Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 17: Proximal Stochastic Dual Coordinate Ascent

Regularized Loss Minimization

In this lecture, we still consider the composite optimization problem, but with an added finite sum structure as follows,

$$\phi(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(X_i^\top w) + \lambda g(w), \tag{1}$$

where $w \in \mathbb{R}^d$ is the model parameter:
We assume that $g(w)$ is strongly convex.

Dual Decomposition

In order to derive the dual formulation of (1), we use the decomposition technique, and rewrite it as:

$$\phi(w, \{u_i\}) = \frac{1}{n}\sum_{i=1}^{n} f_i(u_i) + \lambda g(w), \quad \text{subject to } \forall i, X_i^\top w = u_i.$$

Here we have n dual variables $\{\alpha_i\}_{1,\ldots,n}$, and each $\alpha_i \in \mathbb{R}^k$. One dual variable for each constraint.

Dual Formulation

The dual objective function is defined as:

$$\phi_D(\alpha) = \min_{w,\{u_i\}} L(w, \{u_i\}, \alpha) = \frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n} X_i \alpha_i\right). \tag{2}$$

In this case, we may define the primal solution w from the dual variables as follows:

$$w = \nabla g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n} X_i \alpha_i\right). \tag{3}$$

Example

In ridge regression, we have $k = 1$ and loss is:

$$f_i(u) = \frac{1}{2}(u - y_i)^2,$$

and regularizer is

$$g(w) = \frac{1}{2}\|w\|_2^2.$$

The primal problem is:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2.$$

The dual problem is:

$$\frac{1}{n}\sum_{i=1}^{n} - \left[-\alpha_i^\top y_i + \frac{1}{2}\alpha_i^2\right] - \frac{1}{2\lambda n^2}\left\|\sum_{i=1}^{n} x_i \alpha_i\right\|_2^2,$$

where each $\alpha_i \in \mathbb{R}$.

Example

Consider regularized multi-class logistic regression, with training data $\{(x_i, y_i)\}$. The input features are $X_i = [\psi(x_i, j)]_{j=1,\ldots,k}$, where each $\psi(x, y) \in \mathbb{R}^d$ corresponds to the feature vector of data x for class y. The label $y_i \in \{1, \ldots, k\}$ is the class label. The loss functions are

$$f_i(u) = -u_{y_i} + \ln \sum_{y'=1}^{k} \exp(u_{y'}),$$

and the regularizer is

$$g(w) = \frac{1}{2}\|w\|_2^2.$$

The primal problem is

$$\frac{1}{n}\sum_{i=1}^{n}\left[-\psi(x_i, y_i)^\top w + \ln \sum_{y'=1}^{k} \exp(\psi(x_i, y')^\top w)\right] + \frac{\lambda}{2}\|w\|_2^2.$$

Prox-SDCA

The dual coordinate ascent (DCA) method maximizes the dual problem (2) by optimizing one α_i at a time for a chosen i, while keeping α_j with $j \neq i$ fixed.

We focus on a *stochastic* version of DCA, called SDCA, in which at each round we choose which dual variable α_i to optimize uniformly at random.

$$\frac{1}{n}\sum_{j=1}^{n} -f_j^*(-\alpha_j + \Delta\alpha_j \delta_i^j) - \lambda g^* \left(\frac{1}{\lambda n}\sum_{j=1}^{n} X_j \alpha_j + \frac{1}{\lambda n} X_i \Delta\alpha_i \right).$$

$\delta_i^j = 1$ when $i = j$ and $\delta_i^j = 0$ otherwise.

Motivation of SDCA

The idea of Prox-SDCA algorithm can be described as follows. Consider the maximal increase of the dual objective, where we only allow to change the *i*'th component of α. At step *t*, let

$$v^{(t-1)} = (\lambda n)^{-1} \sum_i X_i \alpha_i^{(t-1)}$$

and let

$$w^{(t-1)} = \nabla g^*(v^{(t-1)}).$$

We will update the *i*-th dual variable $\alpha_i^{(t)} = \alpha_i^{(t-1)} + \Delta\alpha_i$, in a way that will lead to a sufficient increase of the dual objective.

