Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization

Shai Shalev-Shwartz & Tong Zhang

Presented by Sijing Li & Yifei Liu

October 24, 2016

Outline

1. Introduction
2. Related Work
3. Basic Results
4. Examples
5. Experimental Results

Introduction

- $x_1, \dots, x_n$: vectors in $\mathbb{R}^d$
- $\phi_1, \dots, \phi_n$: a sequence of scalar convex functions
- Goal: find $w^* = \operatorname{argmin}_{w \in \mathbb{R}^d} P(w)$

$$P(w) = [\frac{1}{n} \sum_{i=1}^{n} \phi_i(w^T x_i) + \frac{\lambda}{2} ||w||^2]. \tag{1}$$

- A solution w is ϵ_P-sub-optimal if $P(w) - P(w^*) \leq \epsilon_P$.

Introduction

A simple approach for solving SVM is stochastic gradient descent (SGD) .
SGD finds an ϵ_P-sub-optimal solution in time $\tilde{O}(1/(\lambda\epsilon_P))$.
Disadvantages of SGD:

- does not have a clear stopping criterion;
- too aggressive at the beginning of the optimization process, especially when λ is very small;
- slow convergence in more accurate solutions.

Introduction

An alternative approach is dual coordinate ascent (DCA), which solves a *dual* problem of (1).

- for each i let $\phi_i^* : \mathbb{R} \to \mathbb{R}$ be the convex conjugate of ϕ_i, namely, $\phi_i^*(u) = \max_z(zu - \phi_i(z))$.
- The dual problem is

$$\max_{\alpha \in \mathcal{R}^n} D(\alpha) \quad \text{where} \quad D(\alpha) = [\frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2}||\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i||^2]. \qquad (2)$$

Introduction

- Define

$$w(\alpha) = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i, \tag{3}$$

it is known that $w(\alpha^*) = w^*$, where α^* is an optimal solution of (2).

- It is also known that $P(w^*) = D(\alpha^*)$ which immediately implies that for all w and α, they have $P(w) \geq D(\alpha)$, and hence the duality gap defined as

$$P(w(\alpha)) - D(\alpha),$$

and

$$P(w(\alpha)) - D(\alpha) \geq P(w(\alpha)) - P(w^*).$$

Introduction

Their main findings are: in order to achieve a duality gap of ϵ,

- For L-Lipschitz loss functions, they obtain the rate of $\tilde{O}(n + L^2/(\lambda\epsilon))$.
- For $(1/\gamma)$-smooth loss functions, they obtain the rate of $\tilde{O}((n + 1/(\lambda\epsilon)) \log(1/\epsilon))$.
- For loss functions which are almost everywhere smooth (such as the hinge-loss), they can obtain rate better than the above rate for Lipschitz loss.

Related Work

Table: Lipschitz loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
SDCA	primal-dual	$\tilde{O}(n + \frac{1}{\lambda\epsilon})$ or faster

Table: Smooth loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
SDCA	primal-dual	$\tilde{O}((n + \frac{1}{\lambda}) \log \frac{1}{\epsilon})$

Basic Results: Vanilla SDCA

Algorithm 1 Procedure SDCA($\alpha^{(0)}$)

```
**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 $\bar{w}$
**Output (Random option):**
Let $\bar{\alpha} = \alpha^{(t)}$ and $\bar{w} = w^{(t)}$ for some random $t \in T_0 + 1, \dots, T$
return $\bar{w}$
```

Vanilla SDCA

Theorem 2

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

$$T \geq T_0 + n + \frac{4L^2}{\lambda \epsilon_P} \geq \max(0, \lceil n \log(0.5\lambda n L^{-2}) \rceil) + n + \frac{20L^2}{\lambda \epsilon_P}.$$

Moreover, when $t \geq T_0$, we have **dual sub-optimality bound** of $\mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})] \leq \epsilon_P/2$.

Vanilla SDCA

Theorem 5

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

$$T \geq (n + \frac{1}{\lambda\gamma}) \log((n + \frac{1}{\lambda\gamma}) \cdot \frac{1}{\epsilon_P}).$$

Moreover, To obtain an expected **duality gap** of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_P$, it suffices to have a total number of iterations of $T > T_0$ where

$$T_0 \geq (n + \frac{1}{\lambda\gamma}) \log((n + \frac{1}{\lambda\gamma}) \cdot \frac{1}{(T - T_0)\epsilon_P}).$$

Basic Results: SDCA with SGD initialization

Algorithm 2 Procedure Modified-SGD

```
Initialize: $w^{(0)} = 0$
Iterate: for $t = 1, 2, \ldots, n$:
Find $\alpha_t$ to maximize $-\phi_t^*(-(\alpha_t) - \frac{\lambda t}{2}||w^{(t-1)} + (\lambda t)^{-1}\Delta\alpha_t x_t||^2$.
Let $w^{(t)} = \frac{1}{\lambda t}\sum_{i=1}^{t} \alpha_i x_i$
return $\alpha$
```

