Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers

Taiji Suzuki
Department of Mathematical and Computing Sciences
Tokyo Institute of Technology
Tokyo 152-8552, Japan
s-taiji@is.titech.ac.jp

Abstract

We propose a new stochastic dual coordinate ascent technique that can be applied to a wide range of regularized learning problems. Our method is based on Alternating Direction Method of Multipliers (ADMM) to deal with complex regularization functions such as structured regularizations. Our method can naturally afford mini-batch update and it gives speed up of convergence. We show that, under mild assumptions, our method converges exponentially. The numerical experiments show that our method actually performs efficiently.

1 Introduction

This paper proposes a new stochastic optimization method that shows exponential convergence and can be applied to wide range of regularization functions using the techniques of stochastic dual coordinate ascent with alternating direction method of multipliers. Recently, a lot of stochastic methods have been proposed. Among them, online stochastic optimization is the most basic and successful one. The convergence rate of such a method is $O(1/\sqrt{T})$ for general settings and $O(1/T)$ for strongly convex losses, which are minimax optimal [8]. On the other hand, recently it was shown that, if the sample size is fixed and it is allowed to reuse the observed data, it is possible to develop a stochastic method with exponential convergence rate for a strongly convex objective (Stochastic Average Gradient (SAG) [7], Stochastic Dual Coordinate Ascent (SDCA) [16, 17]). These methods are still stochastic in a sense that one sample or small mini-batch is randomly picked up to be used for each update. However, these methods have some drawbacks. The exponential convergence of SAG is guaranteed only for smooth loss and regularization functions. SDCA method can be applied only to a simple regularization function for which the dual function is easily computed, thus it is hard to apply the method to a complex regularization function such as structured regularization.

In this paper, we propose Stochastic Dual Coordinate Ascent method for Alternating Direction Method of Multipliers (SDCA-ADMM). Our method is similar to SDCA, but inherits a favorable property of ADMM. By combining SDCA and ADMM, our method can be applied to a wide range of regularized learning problems which are given by the following optimization problem:

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} f_i(z_i^\top w) + \psi(B^\top w), \tag{1}$$

where we assume that the *proximal operation* with respect to ψ is easily computed. Here the proximal operation is defined by $\mathrm{prox}(q|\psi) := \arg\min_u \left\{\frac{1}{2}\|q-u\|^2 + \psi(u)\right\}$. This formulation is quite flexible and fit wide range of applications such as structured regularization, dictionary learning, convex tensor decomposition and so on [11, 6, 21, 12]. The purpose of this paper is to give an efficient stochastic optimization method to solve this problem (1). For this purpose, we employ the *dual formulation*. Using the *Fenchel's duality theorem*, we have the following dual formulation.

Lemma 1.

$$\min_{w\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^n f_i(z_i^\top w)+\psi(B^\top w)=-\min_{x\in\mathbb{R}^n,y\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n f_i^*(x_i)+\psi^*\big(\frac{y}{n}\big)\mid Zx+By=0\right\},$$

where f_i^* and ψ^* are the convex conjugates of f_i and ψ respectively [13], and $Z=[z_1,z_2,\ldots,z_n]\in\mathbb{R}^{p\times n}$. Moreover w^*, x^* and y^* are optimal solutions of both sides if and only if $z_i^\top w^*\in\nabla f_i^*(x_i^*)$, $\frac{1}{n}y^*\in\nabla\psi(u)|_{u=B^\top w^*}$, $Zx^*+By^*=0$.*

2 Proposed Method: Stochastic Dual Coordinate Ascent with ADMM

In this section, we present our proposal, stochastic dual coordinate ascent type ADMM. For a positive semidefinite matrix S, we denote by $\|x\|_S:=\sqrt{x^\top Sx}$. Z_I denotes the matrix consisting of columns included in the index set I ($Z_I=[Z_{i_1},\ldots,Z_{i_{|I|}}]$), and $Z_{\backslash I}$ is a matrix obtained by subtracting columns with indexes in I from Z. Similarly, for a vector x, x_I is a vector consisting of components with indexes $i\in I$, $x_I=(x_i)_{i\in I}$, and $x_{\backslash I}$ is a vector obtained by subtracting the components in I from x.

