_i$
Let $w^{(0)} = \nabla g^*(u^{(0)})$
let $\bar{w}^{(0)} = w^{(0)}$
for $t = 1, 2, \ldots, T$ **do**
    Randomly pick i
    Let $\Delta\alpha_i \in \arg\max_{\Delta\alpha_i}\left[-f_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \bar{w}^{(t-1)\top} X_i\Delta\alpha_i - \frac{1}{2\sigma}\|\Delta\alpha_i\|_2^2\right]$
    Let $\alpha_i^{(t)} = \alpha_i^{(t-1)} + \Delta\alpha_i$ and $\alpha_j^{(t)} = \alpha_j^{(t-1)}$ when $j \neq i$
    Let $w^{(t)} = \mathrm{prox}_{\tau g}(w^{(t-1)} + \tau(u^{(t-1)} + X_i\Delta\alpha_i))$
    Let $u^{(t)} = u^{(t-1)} + n^{-1}X_i\Delta\alpha_i$
    Let $\bar{w}^{(t)} = w^{(t)} + \theta(w^{(t)} - w^{(t-1)})$
Return: $\alpha^{(T)}$, $w^{(T)}$

Theorem 2 *([1]) Assume that $f_i^*(\cdot)$ is $1/L$-strongly convex, and $g(\cdot)$ is λ-strongly convex. Let $R = \max_i \|X_i\|_2$. We have*

$$\begin{aligned}&\left(\frac{1}{2\tau}+\lambda\right)\mathbf{E}\|w^{(t)} - w_*\|_2^2 + \left(\frac{1}{4\sigma}+\frac{1}{L}\right)\mathbf{E}\|\alpha^{(t)} - \alpha_*\|_2^2\\ \leq&\theta^t\left(\left(\frac{1}{2\tau}+\lambda\right)\mathbf{E}\|w^{(0)} - w_*\|_2^2 + \left(\frac{1}{4\sigma}+\frac{1}{L}\right)\mathbf{E}\|\alpha^{(0)} - \alpha_*\|_2^2\right).\end{aligned}$$

Assume $R = O(1)$, and $\kappa = L/\lambda$, then SDCA requires

$$O\left((n+\kappa)\log\frac{1}{\epsilon}\right)$$

steps to achieve primal suboptimality of $O(\epsilon)$. On the other hand, SPDC requires only

$$O\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon}\right)$$

steps. Therefore SPDC converges faster if $\kappa \gg n$. In fact, the rate achieved by SPDC is optimal for finite sum problems.

4 Empirical Studies

We study the smoothed hinge loss function $\phi_\gamma(z)$ with $\gamma = 1$, and solves the following $L_1 - L_2$ regularization problem:

$$\min_w \left[\underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi_\gamma(w^\top x_i y_i)}_{f(w)} + \underbrace{\frac{\lambda}{2}\|w\|_2^2 + \mu\|w\|_1}_{g(w)} \right].$$

We compare proximal gradient, accelerated proximal gradient, SDCA, to primal coordinate descent, and dual accelerated gradient descent (SPDC). The results show that SDCA is superior when n is large, and especially when λn is at least $O(1)$ order. It is not as competitive as traditional algorithms when λn is much smaller than 1. This is consistent with the theory. When λn is smaller than 1, SDCA will be better.

Primal CD is not sensitive to λ, and still works well when $\lambda n \ll 1$.

The computational complex of each proximal gradient descent or accelerated proximal gradient descent is $O(nd)$. The expected computational complex of each iteration of CD is $O(np^{-1}\sum_{i=1}^{p} d_i) = O(nd/p)$. Therefore after every p iterations, we have a computationally complexity of $O(nd)$. This is also the same computational cost of SDCA or SPDC after n iterations. In the plots, each epoch is p inner iterations for CD, or n inner iterations for SDCA and SPDC.

References

[1] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *J. Mach. Learn. Res.*, 18(1):29392980, 2017.

