Stochastic Alternating Direction Method of Multipliers

Hua Ouyang † HOUYANG@CC.GATECH.EDU
Niao He ‡ NHE6@ISYE.GATECH.EDU
Long Q. Tran † LTRAN3@GATECH.EDU
Alexander Gray † AGRAY@CC.GATECH.EDU
† School of Computational Science and Engineering, Georgia Tech
‡ H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech

Abstract

The Alternating Direction Method of Multipliers (ADMM) has received lots of attention recently due to the tremendous demand from large-scale and data-distributed machine learning applications. In this paper, we present a stochastic setting for optimization problems with non-smooth composite objective functions. To solve this problem, we propose a stochastic ADMM algorithm. Our algorithm applies to a more general class of convex and nonsmooth objective functions, beyond the smooth and separable least squares loss used in lasso. We also demonstrate the rates of convergence for our algorithm under various structural assumptions of the stochastic function: $O(1/\sqrt{t})$ for convex functions and $O(\log t/t)$ for strongly convex functions. Compared to previous literature, we establish the convergence rate of ADMM for convex problems in terms of both the objective value and the feasibility violation. A novel application named Graph-Guided SVM is proposed to demonstrate the usefulness of our algorithm.

1. Introduction

The Alternating Direction Method of Multipliers (ADMM) (Glowinski & Marroco, 1975; Gabay & Mercier, 1976) is a very simple computational method for optimization proposed in 1970s. It stemmed from the augmented Lagrangian method (also known as the method of multipliers) dating back to late 1960s. The theoretical aspects of ADMM have been studied since 1980s, and its global convergence was established in the literature (Gabay, 1983; Glowinski & Tallec, 1989; Eckstein & Bertsekas, 1992). As reviewed in the comprehensive paper (Boyd et al., 2010), with the ability of dealing with objective functions separately and synchronously , ADMM turned out to be a natural fit in the field of large-scale data-distributed machine learning and big-data related optimization, and therefore received significant amount of attention in the last few years. Considerable work was conducted thereafter. On the theoretical side, ADMM was shown to have an $O(1/N)$ rate of convergence for convex problems (Monteiro & Svaiter, 2010; He & Yuan, 2012a;b; Wang & Banerjee, 2012), where N stands for the number of iterations. When objective functions are strongly convex and Lipschitz smooth, linear convergence rates were reported very recently (Hong & Luo, 2012; Deng & Yin, 2012). On the practical side, ADMM has been applied to a wide range of application domains, such as compressed sensing (Yang & Zhang, 2011), image restoration (Goldstein & Osher, 2009), video processing and matrix completion (Goldfarb et al., 2010). Besides that, many variations of this classical method have been recently developed, such as linearized (Goldfarb et al., 2010; Zhang et al., 2011; Yang & Yuan, 2012), accelerated (Goldfarb et al., 2010) and online (Wang & Banerjee, 2012) ADMM. However, most of these variants including the classic one implicitly assume full accessibilty of true data values, while in reality one can hardly ignore the existence of noise. A more natural way of handling this issue is to consider unbiased or even biased observations of true data, which leads us to the stochastic setting.

1.1. Stochastic Setting for ADMM

In this work, we study a family of convex optimization problems where our objective functions are stochastic and composite. Specifically, we are interested in the

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

following equality-constrained stochastic optimization:

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{\boldsymbol{\xi}}\theta_1(\mathbf{x},\boldsymbol{\xi}) + \theta_2(\mathbf{y}) \text{ s.t. } A\mathbf{x} + B\mathbf{y} = \mathbf{b}, \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^{d_1}, \mathbf{y} \in \mathbb{R}^{d_2}, A \in \mathbb{R}^{m\times d_1}, B \in \mathbb{R}^{m\times d_2}, \mathbf{b} \in \mathbb{R}^m$, $\mathcal{X}$ is a convex compact set, and $\mathcal{Y}$ is a closed convex set. We use the notation θ_1 for both the instance function value $\theta_1(\mathbf{x},\boldsymbol{\xi})$ and its expectation $\theta_1(\mathbf{x}) \equiv \mathbb{E}_{\boldsymbol{\xi}}\theta_1(\mathbf{x},\boldsymbol{\xi})$. We are able to draw a sequence of identical and independent (i.i.d.) observations from the random vector $\boldsymbol{\xi}$ that obeys a fixed but unknown distribution P. When $\boldsymbol{\xi}$ is deterministic, we can recover the traditional problem formulation of ADMM (Boyd et al., 2010). In our most general setting, real-valued functions $\theta_1(\cdot)$ and $\theta_2(\cdot)$ are convex but not necessarily continuously differentiable. We will make additional assumptions in Section 4, in which we suggest more structural information on θ_1.

