Doubly Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization with Factorized Data

Adams Wei Yu[†], **Qihang Lin**[*], **Tianbao Yang**[*]
Carnegie Mellon University[†]
The University of Iowa[*]
weiyu@cs.cmu.edu, {qihang-lin, tianbao-yang}@uiowa.edu

Abstract

We proposed a doubly stochastic primal-dual coordinate optimization algorithm for regularized empirical risk minimization that can be formulated as a saddle-point problem. Different from existing coordinate methods, the proposed method randomly samples both primal and dual coordinates to update solutions, which is a desirable property when applied to data with both a high dimension and a large size. The convergence of our method is established not only in terms of the solution's distance to optimality but also in terms of the primal-dual objective gap. When applied to the data matrix already factorized as a product of two smaller matrices, we show that the proposed method has a lower overall complexity than other coordinate methods, especially, when data size is large.

1 Introduction

Setup We consider the following regularized empirical risk minimization (ERM) problem:

$$\min_{x\in\mathbb{R}^p}\left\{P(x)\equiv\frac{1}{n}\sum_{i=1}^{n}\phi_i(a_i^Tx)+\sum_{j=1}^{p}g_j(x_j)\right\}, \tag{1}$$

where $a_1,\ldots,a_n\in\mathbb{R}^p$ are n data points, $\phi_i:\mathbb{R}\to\mathbb{R}$ is a convex loss function, and $g_j:\mathbb{R}\to\mathbb{R}$ is a function of x_j, the j-th coordinate of x. We further assume that g_j is λ-strongly convex for $j=1,2,\ldots,p$ and ϕ_i is $(1/\gamma)$-smooth for $i=1,2,\ldots,n$. The dual problem of (1) is

$$\max_{y\in\mathbb{R}^n}\left\{D(y)\equiv-g^*\left(-\frac{A^Ty}{n}\right)-\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(y_i)\right\}, \tag{2}$$

where $A=[a_1,a_2,\ldots,a_n]^T\in\mathbb{R}^{n\times p}$ is the data matrix, and ϕ_i^* and g^* are the Fenchel's conjugates of ϕ and g, respectively. We denote the i-th row of A by a_i and the j-th column of A by A^j. Let $\|\cdot\|$ represents ℓ_2-norm. The maximum norm of data points is defined as $R=\max_{i=1,\ldots,n}\|A_i\|$. Both (1) and (2) corresponds to the following *saddle-point* problem

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n}\left\{g(x)+\frac{1}{n}y^TAx-\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(y_i)\right\}. \tag{3}$$

In this paper, we propose an efficient primal dual coordinated descent algorithm for the general problem (3) and also one for a specific problem when the data A is factorized.

Related Works For solving problem (3), efficient deterministic first-order methods have been developed, including smoothing method [15, 3], excessive gap method [14], extragradient method [10, 12], Mirror-Prox method [13] and primal-dual hybrid gradient methods [1, 2, 4]. These approaches

Algorithm 1 Doubly Stochastic Primal-Dual Coordinate (DSPDC) Method

Input: $x^{(-1)} = x^{(0)} = \bar{x}^{(0)} \in \mathbb{R}^p$, $y^{(-1)} = y^{(0)} = \bar{y}^{(0)} \in \mathbb{R}^n$, and positive parameters (θ, τ, σ)

For $t = 0, 1, 2, \ldots, T-1$

Uniformly and randomly choose two sets of indices $I \subset \{1, 2, \ldots, n\}$ and $J \subset \{1, 2, \ldots, p\}$ of sizes m and q, respectively.

$$y_i^{(t+1)} = \begin{cases} \arg\max_{\beta \in \mathbb{R}} \left\{ \frac{1}{n} \langle A_i, \bar{x}^{(t)} \rangle \beta - \frac{\phi_i^*(\beta)}{n} - \frac{1}{2\sigma} (\beta - y_i^{(t)})^2 \right\} & \text{if } i \in I, \\ y_i^{(t)} & \text{if } i \notin I, \end{cases} \tag{4}$$

