

Randomized Dual Coordinate Ascent with Arbitrary Sampling

Zheng Qu ¹ Peter Richtárik ¹ Tong Zhang ²

¹University of Edinburgh

²Rutgers University & Baidu

22nd ISMP
July 12-17 2015, Pittsburgh

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

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 ;
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 sampling (SDCA: [Shalev-Shwartz & Zhang '13])

$$p_i = \mathbf{Prob}(i_t = i) = \frac{1}{n}, \quad \forall i \in [n].$$

Analysis of SDCA and Iprox-SDCA

$$\mathbb{E}_t \left[\underbrace{D(\alpha^{t+1}) - D(\alpha^t)}_{\text{dual value increase at iteration } t} \right] \geq \theta \underbrace{(P(w^t) - D(\alpha^t))}_{\text{primal dual gap at iteration } t}, \quad \forall t \geq 0.$$



$$\mathbb{E} \left[\underbrace{P(w^t) - D(\alpha^t)}_{\text{expected primal dual gap at iteration } t} \right] \leq \frac{1}{\theta} (1 - \theta)^t \underbrace{(D(\alpha^*) - D(\alpha^0))}_{\text{initial dual optimality gap}}, \quad \forall t \geq 0.$$

where

$$\theta = \begin{cases} \min_i \frac{\lambda \gamma}{\|A_i\|^2 + \lambda \gamma n} & \text{if } p_i \sim \frac{1}{n} \text{ (SDCA)} \\ \frac{\lambda \gamma n}{\sum_{i=1}^n (\|A_i\|^2 + \lambda \gamma n)} & \text{if } p_i \sim \|A_i\|^2 + \lambda \gamma n \text{ (Iprox SDCA)} \end{cases}$$

Iteration Complexity Results

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

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

Iteration complexity:

$$T \geq \left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right) \log \left(\left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right) \left(\frac{D(\alpha^*) - D(\alpha^0)}{\epsilon} \right) \right)$$

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

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{\|A_i\|^2}{\sum_{i=1}^n \|A_i\|^2 + \lambda\gamma n},$$

Iteration complexity:

$$T \geq \left(n + \frac{\sum_{i=1}^n \|A_i\|^2}{n\lambda\gamma} \right) \log \left(\left(n + \frac{\sum_{i=1}^n \|A_i\|^2}{n\lambda\gamma} \right) \left(\frac{D(\alpha^*) - D(\alpha^0)}{\epsilon} \right) \right).$$

Main Contributions

Quartz (Randomized dual coordinate ascent method with arbitrary sampling)

- arbitrary sampling:
 - at each iteration, update a **random subset** $S_t \subset [n]$ of dual variables
- direct primal-dual analysis:

$$\mathbb{E}_t [P(w^{t+1}) - D(\alpha^{t+1})] \leq (1-\theta)(P(w^t) - D(\alpha^t)), \quad \forall t \geq 0.$$



$$T \geq \frac{1}{\theta} \log\left(\frac{P(w^0) - D(\alpha^0)}{\epsilon}\right)$$

- additional data-driven speedup

Algorithm (Quartz)

Primal solution

For $t \geq 0$:

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

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. Generate a random subset $S_t \subset [n]$ according to \hat{S} ;
3. For each $i \in S_t$ do:

$$\alpha_i^{t+1} = (1 - \theta p_i^{-1}) \alpha_i^t + \theta p_i^{-1} (-\nabla \phi_i(A_i^\top w^{t+1}))$$

$$p_i \stackrel{\text{def}}{=} \mathbf{Prob}(i \in \hat{S})$$

Interpretation through Fenchel's Duality

Primal dual gap ($\bar{\alpha} = \frac{1}{\lambda n} A\alpha$):

$$\begin{aligned} & P(w) - D(\alpha) \\ &= \lambda \underbrace{(g(w) + g^*(\bar{\alpha}) - \langle w, \bar{\alpha} \rangle)}_{\text{GAP}_g(w, \alpha)} + \frac{1}{n} \sum_{i=1}^n \underbrace{(\phi_i(A_i^\top w) + \phi_i^*(-\alpha_i) + \langle A_i^\top w, \alpha_i \rangle)}_{\text{GAP}_{\phi_i}(w, \alpha_i)} \\ &\geq 0. \end{aligned}$$



Optimality conditions:

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

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

Quartz as a Randomized SOR Method

$$0 = \lambda \underbrace{(g(w) + g^*(\bar{\alpha}) - \langle w, \bar{\alpha} \rangle)}_{GAP_g(w, \alpha)} + \frac{1}{n} \sum_{i=1}^n \underbrace{\phi_i(A_i^\top w) + \phi_i^*(-\alpha_i) + \langle A_i^\top w, \alpha_i \rangle}_{GAP_{\phi_i}(w, \alpha_i)}$$