1.2. Motivations

The stochasticity of the proposed setting is inspired by the structural risk minimization principle (Vapnik, 2000). Under this principle, a statistical learning system's goal is to minimize the *regularized expected risk function*: $R(\mathbf{x}) \equiv \mathbb{E}_{\boldsymbol{\xi}}L(\mathbf{x},\boldsymbol{\xi}) + \Omega(\mathbf{x})$, where $L(\mathbf{x},\boldsymbol{\xi})$ is the *loss* incurred when applying prediction rule $\mathbf{x}$ on a sample $\boldsymbol{\xi}$, and Ω is a regularizer. In the batch learning setting, one uses a set of training samples to minimize the *regularized empirical risk function* $R_{\text{emp}}(\mathbf{x}) \equiv \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{x},\boldsymbol{\xi}_i) + \Omega(\mathbf{x})$. With high probability, R and R_{emp} are close when the number of samples is large (Vapnik, 2000). However, to minimize R_{emp} one has to handle larger amount of samples which becomes less efficient under time and resource constraints. In the *stochastic* setting, in each iteration $\mathbf{x}$ is updated based on one noisy sample drawn from P instead of a finite training set. One obvious advantage is that the update costs much less time and resources than in the batch setting. Another advantage we will show later in this paper is that, when carefully designed, our algorithm optimizes the expected risk *directly* with good rates of convergence.

The proposed stochastic ADMM setting fits perfectly with the regularized expected risk minimization. Putting it into our canonical form (1): $\theta_1(\mathbf{x},\boldsymbol{\xi}) = L(\mathbf{x},\boldsymbol{\xi})$, $\theta_2(\mathbf{y}) = \Omega(\mathbf{y})$, and the constraint becomes $\mathbf{x} = \mathbf{y}$. Beyond this simple formulation, the objective separation of ADMM is so flexible that one can use a more general linear constraint $A\mathbf{x} + B\mathbf{y} = \mathbf{b}$ to model the *complex* structural information encoded in the regularizer $\Omega(\mathbf{x})$. For example, if $\Omega(v_1, v_2) = |v_1 - v_2|$, we could add a variable v_{12}, a linear constraint $v_1 - v_2 = v_{12}$, and simply minimize $\Omega(v_{12}) = |v_{12}|$, which is easier to handle under our stochastic setting for ADMM. More examples will be given in Section 5.

1.3. Our Contributions

We propose a stochastic setting of the ADMM problem and also design the Stochastic ADMM algorithm to solve this problem. A key algorithmic feature of our Stochastic ADMM that distinguishes our method from previous ADMM and its variants is the first-order approximation of θ_1 that we used to modify the augmented Lagrangian. This simple modification is not only necessary for the convergence analysis of our stochastic method, but also makes our method applicable to a more general class of convex objective functions which might not have a closed-form solution in minimizing the augmented θ_1 directly. Moreover, the linearization makes the updates simpler and faster, as demonstrated by the examples in Section 5.

We establish convergence rates under various structural assumptions of θ_1: $O(1/\sqrt{t})$ for convex functions and $O(\log t/t)$ for strongly convex functions in terms of both the objective value and the feasibility violation. By contrast, recent research (He & Yuan, 2012a;b; Wang & Banerjee, 2012) only show the convergence of ADMM *indirectly* in terms of the satisfaction of variational inequalities. We also demonstrate the usefulness of our algorithm with a novel application in Graph-Guided Support Vector Machine.