SDCA with SGD initialization

Algorithm 3 Procedure SDCA with SGD Initialization

```
    **Stage 1:** call Procedure Modified-SGD and obtain $\alpha$
    **Stage 2:** call Procedure SDCA with parameter $\alpha^{(0)} = \alpha$
```

SDCA with SGD initialization

Theorem 11

Assume that ϕ_i is L-**Lipschitz** for all i. In addition, assume that (ϕ_i, x_i) are iid samples from the same distribution for all $i = 1, \ldots, n$. Consider Procedure SDCA with SGD Initialization. To obtain a **duality gap** of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \le \epsilon_P$ at Stage 2, it suffices to have a total number of SDCA iterations of

$$T \ge T_0 + n + \frac{4L^2}{\lambda \epsilon_P} \ge \lceil n \log(\log(en)) \rceil + n + \frac{20L^2}{\lambda \epsilon_P}.$$

Moreover, when $t \ge T_0$, we have **dual sub-optimality bound** of $\mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})] \le \epsilon_P/2$.

Refined analysis for almost smooth loss

- Many loss functions are nearly smooth everywhere.
- A chance of linear convergence.
- For SVM, if $\exists s_0$ such that $\lambda n|{w^*}^T x_i y_i - 1| \geq s_0$ for all i, then linear rate can be obtained.

Examples: SDCA-Perm

Algorithm 4 Procedure SDCA-Perm($\alpha^{(0)}$)

```
**Let** $w^{(0)} = w(\alpha^{(0)})$
**Let** $t = 0$
**Iterate:** for epoch $k = 1, 2, \ldots$
**Let** $\{i_1, \ldots, i_n\}$ be a random permutation of $\{1, \ldots, n\}$
**Iterate:** for $j = 1, 2, \ldots, n$:
$t \leftarrow t + 1$
$i \leftarrow i_j$
Find $\Delta\alpha_i$ to increase dual $(*)$
$\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$
$w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1} \Delta\alpha_i x_i$
**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 $\bar{w}$
**Output (Random option):**
Let $\bar{\alpha} = \alpha^{(t)}$ and $\bar{w} = w^{(t)}$ for some random $t \in T_0 + 1, \ldots, T$
return $\bar{w}$
```

Experimental Results: Data

Data Set	Training Size	Testing Size	Features	Sparsity
astro-ph	29882	32487	99757	0.08%
CCAT	781265	23149	47236	0.16%
cov1	522911	58101	54	22.22%

Linear Convergence For Smooth Hinge-loss

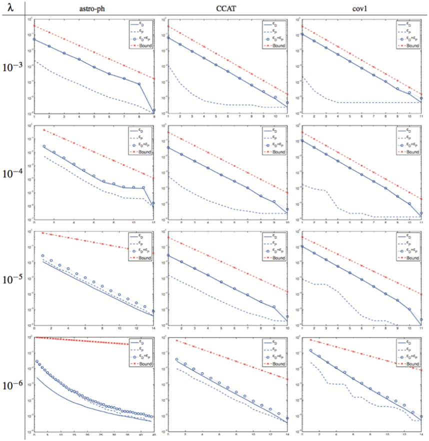

Figure 1: Experiments with the smoothed hinge-loss ($\gamma = 1$). The primal and dual sub-optimality, the duality gap, and our bound are depicted as a function of the number of epochs, on the astro-ph (left), CCAT (center) and cov1 (right) data sets. In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

Convergence For Non-smooth Hinge-loss

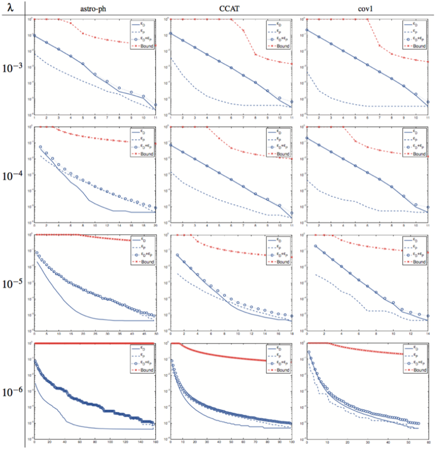

Figure 2: Experiments with the hinge-loss (non-smooth). The primal and dual sub-optimality, the duality gap, and our bound are depicted as a function of the number of epochs, on the astro-ph (left), CCAT (center) and cov1 (right) data sets. In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

Effect Of Smoothness Parameter

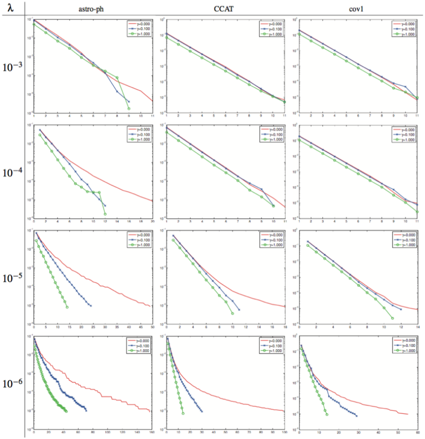

Figure 3: Duality gap as a function of the number of rounds for different values of γ.

