Adaptive Stochastic Alternating Direction Method of Multipliers

Peilin Zhao[*,‡] ZHAOP@I2R.A-STAR.EDU.SG
Jinwei Yang[†] JYANG7@ND.EDU
Tong Zhang[‡] TZHANG@STAT.RUTGERS.EDU
Ping Li[§] PINGLI@STAT.RUTGERS.EDU

[*]Data Analytics Department, Institute for Infocomm Research, A*STAR, Singapore
[†]Department of Mathematics, Rutgers University; and Department of Mathematics, University of Notre Dame, USA
[‡]Department of Statistics & Biostatistics, Rutgers University, USA; and Big Data Lab, Baidu Research, China
[§]Department of Statistics & Biostatistics, Department of Computer Science, Rutgers University; and Baidu Research, USA

Abstract

The *Alternating Direction Method of Multipliers (ADMM)* has been studied for years. Traditional ADMM algorithms need to compute, at each iteration, an (empirical) expected loss function on all training examples, resulting in a computational complexity proportional to the number of training examples. To reduce the complexity, stochastic ADMM algorithms were proposed to replace the expected loss function with a random loss function associated with one uniformly drawn example plus a Bregman divergence term. The Bregman divergence, however, is derived from a simple 2nd-order proximal function, i.e., the half squared norm, which could be a suboptimal choice.

In this paper, we present a new family of stochastic ADMM algorithms with optimal 2nd-order proximal functions, which produce a new family of adaptive stochastic ADMM methods. We theoretically prove that the regret bounds are as good as the bounds which could be achieved by the best proximal function that can be chosen in hindsight. Encouraging empirical results on a variety of real-world datasets confirm the effectiveness and efficiency of the proposed algorithms.

1. Introduction

Originally introduced in (Glowinski & Marroco, 1975; Gabay & Mercier, 1976), the offline/batch Alternating Direction Method of Multipliers (ADMM) stemmed from the augmented Lagrangian method, with its global convergence property established in (Gabay, 1983; Glowinski & Le Tallec, 1989; Eckstein & Bertsekas, 1992). Recent studies have shown that ADMM achieves a convergence rate of $O(1/T)$ (Monteiro & Svaiter, 2013; He & Yuan, 2012) (where T is number of iterations of ADMM), when the objective function is generally convex. Furthermore, ADMM enjoys a convergence rate of $O(\alpha^T)$, for some $\alpha \in (0,1)$, when the objective function is strongly convex and smooth (Luo, 2012; Deng & Yin, 2012). ADMM has shown attractive performance in a wide range of real-world problems such as compressed sensing (Yang & Zhang, 2011), image restoration (Goldstein & Osher, 2009), video processing, and matrix completion (Goldfarb et al., 2013), etc.

From the computational perspective, one drawback of ADMM is that, at every iteration, the method needs to compute an (empirical) expected loss function on all the training examples. The computational complexity is propositional to the number of training examples, which makes the original ADMM unsuitable for solving large-scale learning and big data mining problems. The online ADMM (OADMM) algorithm (Wang & Banerjee, 2012) was proposed to tackle the computational challenge. For OADMM, the objective function is replaced with an online function at every step, which only depends on a single training example. OADMM can achieve an average regret bound of $O(1/\sqrt{T})$ for convex objective functions and $O(\log(T)/T)$ for strongly convex objective functions. Interestingly, although the optimization of the loss function is assumed to be easy in the analysis of (Wang & Banerjee, 2012), this step is actually not necessarily easy in practice. To address this issue, the stochastic ADMM algorithm was proposed, by linearizing the online loss function (Ouyang et al., 2013; Suzuki, 2013). In stochastic ADMM algorithms, the online loss function is firstly uniformly drawn from all the loss functions associated with all the training examples. Then the

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

loss function is replaced with its first order expansion at the current solution plus the Bregman divergence from the current solution. The Bregman divergence is based on a simple proximal function, the half squared norm, so that the Bregman divergence is the half squared distance. In this way, the optimization of the loss function enjoys a closed-form solution. The stochastic ADMM achieves similar convergence rates as OADMM. Using the half square norm as proximal function, however, may be a suboptimal choice. Our paper will address this issue. We should mention that there are other strategies recently adopted to accelerate stochastic ADMM, including stochastic average gradient (Zhong & Kwok, 2014), dual coordinate ascent (Suzuki, 2014), which are still based on half squared distance.

Our contribution. In the previous work (Ouyang et al., 2013; Suzuki, 2013) the Bregman divergence is derived from a simple second order function, i.e., the half squared norm, which could be a suboptimal choice (Duchi et al., 2011). In this paper, we present a new family of stochastic ADMM algorithms with adaptive proximal functions, which can accelerate stochastic ADMM by using adaptive regularization. We theoretically prove that the regret bounds of our methods are as good as those achieved by stochastic ADMM with the best proximal function that can be chosen in hindsight. The effectiveness and efficiency of the proposed algorithms are confirmed by encouraging empirical evaluations on several real-world datasets.

2. Adaptive Stochastic Alternating Direction Method of Multipliers

2.1. Problem Formulation

In this paper, we will study a family of convex optimization problems, where our objective functions are composite. Specially, we are interested in the following equality-constrained optimization task:

$$\min_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in\mathcal{V}} f((\mathbf{w}^\top,\mathbf{v}^\top)^\top) := \mathbb{E}_\xi \ell(\mathbf{w},\xi)+\varphi(\mathbf{v}), \quad \text{s.t. } A\mathbf{w}+B\mathbf{v}=\mathbf{b}, \tag{1}$$

where $\mathbf{w}\in\mathbb{R}^{d_1}$, $\mathbf{v}\in\mathbb{R}^{d_2}$, $A\in\mathbb{R}^{m\times d_1}$, $B\in\mathbb{R}^{m\times d_2}$, $\mathbf{b}\in\mathbb{R}^m$, $\mathcal{W}$ and $\mathcal{V}$ are convex sets. For simplicity, the notation ℓ is used for both the instance function value $\ell(\mathbf{w},\xi)$ and its expectation $\ell(\mathbf{w})=\mathbb{E}_\xi\ell(\mathbf{w},\xi)$. It is assumed that a sequence of identical and independent (i.i.d.) observations can be drawn from the random vector ξ, which satisfies a fixed but unknown distribution. When ξ is deterministic, the above optimization becomes the traditional formulation of ADMM (Boyd et al., 2011). In this paper, we will assume the functions ℓ and φ are convex but not necessarily continuously differentiable. In addition, we denote the optimal solution of (1) as $(\mathbf{w}_*^\top,\mathbf{v}_*^\top)^\top$.