Motivation of Prox-SDCA

The goal of SDCA is to increase the dual objective as much as possible.
Instead of directly maximizing the dual objective function, which may be hard for complex $g(w)$, we try to maximize the following proximal objective which is a lower bound of the dual objective:

$$
\begin{aligned}
&\max_{\Delta\alpha_i \in \mathbb{R}^k} \left[-\frac{1}{n} f_i^*(-(\alpha_i + \Delta\alpha_i)) - \lambda \left(\nabla g^*(v^{(t-1)})^\top (\lambda n)^{-1} X_i \Delta\alpha_i \right.\right. \\
&\qquad \left.\left. + \ \frac{1}{2} \|(\lambda n)^{-1} X_i \Delta\alpha_i\|_2^2 \right) \right] \\
= \ &\max_{\Delta\alpha_i \in \mathbb{R}^k} \left[-f_i^*(-(\alpha_i + \Delta\alpha_i)) - w^{(t-1)\,\top} X_i \Delta\alpha_i - \frac{1}{2\lambda n} \|X_i \Delta\alpha_i\|_2^2 \right].
\end{aligned}
$$

Prox-SDCA

Algorithm 1: Proximal Stochastic Dual Coordinate Ascent

Input: $\phi(\cdot)$, L, λ, $\alpha^{(0)}$, and R such that $\|X_i\|_2 \le R$
Output: $\alpha^{(T)}$, $w^{(T)}$

Let $w^{(0)} = \nabla g^*(\alpha^{(0)})$
for $t = 1, 2, \ldots, T$ **do**
- Randomly pick i
- Find $\Delta\alpha_i$ such as the dual objective is no smaller than one of the following options
- **Option I:**
  - $\Delta\alpha_i \in \arg\max_{\Delta\alpha_i} \left[-f_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - w^{(t-1)\top} X_i \Delta\alpha_i - \frac{1}{2\lambda n} \|X_i \Delta\alpha_i\|_2^2 \right]$
- **Option II:**
  - Let u be s.t. $-u \in \partial f_i(X_i^\top w^{(t-1)})$
  - Let $z = u - \alpha_i^{(t-1)}$
  - Let $s = \arg\max_{s\in[0,1]} \left[-f_i^*(-(\alpha_i^{(t-1)} + sz)) - s\, w^{(t-1)\top} X_i z - \frac{s^2}{2\lambda n} \|X_i z\|_2^2 \right]$
  - Set $\Delta\alpha_i = sz$
- Let $\alpha_i^{(t)} \leftarrow \alpha_i^{(t-1)} + \Delta\alpha_i$ and $\alpha_j^{(t)} = \alpha_j^{(t-1)}$ when $j \ne i$
- Let $v^{(t)} \leftarrow v^{(t-1)} + (\lambda n)^{-1} X_i \Delta\alpha_i$
- Let $w^{(t)} \leftarrow \nabla g^*(v^{(t)})$

Return: $\alpha^{(T)}$, $w^{(T)}$

Example

For example, for ridge regression, we can take option I: and solve

$$\max_{\Delta\alpha_i} \left[\Delta\alpha_i y_i - \frac{1}{2}(\alpha_i^{(t-1)} + \Delta\alpha_i)^2 - w^{(t-1)\top} x_i \Delta\alpha_i - \frac{1}{2\lambda n} \|x_i \Delta\alpha_i\|_2^2 \right],$$

which leads to

$$\begin{aligned}
\Delta\alpha_i =& \frac{\lambda n}{\lambda n + \|x_i\|_2^2} \left[y_i - w^{(t-1)\top} x_i - \alpha_i^{(t-1)} \right] \\
w^{(t)} =& w^{(t-1)} + \frac{1}{\lambda n} x_i \Delta\alpha_i .
\end{aligned}$$

Example: $L_1 - L_2$ Regularized Logistic Regression

Primal: ($y_i \in \{\pm 1\}$)

$$\phi(w) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ln(1 + e^{-w^\top x_i y_i})}_{f_i(w)} + \underbrace{\frac{\lambda}{2} w^\top w + \mu \|w\|_1}_{\lambda g(w)} .$$

Dual: with $\alpha_i y_i \in [0, 1]$

$$\phi_D(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{-\alpha_i y_i \ln(\alpha_i y_i) - (1 - \alpha_i y_i)\ln(1 - \alpha_i y_i)}_{-f_i^*(-\alpha_i)} - \underbrace{\frac{\lambda}{2}\|\mathrm{trunc}(v, \mu/\lambda)\|_2^2}_{\lambda g^*(v)}$$

$$\text{s.t. } v = \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i x_i; \qquad w = \mathrm{trunc}(v, \mu/\lambda)$$

where

$$\mathrm{trunc}(u, \delta)_j = \begin{cases} u_j - \delta & \text{if } u_j > \delta \\ 0 & \text{if } |u_j| \le \delta \\ u_j + \delta & \text{if } u_j < -\delta \end{cases}$$

Convergence

We have the following convergence result for Prox-SDCA.