At each iteration, we randomly choose an index set $I\subseteq\{1,\ldots,n\}$ so that each index i is included in I with probability $1/K$; $P(i\in I)=1/K$ for all $i=1,\ldots,n$. To do so, we suggest the following procedure. We split the index set $\{1,\ldots,n\}$ into K groups $\{I_1,I_2,\ldots,I_K\}$ beforehand, and then pick up uniformly $k\in\{1,\ldots,K\}$ and set $I=I_k$ for each iteration. Each sub-batch I_k can have different cardinality from others, but the probability $P(i\in I)$ should be uniform for all $i=1,\ldots,n$. The update rule of our method is given as follows:

$$\begin{aligned}
y^{(t)}&\leftarrow\arg\min_y\Big\{n\psi^*\Big(\frac{y}{n}\Big)-\langle w^{(t-1)},Zx^{(t-1)}+By\rangle+\frac{\rho}{2}\|Zx^{(t-1)}+By\|^2+\frac{1}{2}\|y-y^{(t-1)}\|_Q^2\Big\},\\
x_I^{(t)}&\leftarrow\arg\min_{x_I}\Big\{\sum_{i\in I}f_i^*(x_i)-\langle w^{(t-1)},Z_Ix_I+By^{(t)}\rangle+\frac{\rho}{2}\|Z_Ix_I+Z_{\backslash I}x_{\backslash I}^{(t-1)}+By^{(t)}\|^2\\
&\qquad\qquad+\frac{1}{2}\|x_I-x_I^{(t-1)}\|_{G_{I,I}}^2\Big\},\\
w^{(t)}&\leftarrow w^{(t-1)}-\gamma\rho\{n(Zx^{(t)}+By^{(t)})-(n-n/K)(Zx^{(t-1)}+By^{(t-1)})\},
\end{aligned}$$

where $\rho,\gamma>0$ are given parameters, and Q,G are positive definite matrices. The optimization procedure looks a bit complicated. To simplify the procedure, we set Q and G as

$$Q=\rho(\eta_B\mathrm{I}_d-B^\top B),\ \ G_{I,I}=\rho(\eta_{Z,I}\mathrm{I}_{|I|}-Z_I^\top Z_I),\tag{3}$$

where η_B and $\eta_{Z,I}$ are chosen so that $\eta_B\geq\|B^\top B\|$ and $\eta_{Z,I}\geq\|Z_I^\top Z_I\|$. We define $q^{(t)}=y^{(t-1)}+\frac{B^\top}{\rho\eta_B}\{w^{(t-1)}-\rho(Zx^{(t-1)}+By^{(t-1)})\}$. Then, by the relation $\mathrm{prox}(q|\psi)+\mathrm{prox}(q|\psi^*)=q$ (Theorem 31.5 of [13]) and simple calculations, the update rule of $y^{(t)}$ and $x^{(t)}$ is rewritten as

$$y^{(t)}\leftarrow q^{(t)}-\mathrm{prox}(q^{(t)}|n\psi(\rho\eta_B\ \cdot\ )/(\rho\eta_B)),\tag{4}$$

$$x_I^{(t)}\leftarrow\mathrm{prox}\Big(x_I^{(t-1)}+\frac{Z_I^\top}{\rho\eta_{Z,I}}\{w^{(t-1)}-\rho(Zx^{(t-1)}+By^{(t)})\}\ \Big|\ \frac{\sum_{i\in I}f_i^*}{\rho\eta_{Z,I}}\Big).\tag{5}$$

Note that, since $\sum_{i\in I}f_i^*(x_i)$ is sum of single variable convex functions $f_i^*(x_i)$, the proximal operation in Eq. (5) can be split into the proximal operation with respect to each single variable x_i. This is advantageous for not only the simpleness of the computation but also parallel computation. In summary, our proposed algorithm is given in Algorithm 1.