Before presenting the proposed algorithm, we first introduce some notations. For a positive definite matrix $G\in\mathbb{R}^{d_1\times d_1}$, we define the G-norm of a vector $\mathbf{w}$ as $\|\mathbf{w}\|_G := \sqrt{\mathbf{w}^\top G\mathbf{w}}$. When there is no ambiguity, we often use $\|\cdot\|$ to denote the Euclidean norm $\|\cdot\|_2$. We use $\langle\cdot,\cdot\rangle$ to denote the inner product in a finite dimensional Euclidean space. Let H_t be a positive definite matrix for $t\in\mathbb{N}$. Set the proximal function $\phi_t(\cdot)$, as $\phi_t(\mathbf{w})=\frac{1}{2}\|\mathbf{w}\|_{H_t}^2=\frac{1}{2}\langle\mathbf{w},H_t\mathbf{w}\rangle$. Then the corresponding *Bregman divergence* for $\phi_t(\mathbf{w})$ is defined as

$$\begin{aligned}\mathcal{B}_{\phi_t}(\mathbf{w},\mathbf{u}) &= \phi_t(\mathbf{w})-\phi_t(\mathbf{u})-\langle\nabla\phi_t(\mathbf{u}),\mathbf{w}-\mathbf{u}\rangle\\ &= \frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_{H_t}^2.\end{aligned}$$

2.2. Algorithm

To solve the problem (1), a popular method is Alternating Direction Method of Multipliers (ADMM). ADMM splits the optimization with respect to $\mathbf{w}$ and $\mathbf{v}$ by minimizing the augmented Lagrangian:

$$\min_{\mathbf{w},\mathbf{v}} \mathcal{L}_\beta(\mathbf{w},\mathbf{v},\theta) := \ell(\mathbf{w})+\varphi(\mathbf{v})-\langle\theta,A\mathbf{w}+B\mathbf{v}-\mathbf{b}\rangle + \frac{\beta}{2}\|A\mathbf{w}+B\mathbf{v}-\mathbf{b}\|^2,$$

where $\beta>0$ is a pre-defined penalty. Specifically, the ADMM algorithm minimizes $\mathcal{L}_\beta$ as follows

$$\begin{aligned}\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} \mathcal{L}_\beta(\mathbf{w},\mathbf{v}_t,\theta_t),\\ \mathbf{v}_{t+1} &= \arg\min_{\mathbf{v}} \mathcal{L}_\beta(\mathbf{w}_{t+1},\mathbf{v},\theta_t),\\ \theta_{t+1} &= \theta_t-\beta(A\mathbf{w}_{t+1}+B\mathbf{v}_{t+1}-\mathbf{b}).\end{aligned}$$

At each step, however, ADMM requires calculating the expectation $\mathbb{E}_\xi\ell(\mathbf{w},\xi)$, which may be computationally too expensive, since we may only have an unbiased estimate of $\ell(\mathbf{w})$ or the expectation $\mathbb{E}_\xi\ell(\mathbf{w},\xi)$ is an empirical one for big data problems. To solve this issue, we propose to minimize the following stochastic approximation:

$$\mathcal{L}_{\beta,t}(\mathbf{w},\mathbf{v},\theta) = \langle\mathbf{g}_t,\mathbf{w}\rangle+\varphi(\mathbf{v})-\langle\theta,A\mathbf{w}+B\mathbf{v}-\mathbf{b}\rangle + \frac{\beta}{2}\|A\mathbf{w}+B\mathbf{v}-\mathbf{b}\|^2+\frac{1}{\eta}\mathcal{B}_{\phi_t}(\mathbf{w},\mathbf{w}_t),$$

where $\mathbf{g}_t=\ell'(\mathbf{w}_t,\xi_t)$ and H_t for $\phi_t=\frac{1}{2}\|\mathbf{w}\|_{H_t}^2$ will be specified later. This objective linearizes $\ell(\mathbf{w},\xi_t)$ and adopts a dynamic Bregman divergence function to keep the new model near to the previous one. It is easy to see that this proposed approximation includes the one proposed by (Ouyang et al., 2013) as a special case when $H_t=I$. To minimize the above function, we followed the ADMM algorithm to optimize over $\mathbf{w}$, $\mathbf{v}$, θ sequentially, by fixing the others. In addition, we also need to update H_t for $\mathcal{B}_{\phi_t}$ at every step, which will be specified later. Finally the proposed Adaptive Stochastic Alternating Direction Method of Multipliers (**Ada-SADMM**) is summarized in Algorithm 1.

Algorithm 1 Adaptive Stochastic Alternating Direction Method of Multipliers (**Ada-SADMM**).

Initialize: $\mathbf{w}_1 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{0}$, $\theta_1 = \mathbf{0}$, $H_1 = aI$, and $a > 0$.
for $t = 1, 2, \ldots, T$ **do**
  Compute $\mathbf{g}_t = \ell'(\mathbf{w}_t, \xi_t)$;
  Update H_t and compute $\mathcal{B}_{\phi_t}$;
  $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_{\beta,t}(\mathbf{w}, \mathbf{v}_t, \theta_t)$;
  $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{L}_{\beta,t}(\mathbf{w}_{t+1}, \mathbf{v}, \theta_t)$;
  $\theta_{t+1} = \theta_t - \beta(A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - \mathbf{b})$;
end for

2.3. Analysis

This subsection is devoted to analyzing the expected convergence rate of the iterative solutions of the proposed algorithm for general H_t, $t = 1, \ldots, T$, where the proof techniques in (Ouyang et al., 2013) and (Duchi et al., 2011) are adopted. To accomplish this, we begin with a technical lemma, which will facilitate the later analysis.

Lemma 1. *Let $\ell(\mathbf{w}, \xi_t)$ and $\varphi(\mathbf{w})$ be convex functions and H_t be positive definite, for $t \geq 1$. For Algorithm 1, we have the following inequality*

$$\begin{aligned}&\ell(\mathbf{w}_t) + \varphi(\mathbf{v}_{t+1}) - \ell(\mathbf{w}) - \varphi(\mathbf{v}) + (\mathbf{z}_{t+1} - \mathbf{z})^\top F(\mathbf{z}_{t+1})\\ &\leq \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2} + \frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w})]\\ &\quad+\frac{\beta}{2}(\|A\mathbf{w} + B\mathbf{v}_t - \mathbf{b}\|^2 - \|A\mathbf{w} + B\mathbf{v}_{t+1} - \mathbf{b}\|^2)\\ &\quad+\langle\delta_t, \mathbf{w} - \mathbf{w}_t\rangle + \frac{1}{2\beta}(\|\theta - \theta_t\|^2 - \|\theta - \theta_{t+1}\|^2),\end{aligned}$$

where $\mathbf{z}_t = (\mathbf{w}_t^\top, \mathbf{v}_t^\top, \theta_t^\top)^\top$, $\mathbf{z} = (\mathbf{w}^\top, \mathbf{v}^\top, \theta^\top)^\top$, $\delta_t = \mathbf{g}_t - \ell'(\mathbf{w}_t)$, and $F(\mathbf{z}) = ((-A^\top\theta)^\top, (-B^\top\theta)^\top, (A\mathbf{w} + B\mathbf{v} - \mathbf{b})^\top)^\top$.

