Stochastic Dual Newton Ascent for Empirical Risk Minimization

Zheng Qu[1] Peter Richtárik[1] Martin Takáč[2] Olivier Fercoq[3]

[1] University of Edinburgh [2] Lehigh University [3] Telecom Paris-Tech

Introduction

We study the problem of minimizing the average of a large number of $1/\gamma$-smooth convex functions penalized with a 1-strongly convex regularizer.

$$\min_{w\in\mathbb{R}^d} P(w) := \frac{1}{n}\sum_{i=1}^{n} \phi_i(\mathbf{a}_i^\top \mathbf{w}) + \lambda \mathbf{g}(\mathbf{w}). \qquad \text{(P)}$$

Each $a_i \in \mathbb{R}^d$ and we write $\mathbf{A} = [a_1, \dots, a_n] \in \mathbb{R}^{d\times n}$. Let g^* and $\{\phi_i^*\}_i$ be the Fenchel conjugate functions of g and $\{\phi_i\}_i$, respectively. In the case of g, for instance, we have $g^*(s) = \sup_{w\in\mathbb{R}^d} \langle w, s\rangle - g(w)$.
The (Fenchel) dual problem of (P) can be written as:

$$\max_{\alpha\in\mathbb{R}^n} D(\alpha) := \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\mathbf{A}\alpha\right). \qquad \text{(D)}$$

The Algorithm

Sampling $\hat{S}$: A random subset of $\{1, 2, \dots, n\}$ such that $\forall i : \mathbf{Prob}(i \in \hat{S}) > 0$ and $\mathbf{Prob}(\hat{S} = \emptyset) = 0$.

Algorithm 1: SDNA Algorithm

1: **Initialization:** $\alpha^0 \in \mathbb{R}^n$; $\bar{\alpha}^0 = \frac{1}{\lambda n}\mathbf{A}\alpha^0$
2: **for** $k = 0, 1, 2, \dots$ **do**
3: Primal update: $w^k = \nabla g^*(\bar{\alpha}^k)$
4: **Generate a random set of blocks** $S_k \sim \hat{S}$
5: Compute:

$$\Delta\alpha^k = \arg\min_{\mathbf{h}\in\mathbb{R}^n} \langle \mathbf{A}^\top w^k, \mathbf{I}_{S_k}\mathbf{h}\rangle + \tfrac{1}{2}\mathbf{h}^\top \mathbf{X}_{S_k}\mathbf{h} + \textstyle\sum_{i\in S_k} \phi_i^*(-\alpha_i^k - \mathbf{h}_i)$$

6: Dual update: $\alpha^{k+1} := \alpha^k + (\Delta\alpha^k)_{S_k}$
7: Average update: $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \frac{1}{\lambda n}\sum_{i\in S_k} \Delta\alpha_i^k a_i$
8: **end for**

Where $\mathbf{X} = \mathbf{A}^T\mathbf{A}$ and $\mathbf{X}_{S_k}$ is the matrix obtained from $\mathbf{X}$ retaining elements $\mathbf{X}_{ij}$ for which both $i, j \in S_k$ and zeroing out all other elements.

Iteration Complexity of SDNA

Theorem: Let $\hat{S}$ be a uniform sampling and let $\tau := \mathbf{E}[|\hat{S}|]$. The output sequence $\{w^k, \alpha^k\}_{k\geq 0}$ of Algorithm 1 satisfies:

$$\mathbf{E}[P(w^k) - D(\alpha^k)] \leq \frac{(1-\sigma)^k}{\theta(\hat{S})}(D(\alpha^*) - D(\alpha^0)),$$

where $\sigma := \frac{\tau \min(1, s_1)}{n}$, $\theta(\hat{S}) := \min_i \frac{p_i \lambda\gamma n}{v_i + \lambda\gamma n}$, $s_1 = \lambda_{\min}\left[\left(\frac{1}{\tau\gamma\lambda}\mathbf{E}[(\mathbf{A}^\top\mathbf{A})_{\hat{S}}] + \mathbf{I}\right)^{-1}\right]$ and $v \in \mathbb{R}^n_{++}$ is a vector satisfying:

$$\mathbf{E}[(\mathbf{A}^\top\mathbf{A})_{\hat{S}}] \preceq \mathrm{diag}(p)\cdot\mathrm{diag}(v). \qquad (1)$$