1.4. Related Work

A related setting named online ADMM was proposed in (Wang & Banerjee, 2012). In this setting, one does not assume $\boldsymbol{\xi}$ to be i.i.d., nor the objective to be stochastic, and the minimization of *regret* is concerned: $R(\mathbf{x}_{[1:t]}) \equiv \sum_{k=1}^{t}[\theta_1(\mathbf{x}_k,\boldsymbol{\xi}_k) + \theta_2(\mathbf{y}_k)] - \inf_{A\mathbf{x}+B\mathbf{y}=\mathbf{b}} \sum_{k=1}^{t}[\theta_1(\mathbf{x},\boldsymbol{\xi}_k) + \theta_2(\mathbf{y})]$. Besides that, it also differs from our stochastic ADMM algorithmically: a nonlinearized θ_1 is used in online ADMM, while a linearized one is adopted in our algorithm.

In an independent work (Suzuki, 2013), the author also linearized θ_1, and proposed dual averaging and proximal gradient methods for problem (1). The proposed OPG-ADMM algorithm enjoys the same order of convergence rates as our stochastic ADMM.

1.5. Notations

Throughout this paper, we denote a subgradient of a function f as f'. When f is differentiable, we will use ∇f . We denote by $\theta(\mathbf{u}) \equiv \theta_1(\mathbf{x}) + \theta_2(\mathbf{y})$ the sum of the stochastic and the deterministic functions.

For simplicity and clarity, we will use the following notations to denote stacked vectors or tuples:

$$\mathbf{u} \equiv \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \mathbf{w} \equiv \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\\ \boldsymbol{\lambda}\end{pmatrix}, \mathbf{w}_k \equiv \begin{pmatrix}\mathbf{x}_k\\ \mathbf{y}_k\\ \boldsymbol{\lambda}_k\end{pmatrix}, \bar{\mathbf{u}}_k \equiv \begin{pmatrix}\frac{1}{k}\sum_{i=0}^{k-1}\mathbf{x}_i\\ \frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_i\end{pmatrix},$$
$$\bar{\mathbf{w}}_k \equiv \begin{pmatrix}\frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_i\\ \frac{1}{k}\sum_{i=1}^{k}\mathbf{y}_i\\ \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\lambda}_i\end{pmatrix}, F(\mathbf{w}) \equiv \begin{pmatrix}-A^T\boldsymbol{\lambda}\\ -B^T\boldsymbol{\lambda}\\ A\mathbf{x}+B\mathbf{y}-\mathbf{b}\end{pmatrix}, \mathcal{W} \equiv \begin{pmatrix}\mathcal{X}\\ \mathcal{Y}\\ \mathbb{R}^m\end{pmatrix}. \tag{2}$$

the For a positive semidefinite matrix $G \in \mathbb{R}^{d_1\times d_1}$, we define the G-norm of a vector as $\|\mathbf{x}\|_G := \|G^{1/2}\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T G\mathbf{x}}$. We use $\langle\cdot,\cdot\rangle$ to denote the inner product in a finite dimensional Euclidean space. When there is no ambiguity, we often use $\|\cdot\|$ to denote the Euclidean norm $\|\cdot\|_2$. For a differentiable function $\omega(\cdot)$, Bregman divergence is defined as $D(\mathbf{u},\mathbf{v}) \equiv \omega(\mathbf{u}) - \omega(\mathbf{v}) - \langle\nabla\omega(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle$.

We assume that the optimal solution of (1) exists, and is denoted as $\mathbf{u}_* \equiv (\mathbf{x}_*^T, \mathbf{y}_*^T)^T$. The following quantities appear frequently in our convergence analysis.

$$\boldsymbol{\delta}_k \equiv \theta_1'(\mathbf{x}_{k-1}, \boldsymbol{\xi}_k) - \theta_1'(\mathbf{x}_{k-1}),$$
$$D_{\mathcal{X}} \equiv \sup_{\mathbf{x}_a,\mathbf{x}_b\in\mathcal{X}} \|\mathbf{x}_a - \mathbf{x}_b\|, \quad D_{\mathbf{y}_*,B} \equiv \|B(\mathbf{y}_0 - \mathbf{y}_*)\|. \tag{3}$$

1.6. Assumptions

Before presenting the algorithm and convergence results, we list the following assumptions that will be used in our statements. These assumptions provide bounds for the magnitude and variance of subgradients for the stochastic function.