The proof is in the appendix. We now analyze the convergence behavior of Algorithm 1 and provide an upper bound on the objective value and the feasibility violation.

Theorem 1. *Let $\ell(\mathbf{w}, \xi_t)$ and $\varphi(\mathbf{w})$ be convex functions and H_t be positive definite, for $t \geq 1$. For Algorithm 1, we have the following inequality for any $T \geq 1$ and $\rho > 0$:*

$$\begin{aligned}&\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|] \leq\\ &\quad\frac{1}{2T}\Big\{\mathbb{E}\sum_{t=1}^{T}[\frac{2}{\eta}(\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*))\\ &\quad+\eta\|\mathbf{g}_t\|^2_{H_t^*}] + \beta D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{\beta}\Big\},\end{aligned}\tag{2}$$

where $\bar{\mathbf{u}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t^\top, \frac{1}{T}\sum_{t=2}^{T+1}\mathbf{v}_t^\top\right)^\top$, $\mathbf{u}_* = (\mathbf{w}_*^\top, \mathbf{v}_*^\top)^\top$, and $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = (\frac{1}{T}\sum_{t=2}^{T+1}\mathbf{w}_t, \frac{1}{T}\sum_{t=2}^{T+1}\mathbf{v}_t)$, and $D_{\mathbf{v}_*,B} = \|B\mathbf{v}_*\|$.

Proof. For convenience, we denote $\mathbf{u} = (\mathbf{w}^\top, \mathbf{v}^\top)^\top$, $\bar{\theta}_T = \frac{1}{T}\sum_{t=2}^{T+1}\theta_t$, and $\bar{\mathbf{z}}_T = (\bar{\mathbf{w}}_T^\top, \bar{\mathbf{v}}_T^\top, \bar{\theta}_T^\top)^\top$. With these notations, using convexity of $\ell(\mathbf{w})$ and $\varphi(\mathbf{v})$ and the monotonicity of operator $F(\cdot)$, we have for any $\mathbf{z}$:

$$\begin{aligned}&f(\bar{\mathbf{u}}_T) - f(\mathbf{u}) + (\bar{\mathbf{z}}_T - \mathbf{z})^\top F(\bar{\mathbf{z}}_T)\\ &\leq \frac{1}{T}\sum_{t=1}^{T}[f((\mathbf{w}_t^\top, \mathbf{v}_{t+1}^\top)^\top) - f(\mathbf{u}) + (\mathbf{z}_{t+1} - \mathbf{z})^\top F(\mathbf{z}_{t+1})]\\ &= \frac{1}{T}\sum_{t=1}^{T}[\ell(\mathbf{w}_t) + \varphi(\mathbf{v}_{t+1}) - \ell(\mathbf{w}) - \varphi(\mathbf{v}) + (\mathbf{z}_{t+1} - \mathbf{z})^\top F(\mathbf{z}_{t+1})].\end{aligned}$$

Combining this inequality with Lemma 1 at the optimal solution $(\mathbf{w}, \mathbf{v}) = (\mathbf{w}_*, \mathbf{v}_*)$, we can derive

$$\begin{aligned}&f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + (\bar{\mathbf{z}}_T - \mathbf{z}_*)^\top F(\bar{\mathbf{z}}_T)\\ &\leq \frac{1}{T}\sum_{t=1}^{T}\Big\{\frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*)] + \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2}\\ &+\langle\delta_t, \mathbf{w}_* - \mathbf{w}_t\rangle + \frac{\beta}{2}(\|A\mathbf{w}_* + B\mathbf{v}_t - \mathbf{b}\|^2\\ &-\|A\mathbf{w}_* + B\mathbf{v}_{t+1} - \mathbf{b}\|^2) + \frac{1}{2\beta}(\|\theta - \theta_t\|^2 - \|\theta - \theta_{t+1}\|^2)\Big\}\\ &\leq \frac{1}{T}\Big\{\sum_{t=1}^{T}\big[\frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*)] + \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2}\\ &\quad+\langle\delta_t, \mathbf{w}_* - \mathbf{w}_t\rangle\big] + \frac{\beta}{2}\|A\mathbf{w}_* + B\mathbf{v}_1 - \mathbf{b}\|^2 + \frac{1}{2\beta}\|\theta - \theta_1\|^2\Big\}\\ &\leq \frac{1}{T}\Big\{\sum_{t=0}^{T-1}\big[\frac{1}{\eta}(\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*)) + \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2}\\ &\quad+\langle\delta_t, \mathbf{w}_* - \mathbf{w}_t\rangle\big] + \frac{\beta}{2}D^2_{\mathbf{v}_*,B} + \frac{1}{2\beta}\|\theta - \theta_1\|^2\Big\}.\end{aligned}$$

As the above inequality is valid for any θ, it also holds in the ball $\mathrm{B}_\rho = \{\theta : \|\theta\| \leq \rho\}$. Combining with the fact that the optimal solution must also be feasible, it follows that

$$\begin{aligned}&\max_{\theta\in\mathrm{B}_\rho}\{f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + (\bar{\mathbf{z}}_T - \mathbf{z}_*)^\top F(\bar{\mathbf{z}}_T)\}\\ &= \max_{\theta\in\mathrm{B}_\rho}\{f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \bar{\theta}_T^\top(A\mathbf{w}_* + B\mathbf{v}_* - \mathbf{b})\\ &\qquad-\theta^\top(A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b})\}\\ &= \max_{\theta\in\mathrm{B}_\rho}\{f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) - \theta^\top(A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b})\}\\ &= f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|.\end{aligned}$$

Utilizing the above two inequalities, we obtain

$$\begin{aligned}&\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|]\\ &\leq \frac{1}{T}\mathbb{E}\Big\{\sum_{t=1}^{T}\big(\frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*)]\\ &+\frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2}) + \langle\delta_t, \mathbf{w}_* - \mathbf{w}_t\rangle\big) + \frac{\beta}{2}D^2_{\mathbf{v}_*,B} + \frac{1}{2\beta}\|\theta - \theta_1\|^2\Big\}\\ &\leq \frac{1}{2T}\Big\{\mathbb{E}\sum_{t=1}^{T}[\frac{2}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t, \mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1}, \mathbf{w}_*)]\\ &\quad+\eta\|\mathbf{g}_t\|^2_{H_t^*}] + \beta D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{\beta}\Big\},\end{aligned}$$

Note $\mathbb{E}\delta_t = 0$ in the last step. This completes the proof. □

The above theorem allows us to derive regret bounds for a family of algorithms that iteratively modify the proximal functions ϕ_t in attempt to lower the regret bounds. Since the rate of convergence is still dependent on H_t and η, next we are going to choose appropriate positive definite matrix H_t and the constant η to optimize the rate of convergence.

