SDCA

Stochastic Dual Coordinate Ascent

Jingchang Liu

June 29, 2017

University of Science and Technology of China

Table of Contents

Lagrangian Duality

Dual Problem

Primal Problem

$$\begin{aligned} min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, i = 1, 2 \cdots, m \\ & h_i(x) = 0, i = 1, 2, \cdots, p \end{aligned}$$

Lagrangian Function

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x), \lambda_i \geq 0$$

Dual Fucntion

$$g(\lambda, v) = \inf_{x \in D} L(x, \lambda, v)$$

$g(\lambda, v)$ is a concave function.

SDCA

Reference

Stochastic Dual Coordinate AscentMethods for Regularized Loss Minimization, Shai Shalev-Shwartz & Tong Zhang, JMLR2013

Optimization Objective

Formulation

$$\min_{w \in \mathbb{R}^d} P(w)$$

$$P(w) := \frac{1}{n} \sum_{i=1}^{n} \phi_i \left(w^T x_i \right) + \frac{\lambda}{2} \|w\|^2$$

Parameters

- $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$, $\phi_1, \phi_2, \cdots, \phi_n$: Scalar convex functions.
- SGD: $O(1/n)$

Examples

- SVM: $\phi_i \left(w^T x_i \right) = \max \left\{ 0, 1 - y_i w^T x_i \right\}$
- Logistic Regression: $\phi_i \left(w^T x_i \right) = \log \left(1 + \exp \left(-y_i w^T x_i \right) \right)$
- Ridge Regression: $\phi_i \left(w^T x_i \right) = \left(w^T x_i - y_i \right)^2$

Dual Problem

$$\max_{\alpha} D\left(\alpha\right)$$

$$D\left(\alpha\right) = \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*\left(-\alpha_i\right) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2$$

Conjugate function: $\phi_i^*\left(u\right) = \max_z\left(zu - \phi_i\left(z\right)\right)$

Derivation

$P\left(w\right) = \frac{1}{n}\sum\limits_{i=1}^{n}\phi_i\left(w^T x_i\right) + \frac{\lambda}{2}\left\|w\right\|^2$ equals to

$$\begin{aligned} P\left(y, z\right) \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}\phi_i\left(z_i\right) + \frac{\lambda}{2}\left\|y\right\|^2 \\ s.t. \qquad & \qquad y^T x_i = z_i, i = 1, 2, \cdots, n \end{aligned}$$

Derivation

$$L(y,z,\alpha) = P(y,z) + \frac{1}{n}\sum_{i=1}^{n}\alpha_i\left(y^T x_i - z_i\right)$$

$$\begin{aligned}
D(\alpha) &= \inf_{y,z} L(y,z,\alpha) \\
&= \frac{1}{n}\sum_{i=1}^{n}\inf_{z_i}\left\{\phi_i(z_i) - \alpha_i z_i\right\} + \inf_{y}\left\{\frac{\lambda}{2}\|y\|^2 + \frac{1}{n}\sum_{i=1}^{n}\alpha_i y^T x_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2
\end{aligned}$$

Relationship

$$w(\alpha) = \frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i$$

Assumptions

L-**Lipschitz continuous**

$$|\phi_i(a) - \phi_i(b)| \leq L\,|a - b|$$

$1/\gamma$-**smooth**

A function $\phi_i : \mathbb{R} \to \mathbb{R}$ is $(1/\gamma)$-smooth if it is differentiable and its derivative is $(1/\gamma)$-Lipschitz.

Remark

if $\phi_i(a)$ is $(1/\gamma)$-smooth, then ϕ_i^* is γ strongly convex.