Assumption 1. *For all* $\mathbf{x}\in\mathcal{X}$, $\mathbb{E}\Big[\|\theta_1'(\mathbf{x},\boldsymbol{\xi})\|^2\Big] \le M^2$.

Assumption 2. *For all* $\mathbf{x}\in\mathcal{X}$,

$$\mathbb{E}\Big[\exp\Big\{\|\theta_1'(\mathbf{x},\boldsymbol{\xi})\|^2/M^2\Big\}\Big] \le \exp\{1\}.$$

Assumption 3. *For all* $\mathbf{x}\in\mathcal{X}$,

$$\mathbb{E}\left[\|\theta_1'(\mathbf{x},\boldsymbol{\xi}) - \theta_1'(\mathbf{x})\|^2\right] \le \sigma^2.$$

2. Stochastic ADMM Algorithm

Directly solving problem (1) can be nontrivial, even if $\boldsymbol{\xi}$ is deterministic and the equality constraint is as simple as $\mathbf{x}-\mathbf{y}=\mathbf{0}$. For example, using the augmented Lagrangian method, one has to minimize the *augmented Lagrangian*:

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \mathcal{L}_\beta(\mathbf{x},\mathbf{y},\lambda) \equiv \min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \Big[\theta_1(\mathbf{x}) + \theta_2(\mathbf{y}) - \langle\boldsymbol{\lambda}, A\mathbf{x}+B\mathbf{y}-\mathbf{b}\rangle + \frac{\beta}{2}\|A\mathbf{x}+B\mathbf{y}-\mathbf{b}\|^2\Big], \tag{4}$$

where β is a pre-defined penalty parameter. This problem is at least not easier than solving the original one. The (deterministic) ADMM (Alg.1) solves this problem in a one-sweep Gauss-Seidel manner: minimizing $\mathcal{L}_\beta$ w.r.t. $\mathbf{x}$ and $\mathbf{y}$ alternatively given the other fixed, followed by a penalty update over the Lagrangian multiplier $\boldsymbol{\lambda}$.

Algorithm 1 Deterministic ADMM

0. Initialize $\mathbf{y}_0$ and $\boldsymbol{\lambda}_0 = 0$.
   for $k = 0, 1, 2, \ldots$ **do**
1. $\quad\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}_k, \boldsymbol{\lambda}_k)$.
2. $\quad\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y}\in\mathcal{Y}} \mathcal{L}_\beta(\mathbf{x}_{k+1}, \mathbf{y}, \boldsymbol{\lambda}_k)$.
3. $\quad\boldsymbol{\lambda}_{k+1} \leftarrow \boldsymbol{\lambda}_k - \beta\,(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b})$.
   end for

A variant deterministic algorithm named linearized ADMM replaces Line 1 of Alg.1 by

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}}\Big[\theta_1(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_k\|_G^2 + \frac{\beta}{2}\left\|(A\mathbf{x}+B\mathbf{y}_k-\mathbf{b}) - \boldsymbol{\lambda}_k/\beta\right\|^2\Big],$$

where $G \in \mathbb{R}^{d_1\times d_1}$ is positive semidefinite. This variant can be regarded as a generalization of the original ADMM. When $G=\mathbf{0}$, it is the same as Alg.1. When $G = rI_{d_1} - \beta A^T A$, it is equivalent to the following linearized proximal point method:

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}}\Big\{\theta_1(\mathbf{x}) + \frac{r}{2}\|\mathbf{x}-\mathbf{x}_k\|^2 + \beta(\mathbf{x}-\mathbf{x}_k)^T\left[A^T(A\mathbf{x_k}+B\mathbf{y}_k-\mathbf{b}-\boldsymbol{\lambda}_k/\beta)\right]\Big\}.$$

Note that the linearization is only applied to the quadratic function $\|(A\mathbf{x}+B\mathbf{y}_k-\mathbf{b}) - \boldsymbol{\lambda}_k/\beta\|^2$, but not to θ_1. This approximation helps in some cases when Line 1 of Alg.1 does not produce a closed-form solution given the quadratic term. For example, let $\theta_1(\mathbf{x}) = \|\mathbf{x}\|_1$ and A not identity.