2.4. Diagonal Matrix Proximal Functions

In this subsection, we restrict H_t to be a diagonal matrix, for two important reasons: (i) the diagonal matrix will provide results easier to understand than that for the general matrix; (ii) for high dimension problem the general matrix may result in prohibitively expensive computational cost, which is not desirable.

Firstly, we notice that the upper bound in the Theorem 1 relies on $\sum_{t=1}^T \|\mathbf{g}_t\|^2_{H_t^*}$. If we assume all the $\mathbf{g}_t$'s are known in advance, we could minimize this term by setting $H_t = \text{diag}(\mathbf{s}), \forall t$. We shall use the following proposition.

Proposition 1. *For any $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_T \in \mathbb{R}^{d_1}$, we have*

$$\min_{diag(\mathbf{s})\succeq 0,\ \mathbf{1}^\top \mathbf{s}\le c} \sum_{t=1}^T \|\mathbf{g}_t\|^2_{diag(\mathbf{s})} = \frac{1}{c}\Big(\sum_{i=1}^{d_1} \|\mathbf{g}_{1:T,i}\|\Big)^2,$$

where $\mathbf{g}_{1:T,i} = (g_{1,i}, \dots, g_{T,i})^\top$ and the minimum is attained at $s_i = c\|\mathbf{g}_{1:T,i}\| / \sum_{j=1}^{d_1} \|\mathbf{g}_{1:T,j}\|$.

We omit proof of this proposition, since it is easy to derive. Since we do not have all the $\mathbf{g}_t$'s in advance, we receives the stochastic (sub)gradients $\mathbf{g}_t$ sequentially instead. As a result, we propose to update the H_t incrementally as $H_t = aI + \text{diag}(\mathbf{s}_t)$, where $s_{t,i} = \|\mathbf{g}_{1:t,i}\|$ and $a \ge 0$. For these H_ts, we have the following inequality

$$\sum_{t=1}^T \|\mathbf{g}_t\|^2_{H_t^*} = \sum_{t=1}^T \langle \mathbf{g}_t, (aI + \text{diag}(\mathbf{s}_t))^{-1}\mathbf{g}_t\rangle \le \sum_{t=1}^T \langle \mathbf{g}_t, \text{diag}(\mathbf{s}_t)^{-1}\mathbf{g}_t\rangle \le 2\sum_{i=1}^{d_1} \|\mathbf{g}_{1:T,i}\|, \quad (3)$$

where the last inequality used Lemma 4 in (Duchi et al., 2011), which implies this update is a nearly optimal update method for the diagonal matrix case. Finally, the adaptive stochastic ADMM with diagonal matrix update (**Ada-SADMM**$_{diag}$) is summarized into the Algorithm 2.

For the convergence rate of the proposed Algorithm 2, we have the following specific theorem.

Theorem 2. *Let $\ell(\mathbf{w}, \xi_t)$ and $\varphi(\mathbf{w})$ be convex functions for any $t > 0$. Then for Algorithm 2, we have the following inequality for any $T \ge 1$ and $\rho > 0$*

$$\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|] \le \frac{1}{2T}\Big(\mathbb{E}[2\eta\sum_{i=1}^{d_1}\|\mathbf{g}_{1:T,i}\| + \frac{2}{\eta}\max_{t\le T}\|\mathbf{w}_t - \mathbf{w}_*\|^2_\infty \sum_{i=1}^{d_1}\|\mathbf{g}_{1:T,i}\|] + \beta D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{\beta}\Big).$$

If we further set $\eta = D_{\mathbf{w},\infty}/\sqrt{2}$ where $D_{\mathbf{w},\infty} = \max_{\mathbf{w},\mathbf{w}'} \|\mathbf{w} - \mathbf{w}'\|_\infty$, then we have

$$\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|] \le \frac{1}{T}\big(\sqrt{2}\mathbb{E}[D_{\mathbf{w},\infty}\sum_{i=1}^{d_1}\|\mathbf{g}_{1:T,i}\|] + \frac{\beta}{2}D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{2\beta}\big).$$

The proof is in the appendix.

Algorithm 2 Adaptive Stochastic ADMM with Diagonal Matrix Update (**Ada-SADMM**$_{diag}$).

Initialize: $\mathbf{w}_1 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{0}$, $\theta_1 = \mathbf{0}$, and $a > 0$.
for $t = 1, 2, \dots, T$ **do**
  Compute $\mathbf{g}_t = \ell'(\mathbf{w}_t, \xi_t)$;
  Update $H_t = aI + \text{diag}(\mathbf{s}_t)$, where $s_{t,i} = \|\mathbf{g}_{1:t,i}\|$;
  $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \mathcal{L}_{\beta,t}(\mathbf{w}, \mathbf{v}_t, \theta_t)$;
  $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{L}_{\beta,t}(\mathbf{w}_{t+1}, \mathbf{v}, \theta_t)$;
  $\theta_{t+1} = \theta_t - \beta(A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - \mathbf{b})$;
end for

2.5. Full Matrix Proximal Functions

In this subsection, we derive and analyze new updates when we estimate a full matrix H_t for the proximal function instead of a diagonal one. Although full matrix computation may not be attractive for high dimension problems, it may be helpful for tasks with low dimension. Furthermore, it will provide us with a more complete insight. Similar with the analysis for the diagonal case, we first introduce the following proposition (Lemma 15 in (Duchi et al., 2011)).

Proposition 2. *For any $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_T \in \mathbb{R}^{d_1}$, we have the following equality*

$$\min_{S\succeq 0,\ \text{tr}(S)\le c} \sum_{t=1}^T \|\mathbf{g}_t\|^2_{S^{-1}} = \frac{1}{c} tr(G_T),$$

where $G_T = \sum_{t=1}^T \mathbf{g}_t\mathbf{g}_t^\top$. and the minimizer is attained at $S = cG_T^{1/2}/\text{tr}(G_T^{1/2})$. If G_T is not of full rank, then we use its pseudo-inverse to replace its inverse in the minimization problem.