Algorithms

Let $w^{(0)} = w(\alpha^{(0)})$

Iterate: for $t = 1, 2, \dots, T$:

Randomly pick i

Find $\Delta\alpha_i$ to maximize $-\phi_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i\|^2$

$\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$

$w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i$

Output (Averaging option):

Let $\bar{\alpha} = \frac{1}{T-T_0}\sum_{i=T_0+1}^{T}\alpha^{(t-1)}$

Let $\bar{w} = w(\bar{\alpha}) = \frac{1}{T-T_0}\sum_{i=T_0+1}^{T} w^{(t-1)}$

return $\bar{w}$

Figure 1: Procedure SDCA

Theorem

Th1

Consider Procedure SDCA with $\alpha^{(0)} = 0$. Assume that ϕ_i is L-Lipschitz for all i. To abtain a duality gap of $\mathbb{E}\left[P\left(\bar{w}\right) - D\left(\bar{\alpha}\right)\right] \leq \varepsilon$, it suffices to have a total number of iterations of

$$T \geq T_0 + n\frac{4L^2}{\lambda\varepsilon}$$

Th2

Consider Procedure SDCA with $\alpha^{(0)} = 0$. Assume that ϕ_i is $(1/\gamma)$-smooth for all i. To abtain a duality gap of $\mathbb{E}\left[P\left(\bar{w}\right) - D\left(\bar{\alpha}\right)\right] \leq \varepsilon$, it suffices to have a total number of iterations of

$$T \geq \left(n + \frac{1}{\lambda\gamma}\right)\log\left(\left(n + \frac{1}{\lambda\gamma}\right) \cdot \frac{1}{\varepsilon}\right)$$

Linear Convergence For Smooth Hinge-Loss

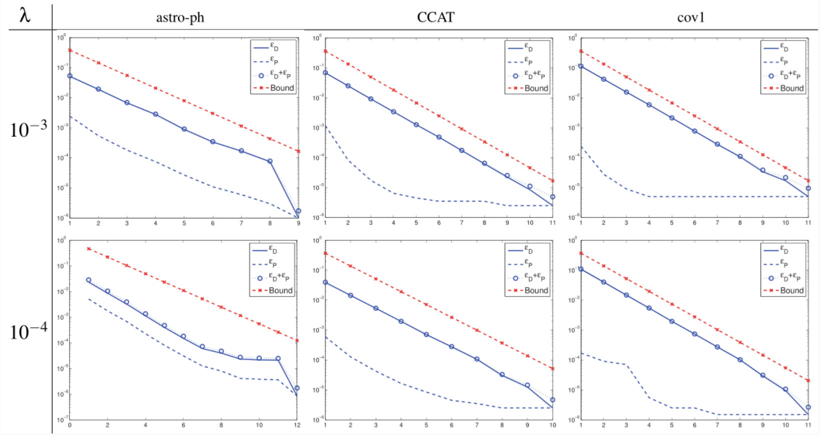


Figure 2: Experiments with the smoothed hinge-loss ($\gamma = 1$).

Convergence For Non-smooth Hinge-loss

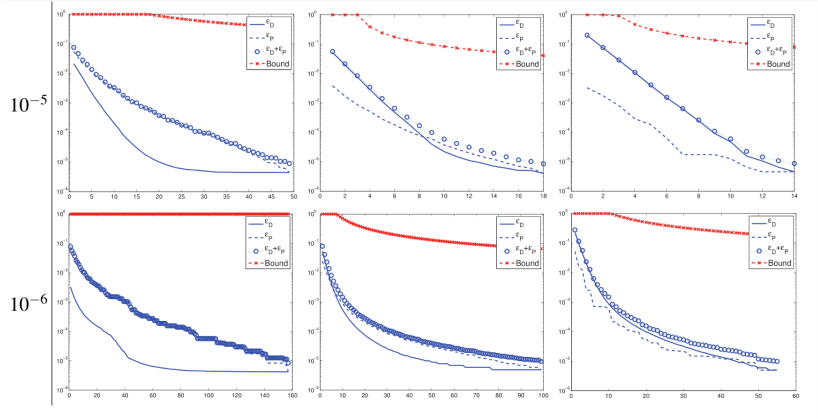


Figure 3: Experiments with the hinge-loss (non-smooth)