$$\bar{y}^{(t+1)} = y^{(t)} + \frac{n}{m}(y^{(t+1)} - y^{(t)}), \tag{5}$$

$$x_j^{(t+1)} = \begin{cases} \arg\min_{\alpha \in \mathbb{R}} \left\{ \frac{1}{n} \langle A^j, \bar{y}^{(t+1)} \rangle \alpha + g_j(\alpha) + \frac{1}{2\tau} (\alpha - x_j^{(t)})^2 \right\} & \text{if } j \in J, \\ x_j^{(t)} & \text{if } j \notin J, \end{cases} \tag{6}$$

$$\bar{x}^{(t+1)} = x^{(t)} + (\theta + 1)(x^{(t+1)} - x^{(t)}). \tag{7}$$

Output: $x^{(T)}$ and $y^{(T)}$

need to evaluate the full (sub)gradient of objective function at each iteration which becomes prohibitive when primal dimension p or dual dimension n are both large. Recent years there have seen an increased interest in stochastic variance reduced gradient methods [8, 24, 17, 9] and incremental gradient methods [19, 6, 11] that makes use of all instances in computing the stochastic gradient, which can accelerate the conventional stochastic gradient decent method. *Stochastic coordinate* methods work by updating randomly sampled coordinates of decision variables [16, 18, 20]. In [7] the authors showed that randomized (block) coordinate descent methods can be accelerated by parallelization when applied to the problem of minimizing the sum of a partially separable smooth convex function and a simple separable convex function. Shalev-Shwartz & Zhang [22, 21, 23] proposed stochastic dual coordinate ascent (SDCA) and its mini-batch, accelerated and proximal variants to maximize the dual formulation (2). Zhang & Xiao [26] and Dong & Lan [5] both proposed stochastic primal-dual coordinate method for (3), which alternates between maximizing over a randomly chosen dual variable and minimizing over all the primal variables. However, all these need to update either full primal or full dual coordinates, which can still have a high computational cost in each iteration when data has both a large size and a high dimension.

2 Primal-Dual Algorithm for General Data Matrix

In this section, we propose a doubly stochastic primal-dual coordinate method in Algorithm 1 for problem (3). When ϕ_i is a $(1/\gamma)$-smooth and g is λ-strongly convex, the saddle-point problem (3) has a unique solution denoted by $(x^\star, y^\star)$ with $x^\star$ and $y^\star$ being the optimal primal and dual solutions for (1) and (2), respectively. The *condition number* of problem (3) is defined as $\kappa = \frac{R^2}{\lambda\gamma}$. Algorithm 1 requires three control parameters θ, τ and σ and its convergence is obtained after a proper choice of these parameters as shown in Theorem 1. All the proofs for the theorems here are deferred to our long version manuscript [25].

Theorem 1. *Suppose the parameters θ, τ and σ in Algorithm 1 are chosen so that*

$$\theta = \frac{p}{q} - \frac{p/q}{\frac{R}{\sqrt{\lambda\gamma}} \sqrt{\frac{n}{m}\frac{p}{q}} + \max\{\frac{n}{m}, \frac{p}{q}\}}, \qquad \tau\sigma = \frac{nq}{4pR^2}, \qquad \frac{p}{2q\lambda\tau} + \frac{p}{q} = \frac{n^2}{2m\gamma\sigma} + \frac{n}{m}, \tag{8}$$

where the last two equations are equivalent to

$$\tau = \frac{p}{q\lambda} \left(\left(\frac{n}{m} - \frac{p}{q} \right) + \sqrt{\left(\frac{n}{m} - \frac{p}{q} \right)^2 + \frac{4np^2R^2}{mq^2\lambda\gamma}} \right)^{-1} \sigma = \frac{n^2}{m\gamma} \left(\left(\frac{p}{q} - \frac{n}{m} \right) + \sqrt{\left(\frac{n}{m} - \frac{p}{q} \right)^2 + \frac{4np^2R^2}{mq^2\lambda\gamma}} \right)^{-1}. \tag{9}$$