Because the (sub)gradients are received sequentially, we propose to update the H_t incrementally as $H_t = aI + G_t^{\frac{1}{2}}$, $G_t = \sum_{i=1}^t \mathbf{g}_i\mathbf{g}_i^\top$, $t = 1, \dots, T$. For these H_ts, we have the following inequalities

$$\sum_{t=1}^{T}\|\mathbf{g}_t\|^2_{H_t^*} \leq \sum_{t=1}^{T}\|\mathbf{g}_t\|^2_{S_t^{-1}} \leq 2\sum_{t=1}^{T}\|\mathbf{g}_t\|^2_{S_T^{-1}} = 2\mathrm{tr}(G_T^{\frac{1}{2}}), \quad (4)$$

where the last inequality used Lemma 10 in (Duchi et al., 2011), which implies this update is a nearly optimal update method for the full matrix case. Finally, the adaptive stochastic ADMM with full matrix update is summarized in Algorithm 3.

Algorithm 3 Adaptive Stochastic ADMM with Full Matrix Update (**Ada-SADMM**$_{full}$).

Initialize: $\mathbf{w}_1 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{0}$, $\theta_1 = \mathbf{0}$, $G_0 = 0$, and $a > 0$
for $t = 1, 2, \ldots, T$ **do**
  Compute $\mathbf{g}_t = \ell'(\mathbf{w}_t, \xi_t)$;
  Update $G_t = G_{t-1} + \mathbf{g}_t\mathbf{g}_t^\top$;
  Update $H_t = aI + S_t$, where $S_t = G_t^{\frac{1}{2}}$;
  $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \mathcal{L}_{\beta,t}(\mathbf{w}, \mathbf{v}_t, \theta_t)$;
  $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{L}_{\beta,t}(\mathbf{w}_{t+1}, \mathbf{v}, \theta_t)$;
  $\theta_{t+1} = \theta_t - \beta(A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - \mathbf{b})$;
end for

For the convergence rate of the above proposed Algorithm 3, we have the following specific theorem.

Theorem 3. *Let $\ell(\mathbf{w}, \xi_t)$ and $\varphi(\mathbf{w})$ be convex functions for any $t > 0$. Then for Algorithm 3, we have the following inequality for any $T \geq 1$, $\rho > 0$,*

$$\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho \parallel A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b} \parallel] \leq \frac{1}{2T}\{\mathbb{E}[2\eta\mathrm{tr}\ (G_T^{\frac{1}{2}}) + \frac{1}{\eta}\max_{t\leq T}\|\mathbf{w}_* - \mathbf{w}_t\|^2\mathrm{tr}\ (G_T^{\frac{1}{2}})] + \beta D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{\beta}\}.$$

Furthermore, if we set $\eta = D_{\mathbf{w},2}/2$, where $D_{\mathbf{w},2} = \max_{\mathbf{w}_1,\mathbf{w}_2}\|\mathbf{w}_1 - \mathbf{w}_2\|$, then we have

$$\mathbb{E}[f(\bar{\mathbf{u}}_T) - f(\mathbf{u}_*) + \rho\|A\bar{\mathbf{w}}_T + B\bar{\mathbf{v}}_T - \mathbf{b}\|] \leq \frac{1}{T}(\sqrt{2}\mathbb{E}[D_{\mathbf{w},2}\mathrm{tr}\ (G_T^{1/2})] + \frac{\beta}{2}D^2_{\mathbf{v}_*,B} + \frac{\rho^2}{2\beta}).$$

The proof is in the appendix.

3. Experiment

In this section, we evaluate the empirical performance of the proposed adaptive stochastic ADMM algorithms for solving Graph-Guided SVM (GGSVM) tasks, which is formulated as the following problem (Ouyang et al., 2013):

$$\min_{\mathbf{w},\mathbf{v}} \frac{1}{n}\sum_{i=1}^{n}[1 - y_i\mathbf{x}_i^\top\mathbf{w}]_+ + \frac{\gamma}{2}\|\mathbf{w}\|^2 + \nu\|\mathbf{v}\|_1,$$
$$s.t.\ F\mathbf{w} - \mathbf{v} = 0,$$

where $[z]_+ = \max(0, z)$ and the matrix F is constructed based on a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. For this graph, $\mathcal{V} = \{w_1, \ldots, w_{d_1}\}$ is a set of variables and $\mathcal{E} = \{e_1, \ldots, e_{|\mathcal{E}|}\}$, where $e_k = \{i, j\}$ is assigned with a weight α_{ij}. And the corresponding F is in the form: $F_{ki} = \alpha_{ij}$ and $F_{kj} = -\alpha_{ij}$. To construct a graph for a given dataset, we adopt the sparse inverse covariance estimation (Friedman et al., 2008) and determine the sparsity pattern of the inverse covariance matrix Σ^{-1}. Based on the inverse covariance matrix, we connect all index pairs (i, j) with $\Sigma^{-1}_{ij} \neq 0$ and assign $\alpha_{ij} = 1$.

3.1. Experimental Testbed and Setup

To examine the performance, we test all the algorithms on six real-world datasets from web machine learning repositories, which are listed in the Table 1. The "news20" dataset was downloaded from www.cs.nyu.edu/~roweis/data.html. All other datasets were downloaded from the LIBSVM website. For each dataset, we randomly divide it into two folds: training set with 80% of examples and test set with the rest.

Table 1. Several real-world datasets in our experiments.

Dataset	# examples	# features
a9a	48,842	123
mushrooms	8,124	112
news20	16,242	100
splice	3,175	60
svmguide3	1,284	21
w8a	64,700	300

To make a fair comparison, all algorithms adopt the same experimental setup. In particular, we set the penalty parameter $\gamma = \nu = 1/n$, where n is the number of training examples, and the trade-off parameter $\beta = 1$. In addition, we set the step size parameter $\eta_t = 1/(\gamma t)$ for SADMM according to the theorem 2 in (Ouyang et al., 2013). Finally, the smooth parameter a is set as 1, and the step size for adaptive stochastic ADMM algorithms are searched from $2^{[-5:5]}$ using cross validation.

All the experiments were conducted with 5 different random seeds and 2 epochs ($2n$ iterations) for each dataset. All the result were reported by averaging over these 5 runs. We evaluated the learning performance by measuring objective values, i.e., $f(\mathbf{u})$, and test error rates on the test datasets. In addition, we also evaluate computational efficiency of all the algorithms by their running time. All experiments were run in Matlab over a machine of 3.4GHz CPU.

3.2. Performance Evaluation

The figure 1 shows the performance of all the algorithms in comparison over trials, from which we can draw several observations. Firstly, the left column shows the objective

Table 2. Evaluation of stochastic ADMM algorithms on the real-world data sets.