Comparison with Mini-Batch SDCA

Algorithm 2: Minibatch SDCA

1: **Parameters:** uniform sampling $\hat{S}$, vector $v \in \mathbb{R}^n_{++}$
2: **Initialization:** $\alpha^0 \in \mathbb{R}^n$; set $\bar{\alpha}^0 = \frac{1}{\lambda n}\mathbf{A}\alpha^0$
3: **for** $k = 0, 1, 2, \dots$ **do**
4: Primal update: $w^k = \nabla g^*(\bar{\alpha}^k)$
5: Generate a random set of blocks $S_k \sim \hat{S}$
6: Compute for each $i \in S_k$

$$h_i^k = \arg\min_{h_i\in\mathbb{R}} h_i(a_i^\top w^k) + \frac{v_i}{2}|h_i|^2 + \phi_i^*(-\alpha_i^k - h_i)$$

7: Dual update: $\alpha^{k+1} := \alpha^k + \sum_{i\in S_k} h_i^k e_i$
8: Average update: $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \frac{1}{\lambda n}\sum_{i\in S_k} h_i^k a_i$
9: **end for**

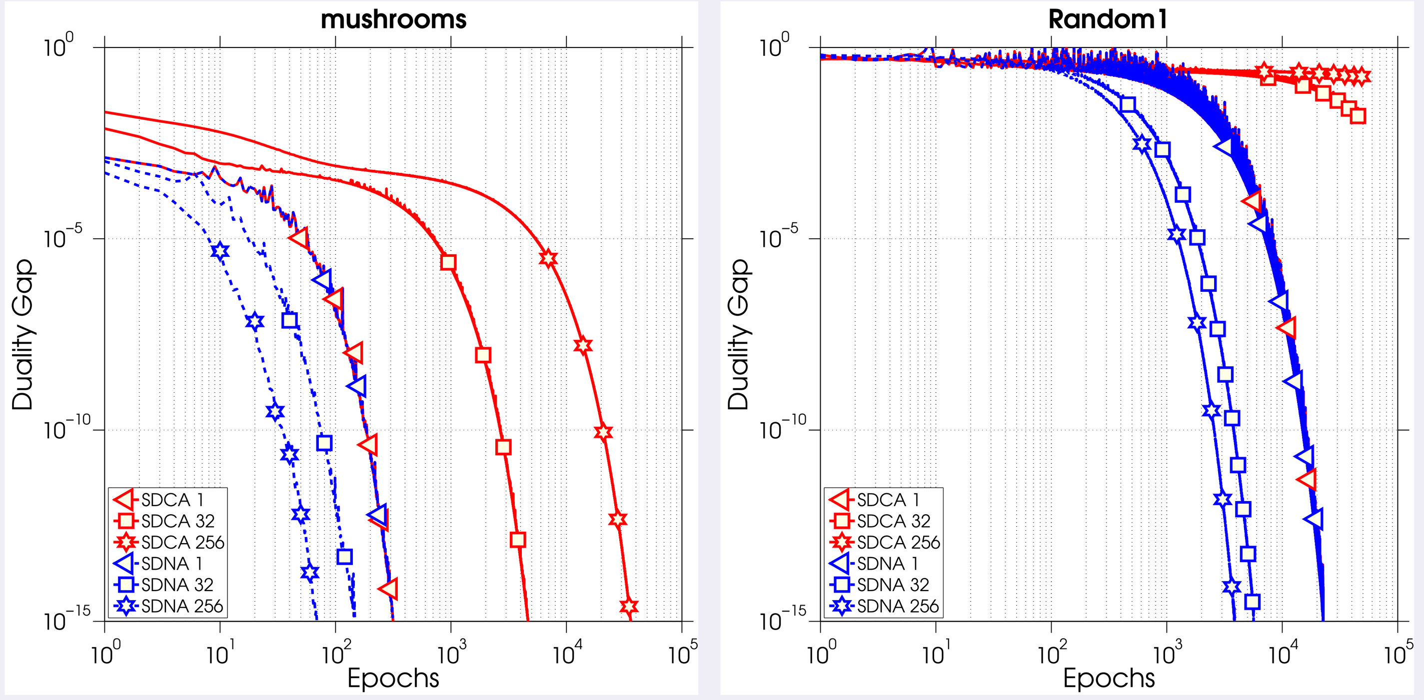
Theorem: If (1) holds, then the output sequence $\{w^k, \alpha^k\}_{k\geq 0}$ of Algorithm 2 satisfies:

$$\mathbf{E}[P(w^k) - D(\alpha^k)] \leq \frac{(1-\theta(\hat{S}))^k}{\theta(\hat{S})}\left(D(\alpha^*) - D(\alpha^0)\right).$$