3 Linear Convergence of SDCA-ADMM

In this section, the convergence rate of our proposed algorithm is given. Indeed, the convergence rate is exponential (R-linear). To show the convergence rate, we assume some conditions. First, we assume that there exits an unique optimal solution w^* and $B^\top$ is injective (B is not necessarily injective). Moreover, we assume the uniqueness of the dual solution x^*, but don't assume the uniqueness of y^*. We denote by the set of dual optimum of y as $\mathcal{Y}^*$ and assume that $\mathcal{Y}^*$ is compact. Then, by Lemma 1, we have that

$$z_i^\top w^*\in\nabla f_i^*(x_i^*),\ \ y^*/n\in\nabla\psi(u)|_{u=B^\top w^*}.\tag{6}$$

*The convex conjugate function f^* of f is defined by $f^*(y):=\sup_x\{x^\top y-f(x)\}$.

Algorithm 1 SDCA-ADMM

Input: $\rho, \eta > 0$
Initialize $x_0 = \mathbf{0}$, $y_0 = \mathbf{0}$, $w_0 = \mathbf{0}$ and $\{I_1, \ldots, I_K\}$.
for $t = 1$ **to** T **do**
Choose $k \in \{1, \ldots, K\}$ uniformly at random, set $I = I_k$, and observe the training samples $\{(x_i, y_i)\}_{i \in I}$.
Set $q^{(t)} = y^{(t-1)} + \frac{B^\top}{\rho \eta_B}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t-1)})\}$.
Update $y^{(t)} \leftarrow q^{(t)} - \mathrm{prox}(q^{(t)} | n\psi(\rho\eta_B \ \cdot \ )/(\rho\eta_B))$
Update $x_I^{(t)} \leftarrow \mathrm{prox}\Big(x_I^{(t-1)} + \frac{Z_I^\top}{\rho\eta_{Z,I}}\{w^{(t-1)} - \rho(Zx^{(t-1)} + By^{(t)})\} \ \Big| \ \frac{\sum_{i \in I} f_i^*}{\rho \eta_{Z,I}}\Big)$.
Update $w^{(t)} \leftarrow w^{(t-1)} - \gamma\rho\{n(Zx^{(t)} + By^{(t)}) - (n - n/K)(Zx^{(t-1)} + By^{(t-1)})\}$.
end for
output $w^{(T)}$.

By the convex duality arguments, this implies that $x_i^* \in \nabla f_i(u)|_{u = z_i^\top w^*}$, $B^\top w^* \in \nabla \psi^*(u)|_{u = y^*/n}$. Moreover, we assume the following condition.

Assumption 1. *There exits $v > 0$ such that, $\forall x_i \in \mathbb{R}$,*

$$f_i^*(x_i) - f_i^*(x_i^*) \geq \langle \nabla f_i^*(x_i^*), x_i - x_i^* \rangle + \frac{v\|x_i - x_i^*\|^2}{2}.$$

There exit $h > 0$ and $v_\psi > 0$ such that, for all y, w, there exists $y^* \in \mathcal{Y}^*$ such that

$$\psi^*(y/n) - \psi^*(y^*/n) \geq \langle B^\top w^*, y/n - y^*/n \rangle + \frac{v_\psi}{2}\|P_{\mathrm{Ker}(B)}(y/n - y^*/n)\|^2, \tag{7}$$

$$\psi(u) - \psi(B^\top w^*) \geq \langle y^*/n, u - B^\top w^* \rangle + \frac{h}{2}\|u - B^\top w^*\|^2, \tag{8}$$

where $P_{\mathrm{Ker}(B)}$ is the projection matrix to the kernel of B.

Note that these conditions should be satisfied only around the optimal solutions (x^*, y^*) and w^*. It does not need to hold for every point, thus is much weaker than the ordinary strong convexity. The condition (7) is satisfied, for example, by ℓ_1-regularization. The quadratic term in the right hand side of the condition (7) is restricted on $\mathrm{Ker}(B)$. Thus, if $B = \mathrm{I}_p$, this condition is always satisfied. The assumption (8) is the strongest one, but this is satisfied by just adding a small quadratic function.