For each $t \geq 0$, Algorithm 1 guarantees

$$\left(\frac{p}{2q\tau}+\frac{p\lambda}{q}\right)\mathbb{E}\|x^\star-x^{(t)}\|^2+\left(\frac{n}{4m\sigma}+\frac{\gamma}{m}\right)\mathbb{E}\|y^\star-y^{(t)}\|^2$$
$$\leq \left(1-\frac{1}{\max\left\{\frac{p}{q},\frac{n}{m}\right\}+\frac{R}{\sqrt{\lambda\gamma}}\sqrt{\frac{n}{m}\frac{p}{q}}}\right)^t\left[\left(\frac{p}{2q\tau}+\frac{p\lambda}{q}\right)\|x^\star-x^{(0)}\|^2+\left(\frac{n}{2m\sigma}+\frac{\gamma}{m}\right)\|y^\star-y^{(0)}\|^2\right].$$

Besides the distance to the saddle-point $(x^\star, y^\star)$, a more useful quality measure for the solution $(x^{(t)}, y^{(t)})$ is its primal-dual objective gap, $P(x^{(t)}) - D(y^{(t)})$, because it can be evaluated in each iteration and used as a stopping criterion in practice. The next theorem establishes the convergence rate of the primal-dual objective gap ensured by DSPDC.

Theorem 2. *Suppose the parameters τ and σ in Algorithm 1 are chosen as (9) and θ is chosen as*

$$\theta = \frac{p}{q} - \frac{p/q}{\frac{2R}{\sqrt{\lambda\gamma}}\sqrt{\frac{n}{m}\frac{p}{q}}+2\max\{\frac{n}{m},\frac{p}{q}\}}. \tag{10}$$

Algorithm 1 guarantees

$$\mathbb{E}\left[P(x^{(t)})-D(y^{(t)})\right]$$
$$\leq \left(1-\frac{1}{2\max\{\frac{n}{m},\frac{p}{q}\}+\frac{2R}{\sqrt{\lambda\gamma}}\sqrt{\frac{n}{m}\frac{p}{q}}}\right)^t\left\{\frac{1}{\min\left\{\frac{p}{q},\frac{n}{m}\right\}}+\frac{\max\left\{\frac{R^2}{2\gamma},\frac{R^2}{\lambda n}\right\}}{\min\left\{\frac{\lambda p}{q},\frac{\gamma}{m}\right\}}\right\}$$
$$\left[\left(\frac{p}{2q\tau}+\frac{p\lambda}{2q}\right)\|x^{(0)}-x^\star\|^2+\left(\frac{n}{2m\sigma}+\frac{\gamma}{2m}\right)\|y^{(0)}-y^\star\|^2+\max\left\{\frac{p}{q},\frac{n}{m}\right\}\left(P(x^{(0)})-D(y^{(0)})\right)\right].$$

For strongly convex problem, the convergence of objective value implies that of solution but the opposite is not true. Therefore, Theorem 2 is not a direct consequence of Theorem 1, especially when $P(x)$ or $D(y)$ contains a non-smooth component or is not defined everywhere in $\mathbb{R}^p$ or $\mathbb{R}^n$.

3 Efficient Implementation for Factorized Data Matrix

Now we consider a specific case where the data matrix A in (3) has a factorized structure $A = UV$ where $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times p}$ with $d << \min\{n, p\}$. We can maintain the vectors $\bar{u}^{(t)} = U^T\bar{y}^{(t)}$ and $\bar{v}^{(t)} = V\bar{x}^{(t)}$ and update them in $O(dm)$ and $O(dq)$ time, respectively, in each iteration. Then, we can obtain $\langle A_i, \bar{x}^{(t)}\rangle$ in (4) in $O(dm)$ time by evaluating $\langle U_i, \bar{v}^{(t)}\rangle$ for each $i \in I$, where U_i is the ith row of U. Similarly, we can obtain $\langle A_j, \bar{y}^{(t+1)}\rangle$ in (6) in $O(dq)$ time by taking $\langle V^j, \bar{v}^{(t)}\rangle$ for each $j \in J$, where V^j is the jth column of V. This leads to an efficient implementation of DSPDC whose per-iteration cost is $O(dm + dq)$, lower than the $O(mp)$ cost when A is not factorized. The detailed procedure is shown in Algorithm 2. The similar efficient implementation can be also applied to other coordinate methods such as SPDC, SDCA and ASDCA to obtain a lower computation cost in each iteration. To make a clear comparison between DSPDC and other coordinate methods when applied to factorized data, we summarize their numbers of iterations and per-iteration costs in Table 1. Numerical experiments will be presented in Section 4.

