

Stochastic Dual Coordinate Ascent with Adaptive Probabilities

Dominik Csiba, Zheng Qu and Peter Richtárik

University of Edinburgh

ICML

06-11 July 2015, Lille

Empirical Risk Minimization

ERM:

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

- supervised learning;
- train a linear predictor $w \in \mathbb{R}^d$;
- n training samples $A_1, \dots, A_n \in \mathbb{R}^d$;
- convex and $1/\gamma$ -smooth loss function $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$;
 - ex.: Squared loss ($\phi_i(a) = \frac{1}{2\gamma}(a - b_i)^2$), Logistic loss ($\phi_i(a) = \frac{4}{\gamma} \log(1 + e^a)$), ...
- 1-strongly convex regularizer $g : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$;
 - ex.: L_2 regularization ($g(w) = \frac{1}{2}\|w\|_2^2$), ...

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha) \stackrel{\text{def}}{=} \underbrace{-\lambda g^* \left(\frac{1}{\lambda n} \sum_{i=1}^n A_i \alpha_i \right)}_{\text{smooth}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \phi_i^*(-\alpha_i)}_{\text{strongly convex and separable}}$$

- Optimality conditions:

$$\mathbf{OPT1} : w^* = \nabla g^* \left(\frac{1}{\lambda n} A \alpha^* \right)$$

$$\mathbf{OPT2} : \alpha_i^* = -\nabla \phi_i(A_i^\top w^*), \quad \forall i = 1, \dots, n.$$

Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^*\left(\frac{1}{\lambda n} A \alpha^t\right)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \dots, n\}$;**
2. **Randomly pick $i_t \in \{1, \dots, n\}$ according to a fixed distribution p ;**

3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg \max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^*(-\beta) - (A_{i_t}^\top w^t) \beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Uniform and Importance Sampling

Uniform sampling (SDCA: [Shalev-Shwartz & Zhang 13'],...)

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{1}{n},$$

Iteration complexity:

$$\tilde{O} \left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right)$$

Importance sampling (Iprox-SDCA: [Zhao & Zhang 15'],...)

$$p_i = \mathbf{Prob}(i_t = i) \sim \|A_i\|^2 + \lambda\gamma n,$$

Iteration complexity:

$$\tilde{O} \left(n + \frac{\frac{1}{n} \sum_{i=1}^n \|A_i\|^2}{\lambda\gamma} \right)$$

Adaptive Sampling

- Each dual variable has a natural measure of progress:

$$\kappa_i^t \stackrel{\text{def}}{=} \alpha_i^t + \nabla \phi_i(A_i^\top w^t), \quad i = 1, \dots, n$$

called **dual residue**.

- Optimality conditions:

$$\text{OPT1} : w^* = \nabla g^* \left(\frac{1}{\lambda n} A \alpha^* \right)$$

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

- A sampling distribution p is coherent with κ^t if for all $i \in [n]$:

$$\kappa_i^t \neq 0 \Rightarrow p_i > 0.$$

Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^*(\frac{1}{\lambda n} A \alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \dots, n\}$ according to a fixed distribution p ;**

3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg \max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^*(-\beta) - (A_{i_t}^\top w^t) \beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Adaptive Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^*(\frac{1}{\lambda n} A \alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \dots, n\}$ according to a distribution p^t coherent with dual residue κ^t ;**

3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg \max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^*(-\beta) - (A_{i_t}^\top w^t) \beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Convergence Theorem

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \stackrel{\text{def}}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1} (\|A_i\|^2 + n\lambda\gamma) |\kappa_i^t|^2} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^t (1 - \tilde{\theta}_k) (D(\alpha^*) - D(\alpha^0)),$$

for all $t \geq 0$ where

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

Optimal Adaptive Sampling Probability

$$\begin{aligned} \rho^*(\kappa^t) = \arg \max \quad & \theta(\kappa^t, \rho) \\ \text{s.t.} \quad & \rho \in \mathbb{R}_+^n, \quad \sum_i \rho_i = 1 \\ & \rho \text{ is coherent with } \kappa^t \\ & \theta(\kappa^t, \rho) \leq \min_{i:\kappa_i^t \neq 0} \rho_i \end{aligned}$$

Relaxation:

$$\begin{aligned} \tilde{\rho}^*(\kappa^t) = \arg \max \quad & \theta(\kappa^t, \rho) \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (\rho_i)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \\ \text{s.t.} \quad & \rho \in \mathbb{R}_+^n, \quad \sum_{i=1}^n \rho_i = 1 \end{aligned}$$

$$(\tilde{\rho}^*(\kappa^t))_i \sim |\kappa_i^t| \sqrt{\|A_i\|^2 + n\lambda\gamma}, \quad \forall i \in [n].$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \stackrel{\text{def}}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^t (1 - \tilde{\theta}_k) (D(\alpha^*) - D(\alpha^0)),$$

for all $t \geq 0$ where

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

Heuristic and Efficient Variant of AdaSDCA

AdaSDCA+:

Dual solution

For $t \geq 1$:

1. If $\text{mod}(t, n) = 0$, then

Option I: Adaptive Sampling Probability

Compute dual residue κ^t : $\kappa_i^t = \alpha_i^t + \nabla \phi_i(A_i^\top w^t)$

Set $p_i^t \sim |\kappa_i^t| \sqrt{\|A_i\|^2 + n\lambda\gamma}$

Option II: Importance Sampling Probability

Set $p_i^t \sim \|A_i\|^2 + n\lambda\gamma$

2. Randomly pick $i_t \in \{1, \dots, n\}$ according to p^t
3. Update $\alpha_{i_t}^t$
4. Update Probability: $p^{t+1} \sim (p_1^t, \dots, p_{i_t}^t/m, \dots, p_n^t)$

Computational Cost per Epoch

ALGORITHM	COST OF AN EPOCH
SDCA	$O(\text{nnz})$
Iprox-SDCA	$O(\text{nnz} + n \log(n))$
ADASDCA	$O(n \cdot \text{nnz})$
ADASDCA+	$O(\text{nnz} + n \log(n))$

Table 1: One epoch computational cost of different algorithms

Numerical Experiments

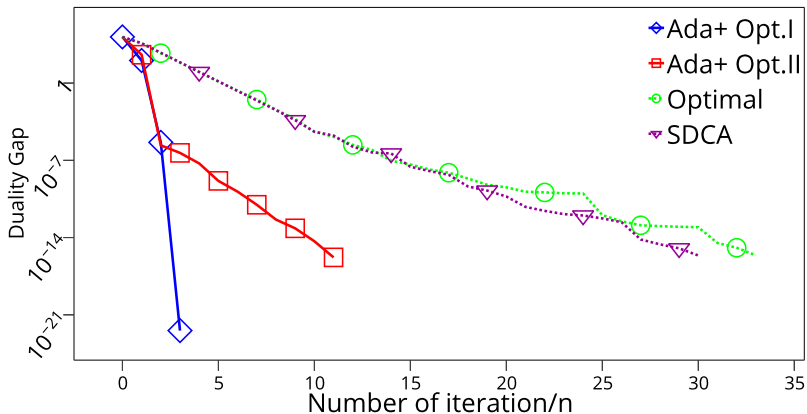


Figure 1: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.