As given in Alg.2, we propose a *Stochastic Alternating Direction Method of Multipliers* (*Stochastic ADMM*) algorithm. Our algorithm shares some features with the classical and the linearized ADMM. One can see that Line 2 and 3 are essentially the same as before. However we have a different updating rule for $\mathbf{x}$ as shown in Line 1, where we define an *approximated augmented Lagrangian*:

$$\hat{\mathcal{L}}_{\beta,k}(\mathbf{x},\mathbf{y},\boldsymbol{\lambda}) \equiv \theta_1(\mathbf{x}_k) + \langle\theta_1'(\mathbf{x}_k,\boldsymbol{\xi}_{k+1}),\mathbf{x}\rangle + \theta_2(\mathbf{y}) - \langle\boldsymbol{\lambda}, A\mathbf{x}+B\mathbf{y}-\mathbf{b}\rangle + \frac{\beta}{2}\|A\mathbf{x}+B\mathbf{y}-\mathbf{b}\|^2 + \frac{\|\mathbf{x}-\mathbf{x}_k\|^2}{2\eta_{k+1}}. \tag{5}$$

There are two differences between $\mathcal{L}_\beta$ (4) and $\hat{\mathcal{L}}_{\beta,k}$ (5). First, we replace $\theta_1(\mathbf{x})$ with a first-order approximation of $\theta_1(\mathbf{x}, \boldsymbol{\xi}_{k+1})$ at $\mathbf{x}_k$: $\theta_1(\mathbf{x}_k) + \mathbf{x}^T\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})$. This approximation has the same flavour of the stochastic mirror descent (Nemirovski et al., 2009) used for solving a one-variable stochastic convex problem. Second, similar to the linearized ADMM, we add an l_2-norm prox-function $\|\mathbf{x} - \mathbf{x}_k\|^2$ but scale it by a time-varying stepsize η_{k+1}. As we will see in Section 3, the choice of this stepsize is crucial in guaranteeing a convergence.

Algorithm 2 Stochastic ADMM

0. Initialize $\mathbf{x}_0, \mathbf{y}_0$ and set $\boldsymbol{\lambda}_0 = 0$.
   for $k = 0, 1, 2, \ldots$ **do**
1. $\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}} \hat{\mathcal{L}}_{\beta,k}(\mathbf{x}, \mathbf{y}_k, \boldsymbol{\lambda}_k)$.
2. $\mathbf{y}_{k+1} \leftarrow \arg\min_{\mathbf{y}\in\mathcal{Y}} \hat{\mathcal{L}}_{\beta,k}(\mathbf{x}_{k+1}, \mathbf{y}, \boldsymbol{\lambda}_k)$.
3. $\boldsymbol{\lambda}_{k+1} \leftarrow \boldsymbol{\lambda}_k - \beta\,(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b})$.
   end for

3. Main Results of Convergence Rates

In this section, we will show that our Stochastic ADMM given in Alg.2 exhibits a rate $O(1/\sqrt{t})$ of convergence in terms of both the objective value *and* the feasibility violation:

$$\mathbb{E}\Big[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2\Big] = O(1/\sqrt{t}).$$

All proofs in this section are provided in a longer version of this paper that is available at arXiv.org.

Before we address the main theorem on convergence rates, we will start with the following simple lemma, which is a very useful result by implementing Bregman divergence as a prox-function in proximal methods.