4 Numerical Experiments

In this section, we conduct numerical experiments to compare the DSPDC method with other two methods, SPCD [26] and SDCA [22] over several real datasets[2] Covtype, RCV1 and Real-sim. We consider the setting of sparse recovery problem after applying randomized feature reduction to binary classification. In particular, let $X \in \mathbb{R}^{n\times p}$ be the original training data, and $G \in \mathbb{R}^{d\times p}$ a Gaussian random matrix. So now $A = UV$ with $U = XG^T, V = G$. The problem of interest is

$$\min_{x\in\mathbb{R}^p}\max_{y\in\mathbb{R}^n}\left\{\frac{\lambda_2}{2}\|x\|_2^2+\lambda_1\|x\|_1+\frac{1}{n}y^TXG^TGx-\frac{1}{n}\sum_{i=1}^n\phi_i^*(y_i)\right\} \tag{17}$$

[1] If $\frac{n}{m} \leq \frac{p}{q}$, we can apply the dual version of DSPDC by switch the updating schemes for x and y.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

Algorithm 2 Efficient Implementation of Algorithm 1 for Factorized Data ($A = UV$)

Input: $x^{(-1)} = x^{(0)} = \bar{x}^{(0)} \in \mathbb{R}^p$, $y^{(-1)} = y^{(0)} = \bar{y}^{(0)} \in \mathbb{R}^n$, and positive control parameters (θ, τ, σ)

Initialize: $u^{(0)} = U^T y^{(0)}$, $v^{(0)} = Vx^{(0)}$, $\bar{u}^{(0)} = U^T \bar{y}^{(0)}$, $\bar{v}^{(0)} = V\bar{x}^{(0)}$

Iterate:

For $t = 0, 1, 2, \ldots, T-1$

Uniformly and randomly choose two sets of indices $I \subset \{1, 2, \ldots, n\}$ and $J \subset \{1, 2, \ldots, p\}$ of sizes m and q, respectively.

$$y_i^{(t+1)} = \begin{cases} \arg\max_{\beta \in \mathbb{R}} \left\{ \frac{1}{n} \langle U_i, \bar{v}^{(t)} \rangle \beta - \frac{\phi_i^*(\beta)}{n} - \frac{1}{2\sigma}(\beta - y_i^{(t)})^2 \right\} & \text{if } i \in I, \\ y_i^{(t)} & \text{if } i \notin I, \end{cases} \tag{11}$$

$$u^{(t+1)} = u^{(t)} + U^T(y^{(t+1)} - y^{(t)}), \tag{12}$$

$$\bar{u}^{(t+1)} = u^{(t)} + \frac{n}{m} U^T(y^{(t+1)} - y^{(t)}), \tag{13}$$

$$x_j^{(t+1)} = \begin{cases} \arg\min_{\alpha \in \mathbb{R}} \left\{ \frac{1}{n} \langle V^j, \bar{u}^{(t+1)} \rangle \alpha + g_i(\alpha) + \frac{1}{2\tau}(\alpha - x_i^{(t)})^2 \right\} & \text{if } j \in J, \\ x_j^{(t)} & \text{if } j \notin J, \end{cases} \tag{14}$$

$$v^{(t+1)} = v^{(t)} + V(x^{(t+1)} - x^{(t)}), \tag{15}$$

$$\bar{v}^{(t+1)} = v^{(t)} + (\theta + 1)V(x^{(t+1)} - x^{(t)}). \tag{16}$$

Output: $x^{(T)}$ and $y^{(T)}$

Algorithm	Number of Iterations	Per-Iteration Cost	Overall Complexity when $m = q = 1$
DSPDC	$\left(\frac{n}{m} + \sqrt{\frac{\kappa n}{m}\frac{p}{q}}\right) \log(\frac{1}{\epsilon})$	$qd + md$	$(nd + \sqrt{\kappa n} pd) \log(\frac{1}{\epsilon})$
SPDC	$\left(\frac{n}{m} + \sqrt{\frac{\kappa n}{m}}\right) \log(\frac{1}{\epsilon})$	$pd + md$	$(npd + \sqrt{\kappa n} pd) \log(\frac{1}{\epsilon})$
SDCA	$(n + \kappa) \log(\frac{1}{\epsilon})$	pd	$(npd + \kappa pd) \log(\frac{1}{\epsilon})$
ASDCA	$(n + \sqrt{\kappa n}) \log(\frac{1}{\epsilon})$	pd	$(npd + \sqrt{\kappa n} pd) \log(\frac{1}{\epsilon})$

Table 1: The complexity to find an ϵ-optimal solution when $A = UV$ and $\frac{n}{m} \geq \frac{p}{q}$.