Effect of Smoothness Parameter

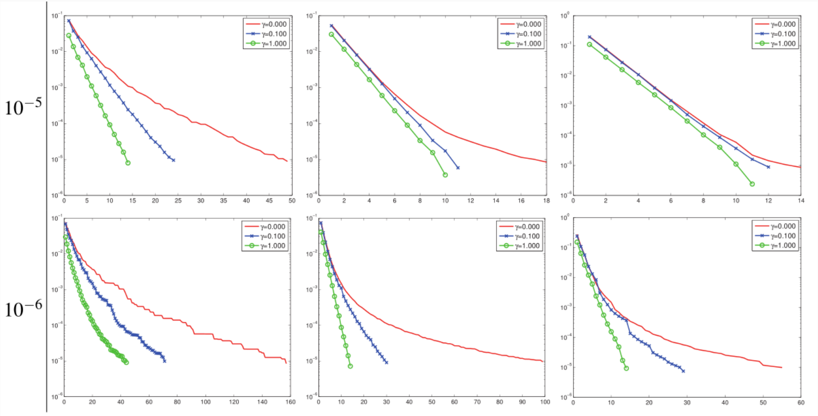


Figure 4: Duality gap as a function of the number of rounds for different values of γ

Comparison To SGD

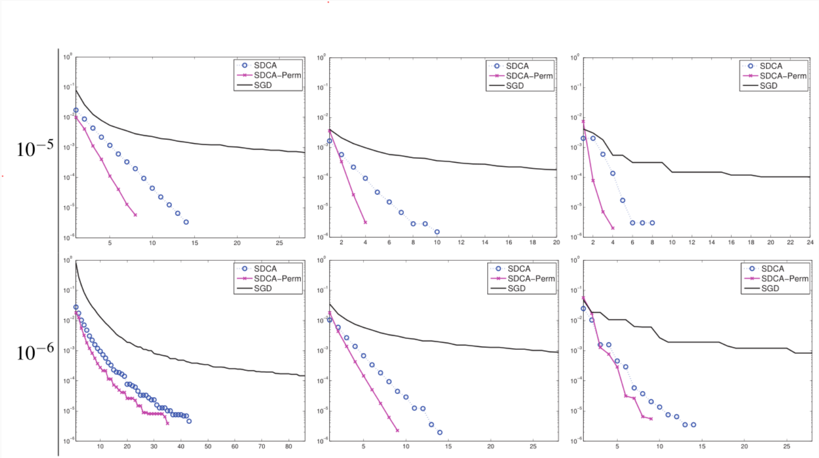


Figure 5: Comparing the primal sub-optimality of SDCA and SGD for the smoothed hinge-loss ($\gamma = 1$)

Asynchronous SDCA

Introduction

Reference

PASSCoDe: Parallel ASynchronous Stochastic dual Co-ordinate Descent

Prime Problem

$$\min_{w \in \mathbb{R}^d} P(w) := \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} l_i \left(w^T x_i \right)$$

Dual Problem

$$\min_{\alpha \in \mathbb{R}^d} D(\alpha) := \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i x_i \right\|^2 + \sum_{i=1}^{n} l_i^* (-\alpha_i)$$

Algorithm

Algorithm 2 Parallel Asynchronous Stochastic dual Co-ordinate Descent (*PASSCoDe*)

Input: Initial $\boldsymbol{\alpha}$ and $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i$

Each thread repeatedly performs the following updates:

step 1: Randomly pick i

step 2: Update $\alpha_i \leftarrow \alpha_i + \Delta\alpha_i$, where

$$\Delta\alpha_i \leftarrow \arg\min_{\delta} \frac{1}{2}\|\boldsymbol{w} + \delta \boldsymbol{x}_i\|^2 + \ell_i^*(-(\alpha_i + \delta))$$

step 3: Update $\boldsymbol{w}$ by $\boldsymbol{w} \leftarrow \boldsymbol{w} + \Delta\alpha_i \boldsymbol{x}_i$

Figure 6: Parallel Asynchronous Stochastic dual Co-ordinate Descent (PASSCoDe)

Operation

PASSCoDe-Lock

- Step 1.5: lock variables in $N_i := \{w_t | (x_i)_t \neq 0\}$
- The locks are then released after step 3.
- May equal to inconsistent read.