Define the dual objective as $F_D(x, y) := \frac{1}{n}\sum_{i=1}^n f_i^*(x_i) + \psi^*(\frac{y}{n}) - \langle w^*, Z\frac{x}{n} - B\frac{y}{n} \rangle$. Note that, by Eq. (6), $F_D(x, y) - F_D(x^*, y^*)$ is always non-negative. Define the block diagonal matrix H as $H_{I,I} = \rho Z_I^\top Z_I + G_{I,I}$ for all $I \in \{I_1, \ldots, I_K\}$ and $H_{i,j} = 0$ for $(i, j) \notin I_k \times I_k$ $(\forall k)$. We define $R_D(x, y, w)$ as $R_D(x, y, w) := F_D(x, y) - F_D(x^*, y^*) + \frac{1}{2n^2\gamma\rho}\|w - w^*\|^2 + \frac{\rho(1-\gamma)}{2n}\|Zx + By\|^2 + \frac{1}{2n}\|x - x^*\|^2_{v\mathrm{I}_p + H} + \frac{1}{2nK}\|y - y^*\|_Q^2$. For a symmetric matrix S, we define $\sigma_{\max}(S)$ and $\sigma_{\min}(S)$ as the maximum and minimum singular value respectively.

Theorem 2. *Suppose that $\gamma = \frac{1}{4n}$, $\eta_{Z,I} > (1 + 2\gamma n(1 - 1/K))\sigma_{\max}(Z_I^\top Z_I)$ for all $I \in \{I_1, \ldots, I_K\}$ and $B^\top$ is injective. Then, under Assumption 1, the dual objective function converges R-linearly: For $C_1 = R_D(x^{(0)}, y^{(0)}, w^{(0)})$, we have that $\mathrm{E}[R_D(x^{(T)}, y^{(T)}, w^{(T)})] \leq \left(1 - \frac{\mu}{K}\right)^T C_1$ where $\mu = \min\Big\{\frac{v}{2(v + \sigma_{\max}(H))}, \frac{h\rho\sigma_{\min}(B^\top B)}{2\max\{1/n, 4h\rho, 4h\sigma_{\max(Q)}\}}, \frac{v_\psi}{4\sigma_{\max}(Q)}, \frac{nv\sigma_{\min}(BB^\top)}{\sigma_{\max(Q)}(\rho\sigma_{\max}(Z^\top Z) + 4v)}\Big\}$. In particular, we have that*

$$\mathrm{E}[\|w^{(T)} - w^*\|^2] \leq \frac{n\rho}{2}\left(1 - \frac{\mu}{K}\right)^T C_1.$$

4 Numerical Experiments

In this section, we give numerical experiments to demonstrate the effectiveness of our proposed algorithm on artificial and real data. We compare our SDCA-ADMM with the existing online stochastic optimization methods such as Regularized Dual Averaging (RDA) [3, 23], Online ADMM (OL-ADMM) [22], Online Proximal Gradient descent ADMM (OPG-ADMM) [9, 19] and RDA-ADMM [19]. We also compared our method with batch ADMM (Batch-ADMM) in the artificial data sets. We used sub-batch with size 50 for all the methods including ours. We employed the *smoothed hinge loss* [16, 17] for which the proximal operation is analytically computed.

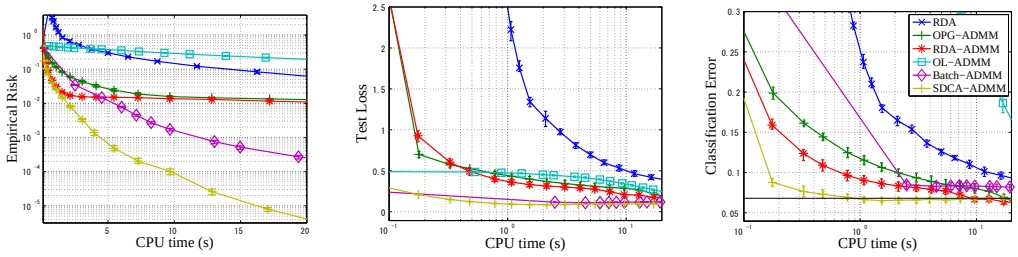

Figure 1: Excess empirical risk, exected loss on the test data and test classification error averaged over 10 independent iteration against CPU time in artificial data.