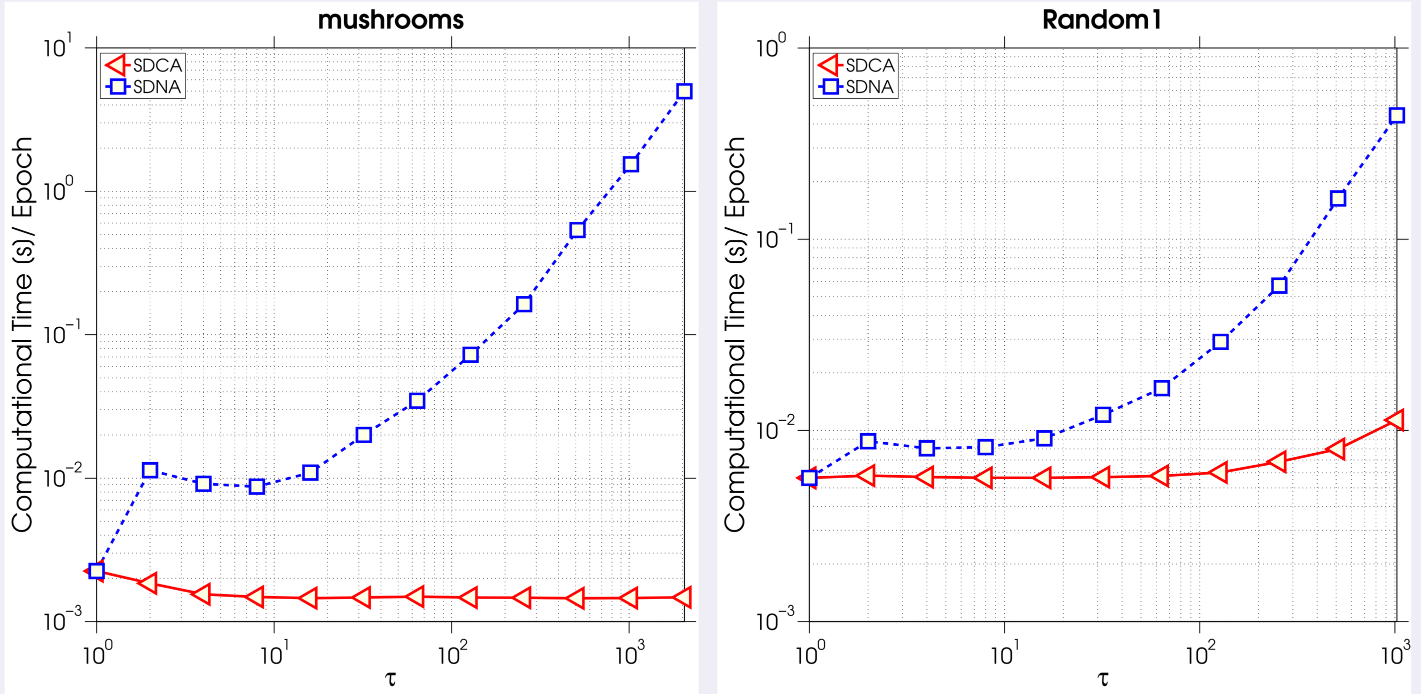
Moreover, $\theta(\hat{S}) \leq \sigma$.