Algorithm	a9a			mushrooms		
	Objective value	Test error rate	Time (s)	Objective value	Test error rate	Time (s)
SADMM	2.6002 ± 0.4271	0.1646 ± 0.0075	56.0914	0.7353 ± 0.2104	0.0350 ± 0.0136	7.6619
Ada-SADMM$_{diag}$	0.3550 ± 0.0001	0.1501 ± 0.0012	94.7619	0.0096 ± 0.0005	0.0006 ± 0.0000	13.0355
Ada-SADMM$_{full}$	0.3545 ± 0.0001	0.1498 ± 0.0013	622.4459	0.0091 ± 0.0002	0.0002 ± 0.0003	67.8198
Algorithm	news20			splice		
	Objective value	Test error rate	Time (s)	Objective value	Test error rate	Time (s)
SADMM	0.5652 ± 0.0151	0.1333 ± 0.0034	13.2948	108.6823 ± 20.9655	0.2454 ± 0.0322	0.9821
Ada-SADMM$_{diag}$	0.3139 ± 0.0003	0.1280 ± 0.0015	22.4788	0.3793 ± 0.0054	0.1578 ± 0.0059	1.3674
Ada-SADMM$_{full}$	0.3204 ± 0.0007	0.1284 ± 0.0016	148.5242	0.3710 ± 0.0014	0.1550 ± 0.0079	7.0392
Algorithm	svmguide3			w8a		
	Objective value	Test error rate	Time (s)	Objective value	Test error rate	Time (s)
SADMM	1.6143 ± 0.3123	0.2161 ± 0.0052	0.1288	0.3357 ± 0.0916	0.0957 ± 0.0012	191.7544
Ada-SADMM$_{diag}$	0.5163 ± 0.0046	0.2056 ± 0.0060	0.2014	0.1526 ± 0.0010	0.0931 ± 0.0005	326.1392
Ada-SADMM$_{full}$	0.5230 ± 0.0044	0.2000 ± 0.0044	0.4602	0.1469 ± 0.0006	0.0929 ± 0.0003	4027.1963

values of the three algorithms. We can observe that the two adaptive stochastic ADMM algorithms converge much faster than SADMM, which shows the effectiveness of exploration of adaptive regularization (Bregman Divergence) to accelerate stochastic ADMM. Secondly, compared with Ada-SADMM$_{diag}$, Ada-SADMM$_{full}$ achieves slightly smaller objective values on most of the datasets, which indicates full matrix is slightly more informative than the diagonal one. This should be due to the lower dimensions of these datasets. Thirdly, the central column provides test error rates of three algorithms, where we observe that the two adaptive algorithms achieve significantly smaller or comparable test error rates at 0.25-th epoch than SADMM at 2-th epoch. This observation indicates that we can terminate the two adaptive algorithms earlier to save time and at the same time achieve similar performance compared with SADMM. Finally, the right column shows the running time of three algorithms, which shows that during the learning process, the Ada-SADMM$_{full}$ is significantly slower while the Ada-SADMM$_{diag}$ is overall efficient compared with SADMM. In summary, the Ada-SADMM$_{diag}$ algorithm achieves a good trade-off between efficiency and effectiveness.

Table 2 summarizes the performance of all the compared algorithms over the six datasets, from which we can make similar observations. This again verifies the effectiveness of the proposed algorithms.

4. Conclusion

ADMM is a popular technique in machine learning. This paper studied a method to accelerate stochastic ADMM with adaptive regularization, by replacing the fixed proximal function with adaptive proximal function. Compared with traditional stochastic ADMM, we show that the proposed adaptive algorithms converge significantly faster through the proposed adaptive strategies. Promising experimental results on a variety of real-world datasets further validate the effectiveness of our techniques.

Acknowledgments

The research of Peilin Zhao and Tong Zhang is partially supported by NSF-IIS-1407939 and NSF-IIS 1250985. The research of Jinwei Yang and Ping Li is partially supported by NSF-III-1360971, NSF-Bigdata-1419210, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137.

Appendix: Some Proofs

Proof for Lemma 1

Proof. Firstly, using the convexity of ℓ and the definition of δ_t, we can obtain

$$\begin{aligned}&\ell(\mathbf{w}_t)-\ell(\mathbf{w})\le\langle\ell'(\mathbf{w}_t),\mathbf{w}_t-\mathbf{w}\rangle\\&=\langle\mathbf{g}_t,\mathbf{w}_{t+1}-\mathbf{w}\rangle+\langle\delta_t,\mathbf{w}-\mathbf{w}_t\rangle+\langle\mathbf{g}_t,\mathbf{w}_t-\mathbf{w}_{t+1}\rangle.\end{aligned}$$

Combining the above inequality with the relation between θ_t and θ_{t+1} will derive

$$\begin{aligned}&\ell(\mathbf{w}_t)-\ell(\mathbf{w})+\langle\mathbf{w}_{t+1}-\mathbf{w},-A^\top\theta_{t+1}\rangle\\&\le\langle\mathbf{g}_t,\mathbf{w}_{t+1}-\mathbf{w}\rangle+\langle\delta_t,\mathbf{w}-\mathbf{w}_t\rangle+\langle\mathbf{g}_t,\mathbf{w}_t-\mathbf{w}_{t+1}\rangle\\&\qquad+\langle\mathbf{w}_{t+1}-\mathbf{w},A^\top[\beta(A\mathbf{w}_{t+1}+B\mathbf{v}_{t+1}-\mathbf{b})-\theta_t]\rangle\\&=\underbrace{\langle\mathbf{g}_t+A^\top[\beta(A\mathbf{w}_{t+1}+B\mathbf{v}_t-\mathbf{b})-\theta_t],\mathbf{w}_{t+1}-\mathbf{w}\rangle}_{L_t}\\&+\underbrace{\langle\mathbf{w}-\mathbf{w}_{t+1},\beta A^\top B(\mathbf{v}_t-\mathbf{v}_{t+1})\rangle}_{M_t}+\langle\delta_t,\mathbf{w}-\mathbf{w}_t\rangle\\&\qquad+\underbrace{\langle\mathbf{g}_t,\mathbf{w}_t-\mathbf{w}_{t+1}\rangle}_{N_t}.\end{aligned}$$