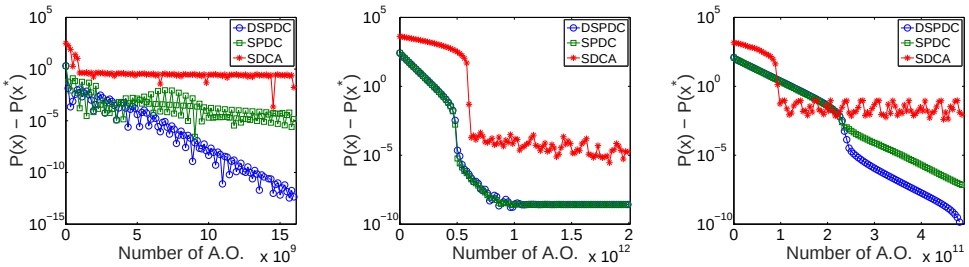

Figure 1: Left: Covtype ($n = 581012, p = 54$). Middle: RCV1 ($n = 20242, p = 47236$). Right: Real-sim ($n = 72309, p = 20958$).

We consider problem (17) with smoothed hinge loss

$$\phi_i(z) = \begin{cases} 0 & \text{if } b_i z \geq 1 \\ \frac{1}{2} - b_i z & \text{if } b_i z \leq 0 \\ \frac{1}{2}(1 - b_i z)^2 & \text{otherwise} \end{cases}, \tag{18}$$

where $b_i \in \{1, -1\}$ is the class label for the ith instance. In all experiments, we choose $d = 20$ and set $\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-2}$ in (17). Since these three sets data are real data, their sizes and dimensions are not in whole thousands. We choose m and q so that n and p can be either dividable by them or has a small division remainder. The numerical performances of the three methods are showed in Figure 1 with the values of m and q stated below. In these three examples, SPDC and DSPDC both outperform SDCA significantly. DSPDC has as similar performance to SPDC on RCV1 Real-sim but has a better performance than SPDC when applied to Covtype. The complementary results could be found in the full version manuscript [25].

References

[1] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[2] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. Technical report, CMAP, Ecole Polytechnique, CNRS, 2014.

[3] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[4] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[5] C. Dang and G. Lan. Randomized first-order methods for saddle point optimization. Technical report, University of Florida, 2014.

[6] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

[7] O. Fercoq and P. Richtrik. Accelerated, parallel and proximal coordinate descent. *CoRR*, abs/1312.5799, 2013.

[8] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323. 2013.

[9] J. Konečný, J. Liu, P. Richtárik, and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. Technical report, the School of Mathematics, University of Edinburg, 2014.

[10] G. Korpelevic. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematcheskie Metody*, 12:747–756, 1976.

[11] G. Lan. An optimal randomized incremental gradient method. Technical report, Department of Industrial and Systems Engineering, University of Florida, 2015.

[12] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of tseng's modified f-b splitting and korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM J. on Optimization*, 21(4):1688–1720.

[13] A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle-point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[14] Y. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Jorunal on Optimization*, 16(1):235–249, 2005.

[15] Y. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

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

[17] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Neural Information Processing Systems (NIPS)*, 2014.

[18] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

[19] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Technical Report HAL 00860051, INRIA, Paris, France, 2013.

[20] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1 regularized loss minimization. In *International Conference on Machine Learning (ICML)*, volume 382, 2009.

[21] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *NIPS*, pages 378–385. 2013.

[22] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

[23] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. In *ICML*, pages 567–599, 2013.

[24] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. arXiv:1403.4699.

[25] A. W. Yu, Q. Lin, and T. Yang. Doubly stochastic primal-dual coordinate method for regularized empirical risk minimization with factorized data. *CoRR*, abs/1508.03390, 2015.

[26] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML*, 2015.