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

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

Quartz:

$$w^t = (1 - \theta) w^{t-1} + \theta \nabla g^*(\bar{\alpha}^{t-1})$$

$$\alpha_i^t = \begin{cases} (1 - \theta p_i^{-1}) \alpha_i^{t-1} + \theta p_i^{-1} (-\nabla \phi_i(A_i^\top w^t)) & , i \in S_t \sim \hat{S} \\ \alpha_i^{t-1} & , i \notin S_t \end{cases}$$

Flexibility of Quartz

Arbitrary proper sampling:

$$p_i \stackrel{\text{def}}{=} \mathbf{Prob}(i \in \hat{S}) > 0, \quad i \in [n].$$

- serial sampling

$$\mathbf{Prob}(|\hat{S}| = 1) = 1$$

- standard mini-batching
 - τ -nice sampling:

$$\mathbf{Prob}(\hat{S} = S) = \frac{1}{C_n^\tau}, \quad \forall S \subset [n], |S| = \tau.$$

- distributed sampling
 - c nodes/machines:

$$\hat{S} = \hat{S}_1 \cup \hat{S}_2 \cdots \cup \hat{S}_c$$

- ...

Formula of θ

$$w^t = (1 - \theta)w^{t-1} + \theta \nabla g^*(\bar{\alpha}^{t-1})$$

$$\alpha_i^t = \begin{cases} (1 - \theta p_i^{-1}) \alpha_i^{t-1} + \theta p_i^{-1} (-\nabla \phi_i(A_i^\top w^t)) & , i \in S_t \sim \hat{S} \\ \alpha_i^{t-1} & , i \notin S_t \end{cases}$$

Theorem (Q., Richtarik & Zhang '14)

Let \hat{S} be a proper sampling. If

$$\theta = \min_i \frac{p_i \lambda \gamma n}{v_i + \lambda \gamma n}$$

where v_1, \dots, v_n satisfy:

$$\mathbb{E} \left\| \sum_{i \in \hat{S}} A_i \alpha_i \right\|^2 \leq \sum_{i=1}^n p_i v_i |\alpha_i|^2, \quad \forall \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq (1 - \theta)^t (P(w^0) - D(\alpha^0)), \quad \forall t \geq 0.$$

Quartz specialized to serial sampling

Lemma

If \hat{S} is a serial proper sampling, then

$$\mathbb{E} \left\| \sum_{i \in \hat{S}} A_i \alpha_i \right\|^2 \leq \sum_{i=1}^n p_i v_i |\alpha_i|^2, \quad \forall \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n,$$

holds for

$$v_i = \|A_i\|^2, \quad \forall i \in [n].$$

Corollary (Q., Richtárik & Zhang '14)

Let \hat{S} be a serial proper sampling. If

$$\theta = \min_i \frac{p_i \lambda \gamma n}{\|A_i\|^2 + \lambda \gamma n},$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq (1 - \theta)^t (P(w^0) - D(\alpha^0)), \quad \forall t \geq 0.$$

Comparison with SDCA

QUARTZ+ uniform sampling

$$T \geq \left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right) \log \left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right)$$

SDCA [S-Shwartz & Zhang '15]

$$T \geq \left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right) \log \left(\left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma} \right) \left(\frac{D(\alpha^*) - D(\alpha^0)}{\epsilon} \right) \right)$$

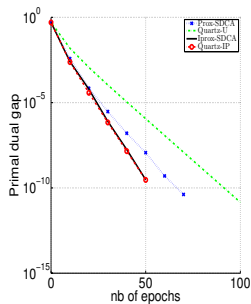
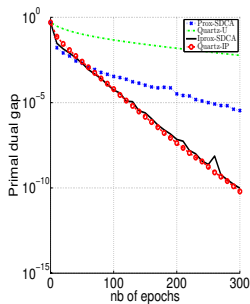
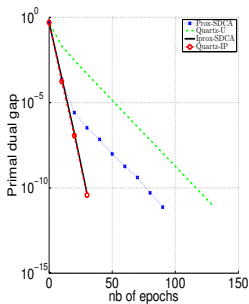
QUARTZ+ importance sampling

$$T \geq \left(n + \frac{\sum_i \|A_i\|^2}{n\lambda\gamma} \right) \log \left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right)$$

lprox-SDCA [Zhao & Zhang '15]

$$T \geq \left(n + \frac{\sum_i \|A_i\|^2}{n\lambda\gamma} \right) \log \left(\left(n + \frac{\sum_i \|A_i\|^2}{n\lambda\gamma} \right) \left(\frac{D(\alpha^*) - D(\alpha^0)}{\epsilon} \right) \right)$$