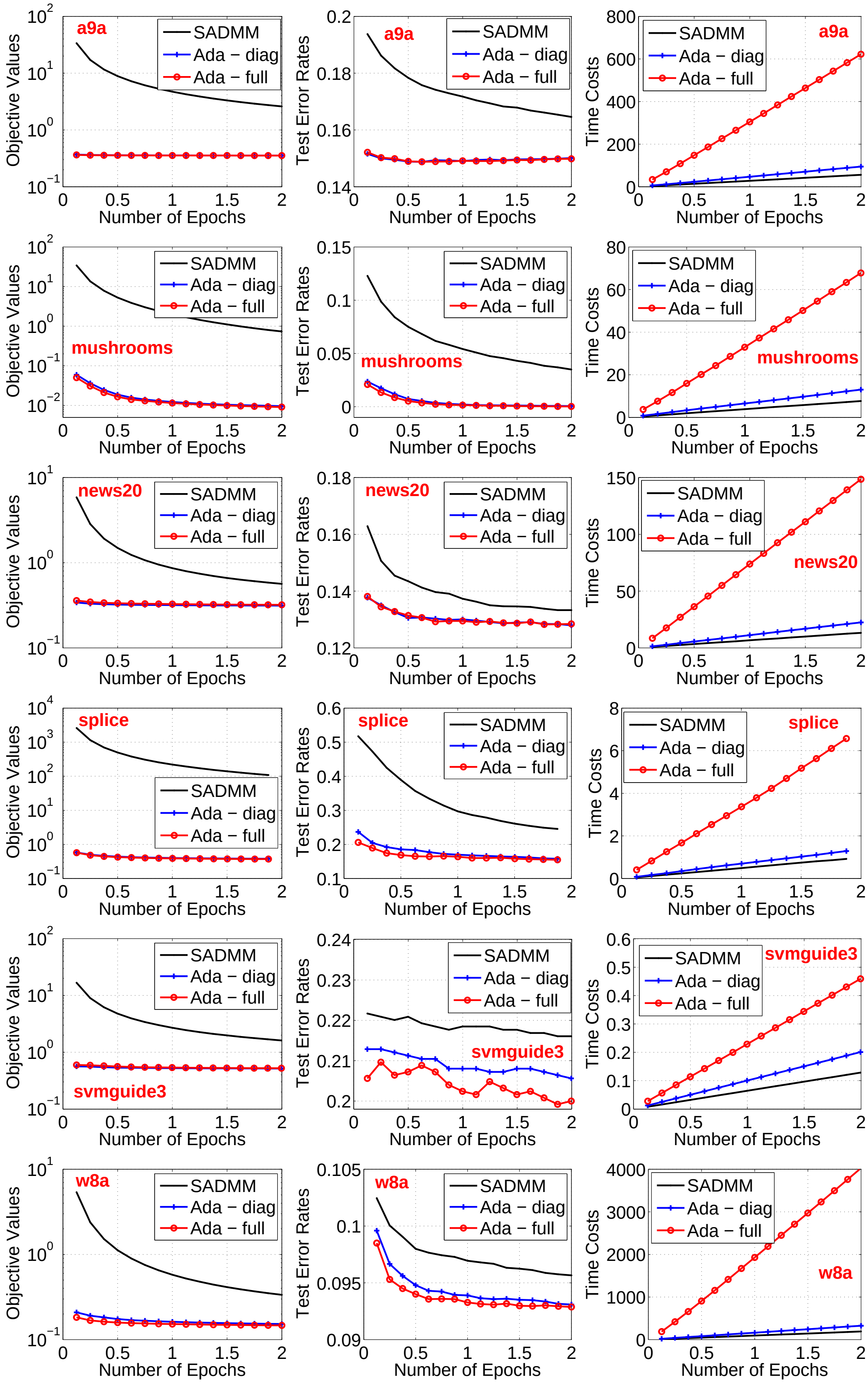

Figure 1. Comparison between SADMM with Ada-SADMM$_{diag}$ (“Ada-diag”) and Ada-SADMM$_{full}$ (“Ada-full”) on 6 real-world datasets. Epoch for the horizontal axis is the number of iterations divided by dataset size. **Left Panels**: Average objective values. **Middle Panels**: Average test error rates. **Right Panels**: Average time costs (in seconds).

To provide an upper bound for the first term L_t, taking $D(\mathbf{u},\mathbf{v}) = \mathcal{B}_{\phi_t}(\mathbf{u},\mathbf{v}) = \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|^2_{H_t}$ and applying Lemma 1 in (Ouyang et al., 2013) to the step of getting $\mathbf{w}_{t+1}$ in Algorithm 1, we will have

$$\begin{aligned}&\langle \mathbf{g}_t + A^\top[\beta(A\mathbf{w}_{t+1} + B\mathbf{v}_t - \mathbf{b}) - \theta_t], \mathbf{w}_{t+1} - \mathbf{w}\rangle\\ &\qquad \le \frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t,\mathbf{w}) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1},\mathbf{w}) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1},\mathbf{w}_t)].\end{aligned}$$

To provide an upper bound for the second term M_t, we can derive as follows

$$\begin{aligned}&\langle \mathbf{w} - \mathbf{w}_{t+1}, \beta A^\top B(\mathbf{v}_t - \mathbf{v}_{t+1})\rangle\\ &= \beta\langle A\mathbf{w} - A\mathbf{w}_{t+1}, B\mathbf{v}_t - B\mathbf{v}_{t+1}\rangle\\ &= \frac{\beta}{2}[(\|A\mathbf{w} + B\mathbf{v}_t - \mathbf{b}\|^2 - \|A\mathbf{w} + B\mathbf{v}_{t+1} - \mathbf{b}\|^2)\\ &+(\|A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - \mathbf{b}\|^2 - \|A\mathbf{w}_{t+1} + B\mathbf{v}_t - \mathbf{b}\|^2)]\\ &\le \frac{\beta}{2}(\|A\mathbf{w} + B\mathbf{v}_t - \mathbf{b}\|^2 - \|A\mathbf{w} + B\mathbf{v}_{t+1} - \mathbf{b}\|^2)\\ &\qquad + \frac{1}{2\beta}\|\theta_{t+1} - \theta_t\|^2.\end{aligned}$$

To drive an upper bound for the final term N_t, we can use Young's inequality to get

$$\begin{aligned}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_{t+1}\rangle &\le \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2} + \frac{\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2_{H_t}}{2\eta}\\ &= \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2} + \frac{\mathcal{B}_{\phi_t}(\mathbf{w}_t,\mathbf{w}_{t+1})}{\eta}.\end{aligned}$$