The experiment on artificial data is a classification problem with overlapped group regularization as performed in [19] where $d = 32 \times 32 = 1024$ and $n = 5120$. The overlapped group regularization is given as $\tilde{\psi}(x) = C(\sum_{i=1}^{32} \|X_{:,i}\| + \sum_{j=1}^{32} \|X_{j,:}\| + 0.01 \times \sum_{i,j} X_{i,j}^2/2)$ where X is the 32×32 matrix obtained by reshaping x. Next, we execute numerical experiments on real data '20 Newsgroups'†. '20 Newsgroups' contains 100 dimensional 12,995 training samples and 3,247 test samples. We constructed a similarity graph between features using graph Lasso as in [9]. Then we imposed the following graph guided regularization: $\tilde{\psi}(w) = C_1 \sum_{i=1}^{p} |w_i| + C_2 \sum_{(i,j)\in E} |w_i - w_j| + 0.01 \times (C_1 \sum_{i=1}^{p} |w_i|^2 + C_2 \sum_{(i,j)\in E} |w_i - w_j|^2)$ where E is the set of edges obtained from the similarity matrix.

We measured the excess empirical risk ($F_P(w^{(t)}) - \min_w F_P(w)$) (or the empirical risk ($F_P(w^{(t)})$) for the real data), the expected loss on the test data ($\mathrm{E}_{(z,y)}[f(y, z^\top w^{(t)})]$), and the classification error ($\mathrm{E}_{(z,y)}[1\{y \neq \mathrm{sign}(z^\top w^{(t)})\}$). Figures 1 and 2 show these three values against CPU time.

We observe that the excess empirical risk of our method, SDCA-ADMM, actually converges linearly while other stochastic methods don't show linear convergence. Although Batch-ADMM also shows linear convergence, SDCA-ADMM is much faster than Batch-ADMM. As for the classification error, existing methods also show nice performances. On the other hand, SDCA-ADMM rapidly converges to a stable state and shows comparable or better classification accuracy than existing methods.

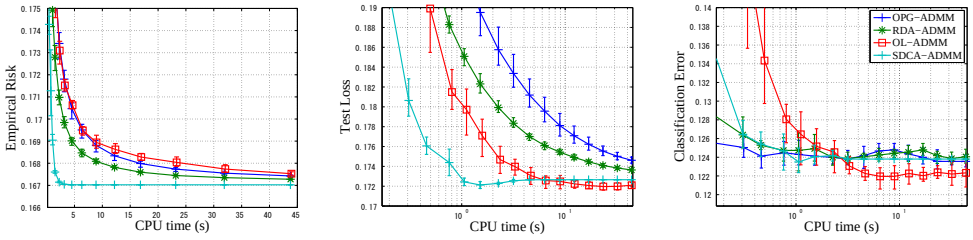

Figure 2: Empirical risk, average loss on the test data and test classification error averaged over 5 independent iteration against CPU time in real data.

5 Conclusion

We proposed a new stochastic dual coordinate ascent technique with alternating direction multiplier method. We have shown theoretically and numerically that our method converges exponentially under some conditions. According to our analysis, the mini-batch method improves the convergence rate.

Acknowledgement This work is partially supported by JSPS KAKENHI 25730013, and the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

†Available at http://www.cs.nyu.edu/~roweis/data.html. We converted the four class classification task into binary classification by grouping category 1,2 and category 3,4 respectively.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010.

[2] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

[3] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2873–2908, 2009.

[4] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers & Mathematics with Applications*, 2:17–40, 1976.

[5] M. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory & Applications*, 4:303–320, 1969.

[6] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[7] N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems 25*, 2013.

[8] A. Nemirovskii and D. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, New York, 1983.

[9] H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[10] M. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, London, New York, 1969.

[11] Z. Qin and D. Goldfarb. Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, 13:1435–1468, 2012.

[12] A. Rakotomamonjy. Applying alternating direction method of multipliers for constrained dictionary learning. *Neurocomputing*, 106:126–136, 2013.

[13] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[14] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:97–116, 1976.

[15] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems 26*, 2013.

[16] S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. Technical report, 2013. arXiv:1211.2717.

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

[18] M. Signoretto, L. D. Lathauwer, and J. Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186, ESAT-SISTA, K.U.Leuven, 2010.

[19] T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[20] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro. Mini-batch primal and dual methods for SVMs. In *the 30th International Conference on Machine Learning*, 2013.

[21] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems 25*, 2011.

[22] H. Wang and A. Banerjee. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[23] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems 23*, 2009.

[24] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems 24*, 2011.