Experimental results



(a) cov1; $n = 522911$;

(b) w8a; $n = 49749$;

(c) ijcnn1; $n = 49990$;

Quartz:

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

SDCA:

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

Quartz Specialized to Mini-batching

Lemma (Richtarik & Takac '12, Fercoq & Richtarik '13)

If \hat{S} is a τ -nice sampling, then

$$\mathbb{E} \left\| \sum_{i \in \hat{S}} A_i \alpha_i \right\|^2 \leq \sum_{i=1}^n p_i v_i |\alpha_i|^2, \quad \forall \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n,$$

holds for

$$v_i = \sum_{j=1}^d \left(1 + \frac{(\omega_j - 1)(\tau - 1)}{n - 1} \right) A_{ji}^\top A_{ji}, \quad i \in [n], \quad (1)$$

where for each $j \in [d]$, ω_j is the number of nonzero blocks in the j -th row of matrix $A = (A_1, \dots, A_n)$, i.e.,

$$\omega_j \stackrel{\text{def}}{=} |\{i \in [n] : A_{ji} \neq 0\}|, \quad j \in [d]. \quad (2)$$

Iteration Complexity

For standard mini-batching of size τ :

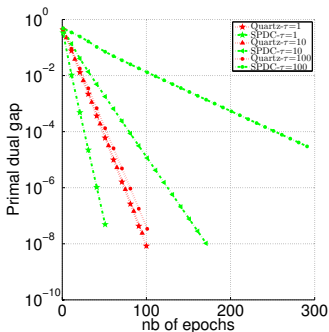
Fully sparse data ($\tilde{\omega} = 1$)	$\left(\frac{n}{\tau} + \frac{\max_i \ A_i\ ^2}{\lambda\gamma\tau} \right) \log\left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right)$
Fully dense data ($\tilde{\omega} = n$)	$\frac{n}{\tau} + \frac{\max_i \ A_i\ ^2}{\lambda\gamma}$
Any data ($1 \leq \tilde{\omega} \leq n$)	$\left(\frac{n}{\tau} + \frac{\left(1 + \frac{(\tilde{\omega}-1)(\tau-1)}{n-1}\right) \max_i \ A_i\ ^2}{\lambda\gamma\tau} \right) \log\left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right)$

Comparison with Accelerated Mini-Batch P-D methods

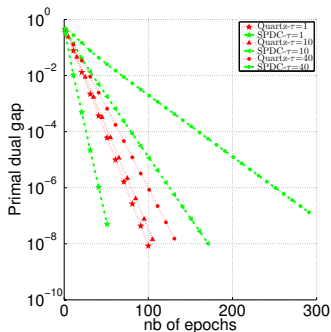
Algorithm	Iteration complexity	g
ASDCA [S-Shwartz & Zhang '13]	$\max \left\{ \frac{n}{\tau}, \sqrt{\frac{n}{\lambda\gamma\tau}}, \frac{1}{\lambda\gamma\tau}, \frac{n^{\frac{1}{3}}}{(\lambda\gamma\tau)^{\frac{2}{3}}} \right\}$	$\frac{1}{2} \ \cdot \ ^2$
SPDC [Zhang & Xiao '14]	$\frac{n}{\tau} + \sqrt{\frac{n}{\lambda\gamma\tau}}$	general
QUARTZ with τ -nice sampling	$\frac{n}{\tau} + \left(1 + \frac{(\tilde{\omega}-1)(\tau-1)}{n-1} \right) \frac{1}{\lambda\gamma\tau}$	general

Acc-Prox-SDCA [Shalev-Shwartz & Zhang '13]; APCG [Lin, Lu & Xiao '14]

Experimental results



(d) $n = 10^5$; sparsity 0.1%



(e) $n = 10^5$; sparsity 1%

Figure 1: Comparison of Quartz with SPDC for different mini-batch size τ in the regime $\kappa = 10n$.

- Summary
 - Quartz: randomized dual coordinate ascent method
 - arbitrary sampling
 - direct primal-dual analysis
 - additional data-driven speedup

- Future work

- Quartz as a randomized SOR:

$$w^t = (1 - \theta)w^{t-1} + \theta \nabla g^*(\bar{\alpha}^{t-1})$$

$$\alpha_i^t = \begin{cases} (1 - \theta p_i^{-1}) \alpha_i^{t-1} + \theta p_i^{-1} (-\nabla \phi_i(A_i^\top w^t)) & , i \in S_t \\ \alpha_i^{t-1} & , i \notin S_t \end{cases}$$

- Accelerated mini-batch dual coordinate ascent+data-driven speedup