Replacing the terms L_t, M_t and N_t with their upper bounds, we will get

$$\begin{aligned}&\ell(\mathbf{w}_t) - \ell(\mathbf{w}) + \langle \mathbf{w}_{t+1} - \mathbf{w}, -A^\top\theta_{t+1}\rangle \le\\ &\frac{1}{\eta}[\mathcal{B}_{\phi_t}(\mathbf{w}_t,\mathbf{w}) - \mathcal{B}_{\phi_t}\mathbf{w}_{t+1},\mathbf{w})] + \frac{\eta\|\mathbf{g}_t\|^2_{H_t^*}}{2} + \langle \delta_t, \mathbf{w} - \mathbf{w}_t\rangle\\ &+\frac{\beta}{2}(\|A\mathbf{w} + B\mathbf{v}_t - \mathbf{b}\|^2 - \|A\mathbf{w} + B\mathbf{v}_{t+1} - \mathbf{b}\|^2)\\ &\qquad + \frac{1}{2\beta}\|\theta_{t+1} - \theta_t\|^2.\end{aligned}$$

Due to the optimality condition of the step of updating $\mathbf{v}$ in Algorithm 1, i.e., $\partial_\mathbf{v}\mathcal{L}_{\beta,t}(\mathbf{w}_{t+1},\mathbf{v}_{t+1},\theta_t)$ and the convexity of φ, we have

$$\varphi(\mathbf{v}_{t+1}) - \varphi(\mathbf{v}) + \langle \mathbf{v}_{t+1} - \mathbf{v}, -B^\top\theta_{t+1}\rangle \le 0.$$

Using the fact $A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - b = (\theta_t - \theta_{t+1})/\beta$, we have

$$\begin{aligned}&\langle \theta_{t+1} - \theta, A\mathbf{w}_{t+1} + B\mathbf{v}_{t+1} - \mathbf{b}\rangle\\ &\qquad = \frac{1}{2\beta}(\|\theta - \theta_t\|^2 - \|\theta - \theta_{t+1}\|^2 - \|\theta_{t+1} - \theta_t\|^2).\end{aligned}$$

Combining the above three inequalities and re-arranging the terms will conclude the proof. □

Proof of Theorem 2

Proof. We have the following inequality

$$\begin{aligned}&2\sum_{t=1}^T[\mathcal{B}_{\phi_t}(\mathbf{w}_t,\mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1},\mathbf{w}_*)]\\ &= \sum_{t=1}^T(\|\mathbf{w}_t - \mathbf{w}_*\|^2_{H_t} - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_t})\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}(\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_{t+1}} - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_t})\\ &= \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{(H_{t+1} - H_t)}\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\max_i(\mathbf{w}_{t+1,i} - \mathbf{w}_{*,i})^2\|\mathbf{s}_{t+1} - \mathbf{s}_t\|_1\\ &= \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_\infty(\mathbf{s}_{t+1} - \mathbf{s}_t)^\top\mathbf{1}\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \max_{t\le T}\|\mathbf{w}_t - \mathbf{w}_*\|^2_\infty\mathbf{s}_T^\top\mathbf{1} - \|\mathbf{w}_1 - \mathbf{w}_*\|^2_\infty\mathbf{s}_1^\top\mathbf{1}\\ &\le \max_{t\le T}\|\mathbf{w}_t - \mathbf{w}_*\|^2_\infty\sum_{i=1}^{d_1}\|\mathbf{g}_{1:T,i}\|,\end{aligned}$$

where the last inequality used $\langle \mathbf{s}_T, \mathbf{1}\rangle = \sum_{i=1}^{d_1}\|\mathbf{g}_{1:T,i}\|$ and $\|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} \le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_\infty\mathbf{s}_1^\top\mathbf{1}$. Plugging the above inequality and the inequality (4) into the inequality (2), will conclude the first part of the theorem. Then the second part is trivial to be derived. □

Proof of Theorem 3

Proof. We consider the sum of the difference

$$\begin{aligned}&2\sum_{t=1}^T[\mathcal{B}_{\phi_t}(\mathbf{w}_t,\mathbf{w}_*) - \mathcal{B}_{\phi_t}(\mathbf{w}_{t+1},\mathbf{w}_*)]\\ &= \sum_{t=1}^T(\|\mathbf{w}_t - \mathbf{w}_*\|^2_{H_t} - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_t})\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}(\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_{t+1}} - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{H_t})\\ &= \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2_{(G_{t+1}^{\frac{1}{2}} - G_t^{\frac{1}{2}})}\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2\lambda_{max}(G_{t+1}^{\frac{1}{2}} - G_t^{\frac{1}{2}})\\ &= \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \sum_{t=1}^{T-1}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2\mathrm{tr}(G_{t+1}^{\frac{1}{2}} - G_t^{\frac{1}{2}})\\ &\le \|\mathbf{w}_1 - \mathbf{w}_*\|^2_{H_1} + \max_{t\le T-1}\|\mathbf{w}_t - \mathbf{w}_*\|^2\mathrm{tr}(G_T^{\frac{1}{2}})\\ &\qquad - \|\mathbf{w}_1 - \mathbf{w}_*\|^2\mathrm{tr}(G_1^{\frac{1}{2}}) \le \max_{t\le T}\|\mathbf{w}_t - \mathbf{w}_*\|^2\mathrm{tr}(G_T^{\frac{1}{2}}).\end{aligned}$$

Plugging the above inequality and the inequality (4) into the inequality (2), will conclude the first part of the theorem. Then the second part is trivial to be derived. □

References

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Deng, Wei and Yin, Wotao. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, DTIC Document, 2012.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Eckstein, Jonathan and Bertsekas, Dimitri P. On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Gabay, Daniel. Chapter ix applications of the method of multipliers to variational inequalities. *Studies in mathematics and its applications*, 15:299–331, 1983.

Gabay, Daniel and Mercier, Bertrand. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Glowinski, R. and Marroco, A. Sur lapproximation, par elements nis dordre un, et la resolution, par penalisationdualite, dune classe de problems de dirichlet non lineares. *Revue Francaise dAutomatique, Informatique, et*, 1975.

Glowinski, Roland and Le Tallec, Patrick. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*, volume 9. SIAM, 1989.

Goldfarb, Donald, Ma, Shiqian, and Scheinberg, Katya. Fast alternating linearization methods for minimizing the sum of two convex functions. *Math. Program.*, 141(1-2): 349–382, 2013.

Goldstein, Tom and Osher, Stanley. The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.

He, Bingsheng and Yuan, Xiaoming. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2): 700–709, 2012.

Luo, Zhi-Quan. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

Monteiro, Renato D. C. and Svaiter, Benar Fux. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

Ouyang, Hua, He, Niao, Tran, Long, and Gray, Alexander G. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 80–88, 2013.

Suzuki, Taiji. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 392–400, 2013.

Suzuki, Taiji. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 736–744, 2014.

Wang, Huahua and Banerjee, Arindam. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1119–1126, 2012.

Yang, Junfeng and Zhang, Yin. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1):250–278, 2011.

Zhong, Wenliang and Kwok, James Tin-Yau. Fast stochastic alternating direction method of multipliers. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 46–54, 2014.