Theorem

In Algorithm 1, assume that for all i, f_i is L-smooth. To obtain an expected duality gap of $\mathbf{E}[\phi(w^{(T)}) - \phi_D(\alpha^{(T)})] \leq \epsilon_P$, it suffices to have a total number of iterations of

$$T \geq \left(n + \frac{R^2 L}{\lambda}\right) \log((n + \frac{R^2 L}{\lambda}) \cdot \frac{\phi(w^{(0)}) - \phi_D(\alpha^{(0)})}{\epsilon_P}).$$

Number of data processed:

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

For GD:

$$T = O\left((n \cdot \kappa) \log \frac{1}{\epsilon}\right)$$

Proof Sketch

The key lemma is the following:

Lemma

Assume that ϕ_i^* is γ-strongly-convex (where γ can be zero). Then, for any iteration t and any $s \in [0, 1]$ we have

$$\mathbf{E}[\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)})] \geq \frac{s}{n}\mathbf{E}\left[\phi(w^{(t-1)}) - \phi_D(\alpha^{(t-1)})\right] - \left(\frac{s}{n}\right)^2 \frac{G^{(t)}}{2\lambda},$$

where

$$G^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left(\|X_i\|_2^2 - \frac{\gamma(1-s)\lambda n}{s}\right)\mathbf{E}\left[\|u_i^{(t-1)} - \alpha_i^{(t-1)}\|_2^2\right],$$

where $\|X_i\|_2$ denotes the spectral norm of X_i, and $-u_i^{(t-1)} \in \partial f_i(X_i^\top w^{(t-1)})$.

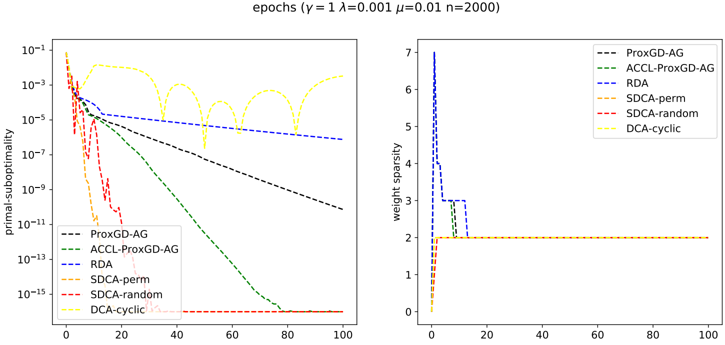
Empirical Studies

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

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

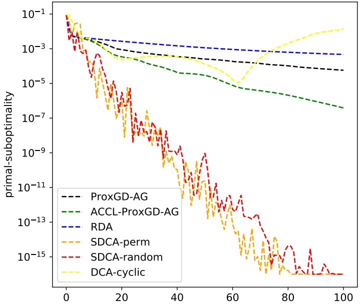
We compare different algorithms

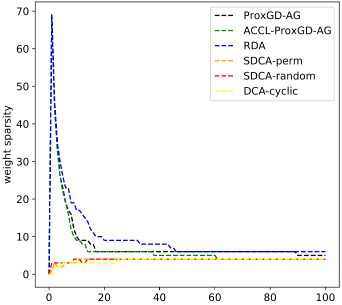
Comparisons



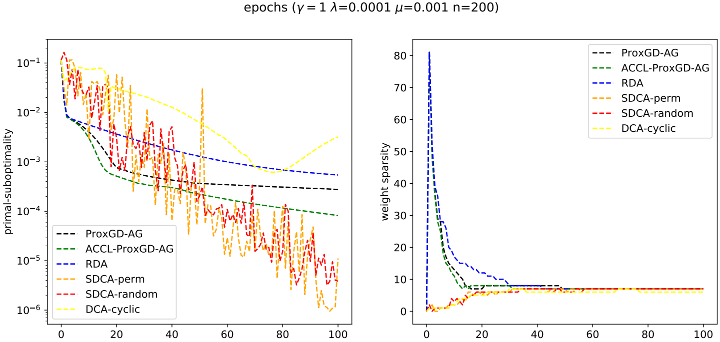
Comparisons

epochs ($\gamma = 1$ λ=0.0001 μ=0.001 n=2000)





Comparisons



Summary

Regularized Loss Minimization

- Finite Sum Structure

Dual Formulation

- n constraints
- n dual variables

Prox-SDCA

- Update one dual variable at a time (one data point)
- convergence: $\tilde{O}(n+\kappa)$.
- Prox-GD: $\tilde{O}(n\kappa)$
- Prox-AGD: $\tilde{O}(n\sqrt{\kappa})$
- Optimal: $\tilde{O}(n+\sqrt{n\kappa})$.