Stochastic Dual Coordinate Ascent with Adaptive Probabilities [1]

Dominik Csiba, University of Edinburgh
Zheng Qu, University of Edinburgh
Peter Richtárik, University of Edinburgh

Problem

We are solving the **E**mpirical **R**isk **M**inimization problem

$$\min_{w\in\mathbb{R}^d}\left[P(w)\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\phi_i(A_i^\top w)+\lambda g(w)\right]$$

$\phi_1,\dots,\phi_n:\mathbb{R}\to\mathbb{R}$ are $1/\gamma$-smooth convex functions ("risk")

$\lambda>0$ ("regularization parameter")

P is a strongly convex function ("regularized empirical risk")

g is a 1-strongly convex function ("regularizer")

using the corresponding dual problem

$$\max_{\alpha\in\mathbb{R}^n}\left[D(\alpha)=-\lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^n A_i\alpha_i\right)-\frac{1}{n}\sum_{i=1}^n\phi_i^*(-\alpha_i)\right]$$

$h^*:\mathbb{R}^k\to\mathbb{R}$ is the convex (Fenchel) conjugate of $h:\mathbb{R}^k\to\mathbb{R}$ defined as $h^*(u)=\sup_s\{s^\top u-h(s)\}$.

D is a strongly concave function

$A_1,\dots,A_n\in\mathbb{R}^d$ ("examples")

Why ERM?

Setup: Object-label pairs $(A_i,y_i)\in\mathbb{R}^d\times\mathbb{R}$ appear naturally in the world with unknown distribution $\mathcal{D}$.
Goal: Find (**train**) a vector $w\in\mathbb{R}^d$ (**linear predictor**), such that, in some sense, for $(A_i,y_i)\sim\mathcal{D}$ we get

$$A_i^\top w\ \approx\ y_i.$$

This allows us to **predict** the **label** y_i by observing the **example** A_i. More precisely, we wish to find w solving

$$\min_w \mathbf{E}_{(A_i,y_i)\sim\mathcal{D}}\left[loss(A_i^T w,y_i)\right],$$

where *loss* is an appropriately chosen **loss function**.
ERM paradigm: Collect i.i.d. samples $(A_i,y_i)\sim\mathcal{D}$, $i=1,2,\dots,n$, and replace the expectation with the **sample average:**

$$\min_{w\in\mathbb{R}^d}\left[\frac{1}{n}\sum_{i=1}^n\phi_i(A_i^\top w)\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n loss(A_i^\top w,y_i)\right].$$

Specific risk functions ϕ_i lead to: **least squares**, **logistic regression**, **SVM**, ...

Main contributions

- **Two new algorithms**
  - **AdaSDCA** theoretical (with convergence analysis)
  - **AdaSDCA+** efficient variant of AdaSDCA
- **Adaptivity**
  Our algorithms are **SDCA-like** [2] - they iteratively update a single randomly chosen dual variable. The probability distribution over the dual coordinates **adaptively changes** on each iteration. Our method is the **first method with a theoretical guarantee with an adaptive probability law.**
- **Convergence rate**
  We prove that **AdaSDCA enjoys better rate than the best known rate for SDCA with fixed sampling [3], [4].**

Key concept: Dual residue

It is well known that the optimal primal-dual pair $(w^*,\alpha^*)\in\mathbb{R}^d\times\mathbb{R}^n$ satisfies the following **optimality conditions**:

$$\textbf{OPT1}: w^*=\nabla g^*\left(\frac{1}{\lambda n}A\alpha^*\right)\stackrel{\text{def}}{=}\nabla g^*\left(\frac{1}{\lambda n}\sum_{i=1}^n A_i\alpha_i^*\right)$$

$$\textbf{OPT2}: \alpha_i^*=-\nabla\phi_i(A_i^\top w^*),\quad \forall i\in[n]\stackrel{\text{def}}{=}\{1\dots n\}.$$

A key concept of this paper is the **dual residue**

$$\kappa_i=\nabla\phi_i(A_i^\top w)+\alpha_i,$$

which is a **measure of optimality** of the current pair of variables (w,α) based on **OPT2**.

AdaSDCA [1]

Init: $v_i=A_i^\top A_i$ for $i\in[n]$; $\alpha^0\in\mathbb{R}^n$; $\bar\alpha^0=\frac{1}{\lambda n}A\alpha^0$
for $t\geq 0$ **do**
  Primal update: $w^t=\nabla g^*(\bar\alpha^t)$ — Maintaining **OPT1**
  Set: $\alpha^{t+1}=\alpha^t$
  Compute residue κ^t:

$$\kappa_i^t=\nabla\phi_i(A_i^\top w^t)+\alpha_i^t,\quad\forall i\in[n]$$

  Compute probability distribution $p^t\in\mathbb{R}_+^n$ coherent with κ^t
  Pick $i\in[n]$ with probability p_i^t
  Compute:

$$\Delta=\arg\max_{h\in\mathbb{R}}\left\{-\phi_i^*(-(\alpha_i^t+h))-A_i^\top w^t h-\tfrac{v_i}{2\lambda n}|h|^2\right\}$$

  Dual update: $\alpha_i^{t+1}=\alpha_i^t+\Delta$
  Average update: $\bar\alpha^t=\bar\alpha^t+\frac{\Delta}{\lambda n}A_i$ — Maintaining $\bar\alpha^t=\frac{1}{\lambda n}A\alpha^0$
end for
Output: w^t,α^t