Lemma 1. *Let $l(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ be a convex function. Denote a subgradient $\mathbf{g} \in \partial l$. Let scalar $s \geq 0$. For any vector $\mathbf{u}$ and $\mathbf{v}$, denote their Bregman divergence as $D(\mathbf{u}, \mathbf{v})$. If $\forall \mathbf{u} \in \mathcal{X}$,*

$$\mathbf{x}^* \equiv \arg\min_{\mathbf{x}\in\mathcal{X}} l(\mathbf{x}) + sD(\mathbf{x}, \mathbf{u}), \tag{6}$$

then

$$\langle \mathbf{g}(\mathbf{x}^*), \mathbf{x}^* - \mathbf{x}\rangle \leq s\left[D(\mathbf{x}, \mathbf{u}) - D(\mathbf{x}, \mathbf{x}^*) - D(\mathbf{x}^*, \mathbf{u})\right].$$

Utilizing the above lemma, we are able to obtain an upper bound of the variation of the Lagrangian function and its first order approximation based on each iteration points.

Lemma 2. *$\forall \mathbf{w} \in \mathcal{W}, k \geq 0$ we have*

$$\theta_1(\mathbf{x}_k) + \theta_2(\mathbf{y}_{k+1}) - \theta(\mathbf{u}) + (\mathbf{w}_{k+1} - \mathbf{w})^T F(\mathbf{w}_{k+1}) \leq \frac{\eta_{k+1}\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \frac{\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2}{2\eta_{k+1}} + \frac{\beta}{2}\left(\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\|^2 - \|A\mathbf{x} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2\right) + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k\rangle + \frac{1}{2\beta}\left(\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_k\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_{k+1}\|_2^2\right). \tag{7}$$

In what follows, we will present our main theorem of the convergence in two fashions, both in terms of expectation and probability satisfication.

Theorem 1. *Let $\eta_k = \frac{D_\mathcal{X}}{M\sqrt{2k}}$ for all $k \geq 1$. Define*

$$M_1(t) \equiv \frac{\sqrt{2}D_\mathcal{X} M}{\sqrt{t}} \quad \text{and} \quad M_2(t) \equiv \frac{\beta D_{\mathbf{y}_*,B}^2 + \rho^2/\beta}{2t}. \tag{8}$$

Then $\forall \rho > 0$ and $t \geq 1$ we have:

(i) Under Assumption 1

$$\mathbb{E}[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|] \leq M_1(t) + M_2(t). \tag{9}$$

(ii) Under Assumption 1 and 2, $\forall \Omega > 0$

$$Prob\Big\{\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\| > (1 + \Omega/2 + 2\sqrt{2\Omega})M_1(t) + M_2(t)\Big\} \leq 2\exp\{-\Omega\}. \tag{10}$$

Remark 1. *Observe that our proof techniques can also be adapted to the deterministic case where no noise takes place. We are able to obtain a similar result for the classic deterministic ADMM:*

$$\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2 \leq \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t}.$$

The positive ρ in the preceding results controls the trade-off between the objective value reduction and the feasibility satisfaction. For a fixed ρ, one can set an optimal $\beta = \rho/D_{\mathbf{y}_*,B}$ such that the upper bound is minimized.

While the resulting $O(1/t)$ rate for the deterministic ADMM is the same as those in the existing literature, the above finding is an advance in the theoretical aspects of ADMM. Our convergence rate for general convex functions is proved in terms of both the objective value and the feasibility violation. By contrast, the existing literature (He & Yuan, 2012a;b; Wang & Banerjee, 2012) only shows the convergence of ADMM in terms of the satisfaction of variational inequalities, which is not a direct measure of how fast an algorithm reaches the optimal solution.

4. Extensions

4.1. Strongly Convex θ_1

When function $\theta_1(\cdot)$ is strongly convex, the convergence rate of Stochastic ADMM can be improved to $O\left(\frac{\log t}{t}\right)$, as shown in the following result.