PASSCode-Atomic

- step 3: For each $j \in N(i)$, Update $w_j \leftarrow w_j + \triangle\alpha_i (x_i)_j$ atomically.

Linear Convergence Rate of PASSCoDe-Atomic

Theorem

If

$$\left(6\tau\left(\tau+1\right)^{2}eM\right)/\sqrt{n}\leq 1$$

and

$$1\geq\frac{2L_{\max}}{R_{\min}^{2}}\left(1+\frac{e\tau M}{\sqrt{n}}\right)\frac{\tau^{2}M^{2}e^{2}}{n}$$

then PASSCoDe-Atomic has a global linear convergence rate in expectation, that is,

$$E\left[D\left(\alpha^{j+1}\right)\right]-D\left(\alpha^{*}\right)\leq\eta\left(E\left[D\left(\alpha^{j}\right)\right]-D\left(\alpha^{*}\right)\right)$$

where α^* is the optimal solution and

$$\eta=1-\frac{\kappa}{L_{\max}}\left(1-\frac{2L_{\max}}{R_{\min}^{2}}\left(1+\frac{e\tau M}{\sqrt{n}}\right)\frac{\tau^{2}M^{2}e^{2}}{n}\right)$$

Convergence and Efficiency

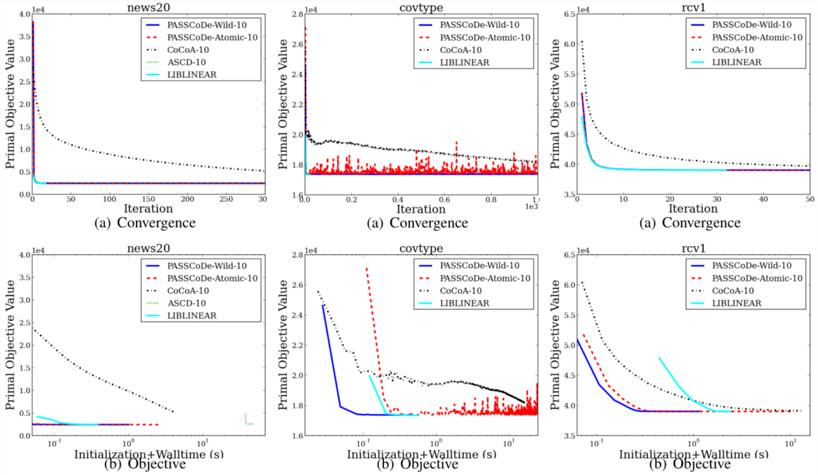
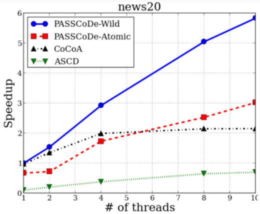


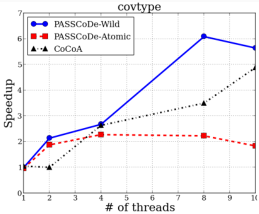
Figure 7: Convergence and Efficiency for news20, covtype, rcv1 datasets

Speedup



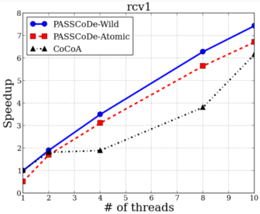
(d) Speedup

Figure 2: news20 dataset



(d) Speedup

Figure 3: covtype dataset



(d) Speedup

Figure 4: rcv1 dataset

Figure 8: Speedup for news20, covtype, rcv1 datasets

Q & A