Numerical Experiments

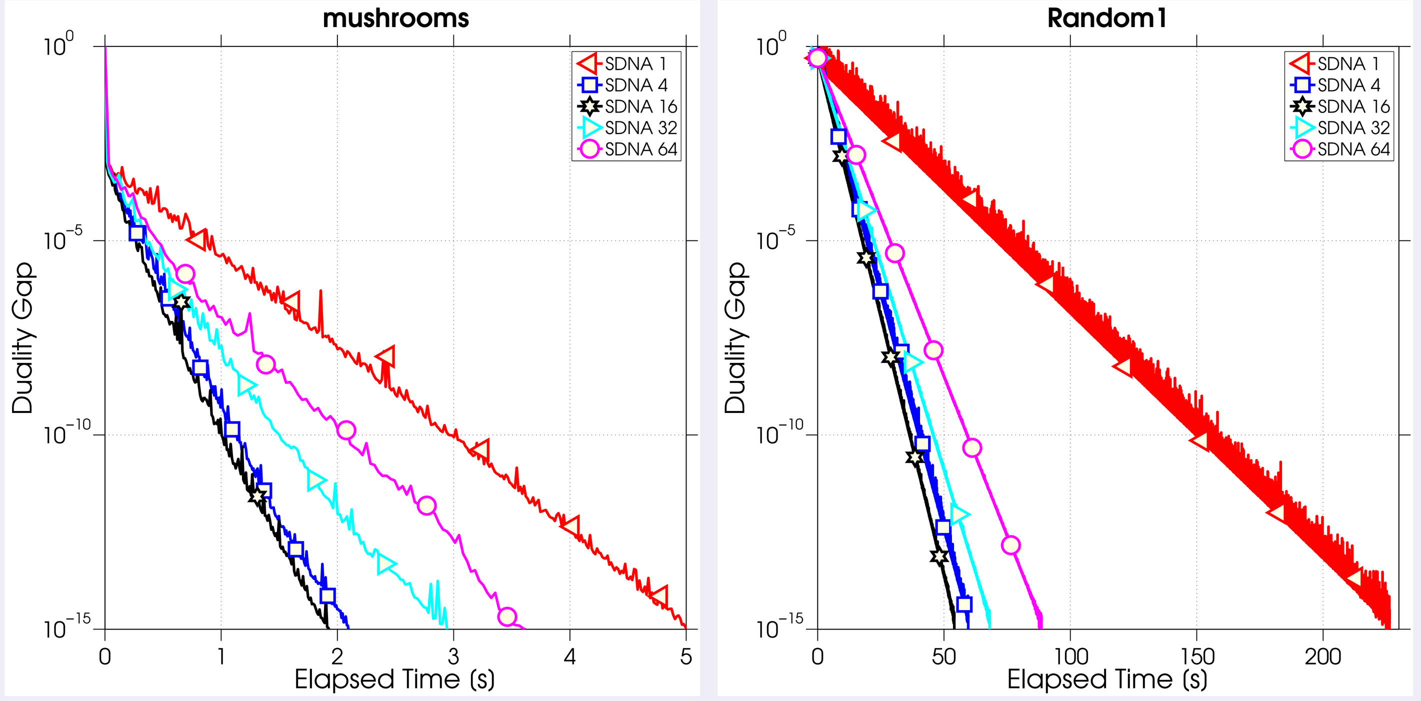
Comparison of SDNA and SDCA for minibatch sizes $\tau = 1, 32, 256$ on a real (left) and synthetic (right) dataset. The methods coincide for $\tau = 1$.



Time it takes for SDNA and SDCA to process a singe epoch as a function of the minibatch size τ.



Runtime of SDNA for minibatch sizes $\tau = 1, 4, 16, 32, 64$.



References

[1] Richtárik, P. and Takáč, M.: On optimal probabilities in stochastic coordinate descent methods, arXiv:1310.3438, 2013.

[2] Richtárik, P. and Takáč, M.: Parallel coordinate descent methods for big data optimization, arXiv:1212.0873, 2012.

[3] Richtárik, P. and Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function, Mathematical Programming, 2012.

[4] Takáč, M., Bijral, A., Richtárik, P. and Srebro, N.: Mini-batch primal and dual methods for SVMs, In ICML, 2013.

[5] Qu, Z., Richtárik, P. and Zhang, T.: Randomized dual coordinate ascent with arbitrary sampling, arXiv:1411.5873, 2014.

[6] Qu, Z., Richtárik, P., Takáč, M. and Fercoq, O.: SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization, arXiv:1502.02268, 2015.