Theorem 2. *When θ_1 is μ-strongly convex with respect to $\|\cdot\|$, taking $\eta_k = \frac{1}{k\mu}$ in Alg.2, under Assumption 1 we have $\forall \rho > 0, t \geq 1$,*

$$\begin{aligned}&\mathbb{E}\left[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2\right] \\ &\leq \frac{M^2 \log t}{\mu t} + \frac{\mu D_{\mathcal{X}}^2}{2t} + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t}.\end{aligned} \tag{11}$$

4.2. Lipschitz Smooth θ_1

Since the bounds given in Theorem 1 are related to the magnitude of subgradients, they do not provide any intuition of the performance in low-noise scenarios. With a Lipschitz smooth function θ_1, we are able to obtain convergence rates in terms of the variations of gradients, as stated in Assumption 3. Besides, under this assumption we are able to replace the unusual definition of $\bar{\mathbf{u}}_k$ in (2) with the following:

$$\bar{\mathbf{u}}_k \equiv \begin{pmatrix} \frac{1}{k}\sum_{i=1}^k \mathbf{x}_i \\ \frac{1}{k}\sum_{i=1}^k \mathbf{y}_i \end{pmatrix}. \tag{12}$$

Theorem 3. *When $\theta_1(\cdot)$ is L-Lipschitz smooth with respect to $\|\cdot\|$, taking $\eta_k = \frac{1}{L + \sigma\sqrt{2k}/D_{\mathcal{X}}}$ in Alg.2, under Assumption 3 we have $\forall \rho > 0, t \geq 1$,*

$$\begin{aligned}&\mathbb{E}\left[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2\right] \\ &\leq \frac{\sqrt{2}D_{\mathcal{X}}\sigma}{\sqrt{t}} + \frac{LD_{\mathcal{X}}^2}{2t} + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t}.\end{aligned} \tag{13}$$

5. Examples and Numerical Evaluations

5.1. Lasso

As one of the many motivating examples given in the review of ADMM (Boyd et al., 2010), the l_1-regularized sparse least squares problem, also known as *lasso*, fits the general class of (1) very naturally. The composite functions can be written as:

$$\theta_1(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{2}\left(l - \mathbf{x}^T\mathbf{s}\right)^2, \quad \theta_2(\mathbf{y}) = \gamma\|\mathbf{y}\|_1, \tag{14}$$

where the training sample $\boldsymbol{\xi}$ contains feature-label pair $\{\mathbf{s}, l\}$ and γ is a regularization parameter. The constraint simply becomes $A = I$, $B = -I$, $\mathbf{b} = \mathbf{0}$. Same as in the deterministic case, applying stochastic ADMM to l_1-regularized problem produces closed-form updating rules. The three updates for (14) are actually very simple:

$$\begin{aligned}\mathbf{x}_{k+1} &\leftarrow \frac{(l_{k+1} - \mathbf{s}_{k+1}^T\mathbf{x}_k)\mathbf{s}_{k+1} + \boldsymbol{\lambda}_k + \beta\mathbf{y}_k + \mathbf{x}_k/\eta_{k+1}}{\beta + 1/\eta_{k+1}}, \\ \mathbf{y}_{k+1} &\leftarrow S_{\frac{\gamma}{\beta}}\left(\mathbf{x}_{k+1} - \boldsymbol{\lambda}_k/\beta\right), \\ \boldsymbol{\lambda}_{k+1} &\leftarrow \boldsymbol{\lambda}_k - \beta(\mathbf{x}_{k+1} - \mathbf{y}_{k+1}),\end{aligned} \tag{15}$$

where the soft-thresholding operator $S_\alpha(\mathbf{x})$ is defined in the same way as in (Boyd et al., 2010):

$$S_\alpha(\mathbf{x}) \equiv \begin{cases} x_i - \alpha, & \text{if } x_i > \alpha \\ 0, & \text{if } |x_i| \leq \alpha \\ x_i + \alpha, & \text{if } x_i < -\alpha \end{cases}, \ \forall i.$$