Effect Of Smoothness Parameter

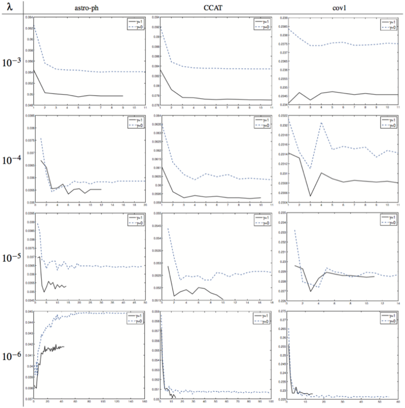

Figure 4: Comparing the test zero-one error of SDCA for smoothed hinge-loss ($\gamma = 1$) and non-smooth hinge-loss ($\gamma = 0$). In all plots the vertical axis is the zero-one error on the test set and the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data). We terminated each method when the duality gap was smaller than 10^{-5}.

Cyclic vs. Stochastic vs. Random Permutation

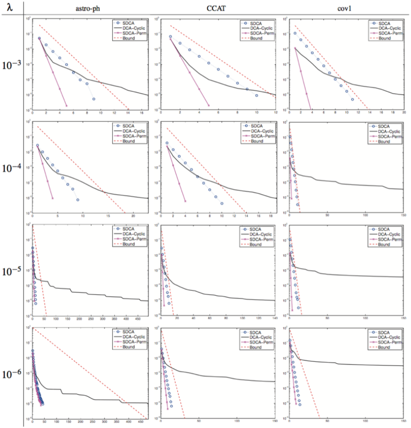

Figure 5: Comparing the duality gap achieved by choosing dual variables at random with repetitions (SDCA), choosing dual variables at random without repetitions (SDCA-Perm), or using a fixed cyclic order. In all cases, the duality gap is depicted as a function of the number of epochs for different values of λ. The loss function is the smooth hinge loss with $\gamma = 1$.

Comparison To SGD

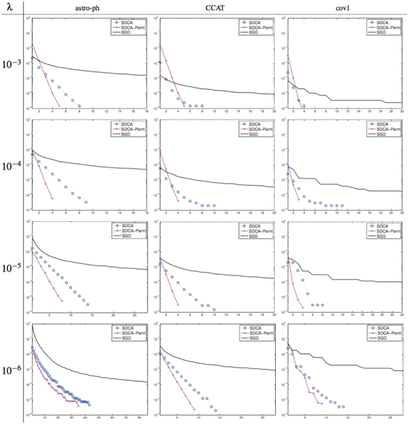

Figure 6: Comparing the primal sub-optimality of SDCA and SGD for the smoothed hinge-loss ($\gamma = 1$). In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

Comparison To SGD

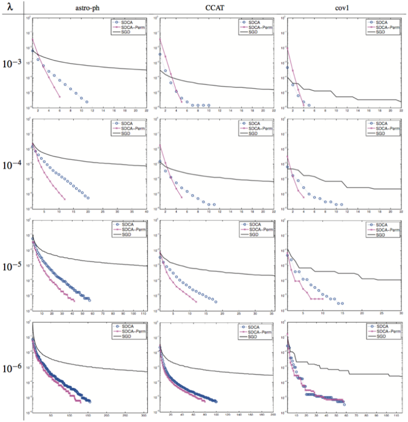

Figure 7: Comparing the primal sub-optimality of SDCA and SGD for the non-smooth hinge-loss ($\gamma = 0$). In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

Comparison To SGD

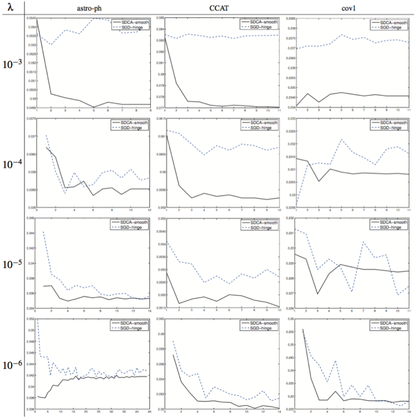

Figure 8: Comparing the test error of SDCA with the smoothed hinge-loss ($\gamma = 1$) to the test error of SGD with the non-smoothed hinge-loss. In all plots the vertical axis is the zero-one error on the test set and the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data). We terminated SDCA when the duality gap was smaller than 10^{-5}.

Take home messages

- Randomization helps!
- Structure helps!

Thank You!