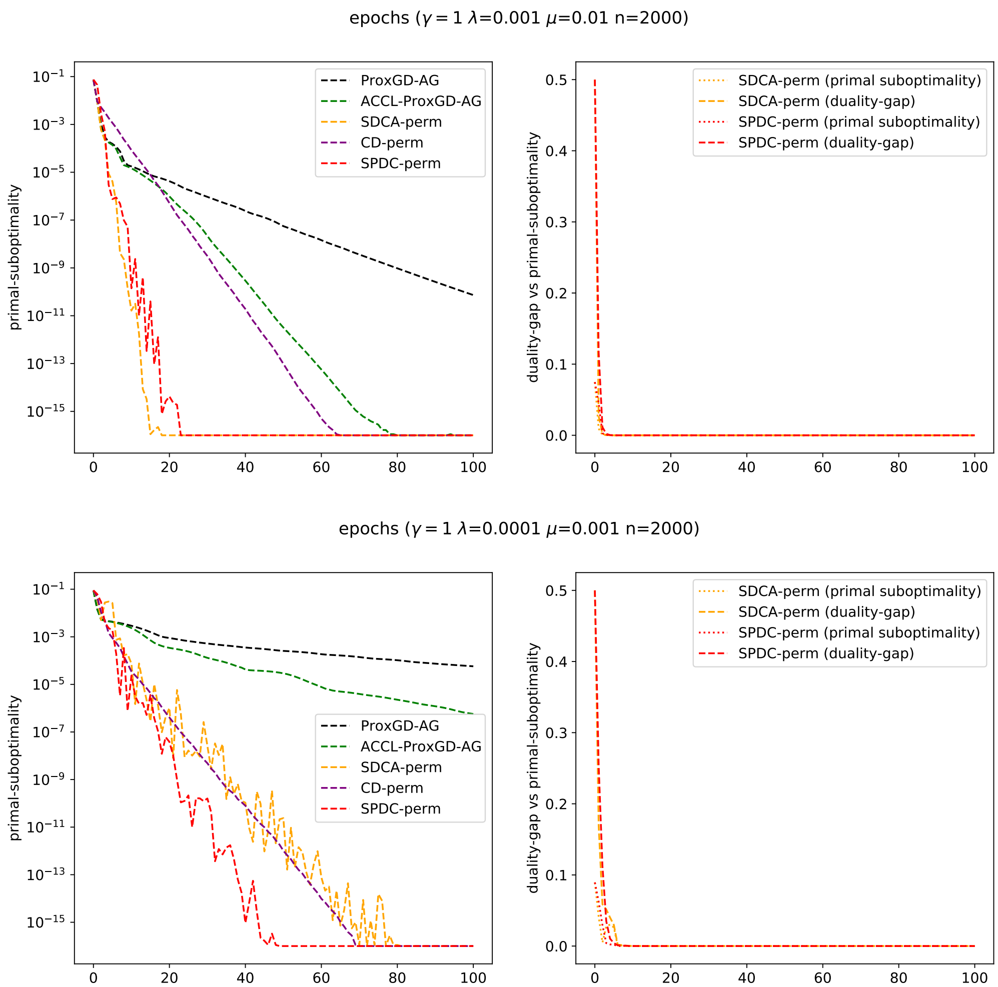

Figure 1: Comparisons of Proximal Gradient, SDCA and primal CD, SPDC

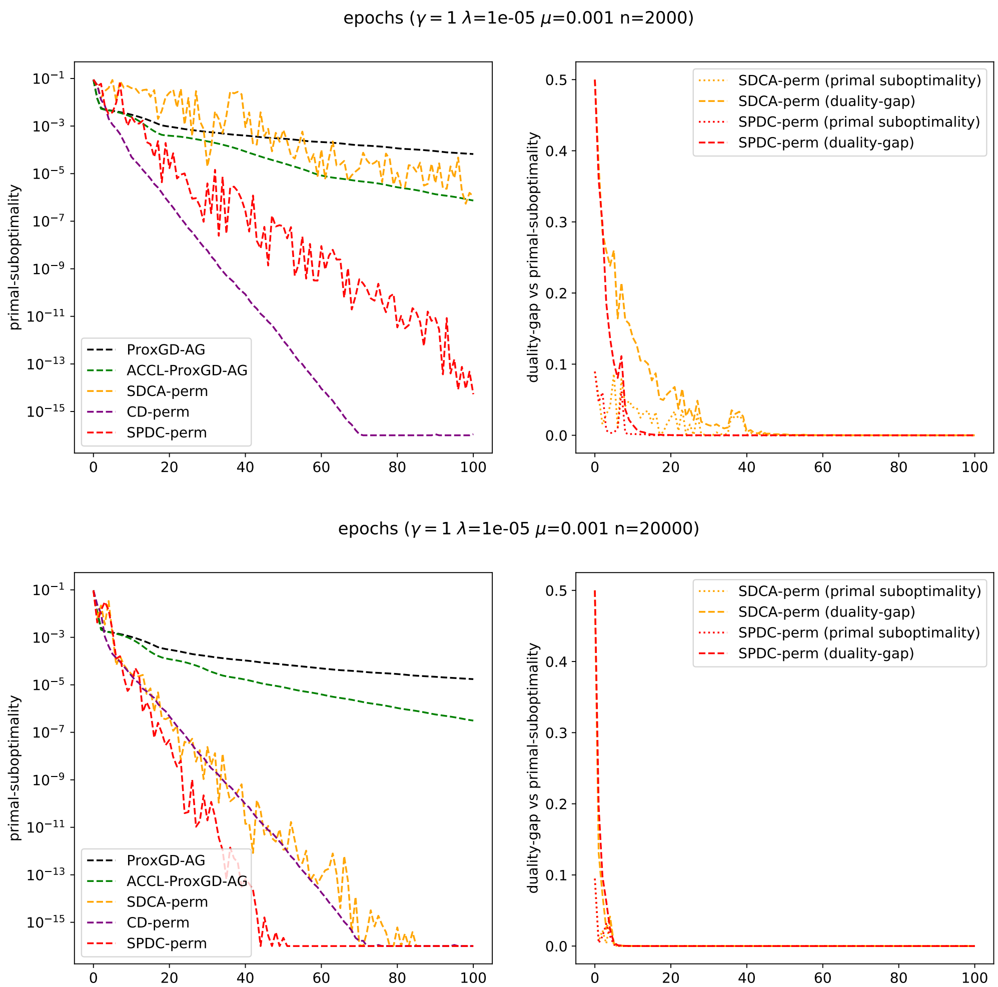

Figure 2: Comparisons of Proximal Gradient, SDCA and primal CD, SPDC