Inefficiency addressed by AdaSDCA+

Definition: p is *coherent* with κ if $\kappa_i\neq 0\Rightarrow p_i>0,\ \forall i$

Convergence results

Theorem Consider AdaSDCA. If at each iteration $t\geq 0$, p^t is coherent with κ^t and

$$\theta(\kappa,p)\stackrel{\text{def}}{=}\frac{n\lambda\gamma\sum_{i:\kappa_i\neq0}|\kappa_i|^2}{\sum_{i:\kappa_i\neq0}p_i^{-1}|\kappa_i|^2(v_i+n\lambda\gamma)}\leq\min_{i:\kappa_i\neq0}p_i,$$

then

$$\mathbb{E}[P(w^t)-D(\alpha^t)]\leq\frac{1}{\tilde\theta_t}\left(\prod_{k=0}^t(1-\tilde\theta_k)\right)\left(D(\alpha^*)-D(\alpha^0)\right),$$

for all $t\geq 0$ where

$$\tilde\theta_t\stackrel{\text{def}}{=}\frac{\mathbb{E}[\theta(\kappa^t,p^t)(P(w^t)-D(\alpha^t))]}{\mathbb{E}[P(w^t)-D(\alpha^t)]}.$$

The **Importance Sampling** [3] satisfies the assumptions of the theorem (but is **suboptimal** and hence **AdaSDCA does better!!**):

$$p_i^*\stackrel{\text{def}}{=}\frac{v_i+n\lambda\gamma}{\sum_{j=1}^n(v_j+n\lambda\gamma)},\quad\forall i\in[n].$$

Corollary Consider **AdaSDCA with importance sampling:** $p^t=p^*$ for all $t\geq 0$. Then

$$\mathbb{E}[P(w^t)-D(\alpha^t)]\leq\frac{1}{\theta_*}(1-\theta_*)^t\left(D(\alpha^*)-D(\alpha^0)\right),$$

where $\theta_*=n\lambda\gamma/\sum_{i=1}^n(v_i+\lambda\gamma n)$. This means that

$$T\geq\left(n+\frac{\frac{1}{n}\sum_{i=1}^n v_i}{\lambda\gamma}\right)\log\left(\frac{c}{\epsilon}\right)\Rightarrow\mathbb{E}[P(w^T)-D(\alpha^T)]\leq\epsilon,$$

where $c>0$ is some constant.

AdaSDCA+ [1]

We introduce an efficient variant of AdaSDCA, which has the **same computational costs as SDCA** with importance sampling, while still maintaining the advantages of adaptivity.

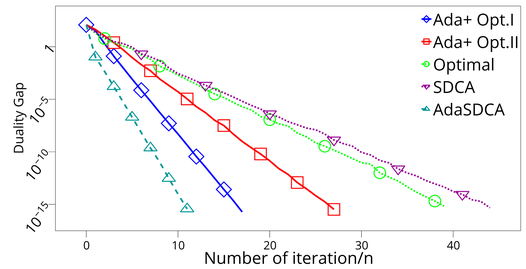
Algorithm	cost of an epoch
SDCA & QUARTZ	$O(\text{nnz}(A))$
IProx-SDCA	$O(\text{nnz}(A)+n\log(n))$
AdaSDCA	$O(n\cdot\text{nnz}(A))$
AdaSDCA+	$O(\text{nnz}(A)+n\log(n))$

Instead of updating the whole probability distribution p^t at each iteration, we divide the algorithm into **epochs**. At the beginning of each epoch we calculate the optimal adaptive probability and during the epoch we **update only one entry of p^t per iteration**. This can be efficiently done using **random counters** [5].

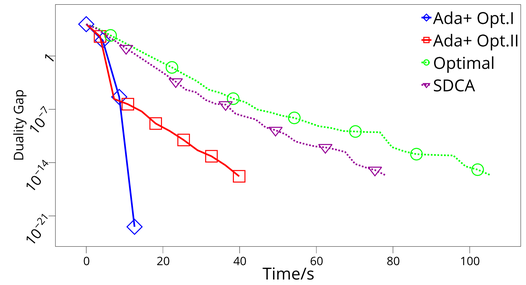
Computational experiments

All of the experiments were done on a laptop.

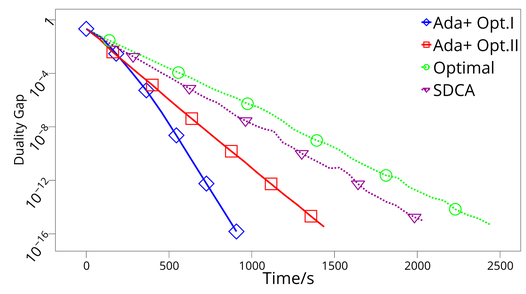
ijcnn1 dataset: $d=22, n=49,990$
quadratic loss with L_2 regularizer, $\lambda=1/n,\gamma=1$



cov1 dataset: $d=54, n=581,012$
Smooth Hinge loss with L_2 regularizer, $\lambda=1/n,\gamma=1$



synthetic dataset: $d=100, n=10^7$, sparsity $=0.1$
Smooth Hinge loss with L_2 regularizer, $\lambda=1/n,\gamma=1$



References

[1] Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. *ICML 2015.*

[2] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

[3] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. *arXiv:1401.2753*, 2014.

[4] Zheng Qu, Peter Richtárik, and Tong Zhang. Randomized Dual Coordinate Ascent with Arbitrary Sampling. *arXiv:1411.5873*, 2014.

[5] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.