Some vector-scaling operations can be saved by by replacing $\boldsymbol{\lambda}_k$ with $\beta\boldsymbol{\zeta}_k$ in (15):

$$\begin{aligned}\mathbf{x}_{k+1} &\leftarrow \frac{(l_{k+1} - \mathbf{s}_{k+1}^T\mathbf{x}_k)\mathbf{s}_{k+1} + \beta(\boldsymbol{\zeta}_k + \mathbf{y}_k) + \mathbf{x}_k/\eta_{k+1}}{\beta + 1/\eta_{k+1}}, \\ \mathbf{y}_{k+1} &\leftarrow S_{\frac{\gamma}{\beta}}\left(\mathbf{x}_{k+1} - \boldsymbol{\zeta}_k\right), \\ \boldsymbol{\zeta}_{k+1} &\leftarrow \boldsymbol{\zeta}_k - (\mathbf{x}_{k+1} - \mathbf{y}_{k+1}).\end{aligned}$$

For simple problems like lasso, it is indeed not necessary to formulate it as a two-variable equality-constrained optimization. Instead, we can directly minimize $\mathbb{E}(l - \mathbf{x}^T\mathbf{s})^2 + \gamma\|\mathbf{x}\|_1$ without any constraint. A popular class of methods for solving this composite-objective problem is called *proximal gradient* (Tseng, 2008; Nemirovski & Yudin, 1983) or *proximal splitting* (Combettes & Pesquet, 2011), which was investigated in various communities (Daubechies et al., 2004; Combettes & Wajs, 2005; Beck & Teboulle, 2009; Nesterov, 2007; Wright et al., 2009). Stochastic and online variants of these methods have also been developed recently, mainly in the large-scale machine learning and optimization literature (Langford et al., 2009; Lan, 2010; Lan & Ghadimi, 2011; Duchi & Singer, 2009; Hu et al., 2009; Xiao, 2010). For comparison purposes, here we take the *online forward-backward splitting* method (FOBOS) (Combettes & Pesquet, 2011; Duchi & Singer, 2009) as a first example. The FOBOS can be regarded as a proximal method with linearization of θ_1:

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{X}} \langle\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}\rangle + \theta_2(\mathbf{x}) + \frac{\|\mathbf{x} - \mathbf{x}_k\|^2}{2\eta_{k+1}}. \tag{16}$$

Comparing this method with our Alg.2, we can see that (16) is actually a special Stochastic ADMM that enforces $\mathbf{x}_k = \mathbf{y}_k$ (hence $\boldsymbol{\lambda}_k = \boldsymbol{\zeta}_k = \mathbf{0}$) in every iteration k. Note that this constraint feasibility is easy to enforce only because lasso comes with an extremely

simple constraint $\mathbf{x} = \mathbf{y}$. One of the most attractive features of (16) is its closed form solution for lasso in terms of soft-thresholding:

$$\mathbf{x}_{k+1} \leftarrow S_{\gamma\eta_{k+1}}\left[\mathbf{x}_k + \eta_{k+1}\left(l_{k+1} - \mathbf{s}_{k+1}^T\mathbf{x}_k\right)\mathbf{s}_{k+1}\right].$$

As we will see in our next example (Sec.5.2), with complex constraints, applying proximal splitting methods might not produce closed-form updates.

The second algorithm we are going to compare with is the *online ADMM* (Wang & Banerjee, 2012), which was proposed under a related but different setting of online learning. In this algorithm, the first-order approximation $\theta_1(\mathbf{x}_k) + \langle\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}\rangle$ is replaced by the exact function $\theta_1(\mathbf{x}, \boldsymbol{\xi}_k)$, which is a very straightforward "stochastization" of the deterministic ADMM. Applying this algorithm to lasso yields the following update for $\mathbf{x}$:

$$\mathbf{x}_{k+1} \leftarrow \left[\mathbf{s}_{k+1}\mathbf{s}_{k+1}^T + (\beta + 1/\eta_{k+1})\, I\right]^{-1}\mathbf{u},$$

while $\mathbf{u} \equiv l_{k+1}\mathbf{s}_{k+1} + \beta(\boldsymbol{\zeta}_k + \mathbf{y}_k) + \frac{\mathbf{x}_k}{\eta_{k+1}}$, and the updates for $\mathbf{y}$ and $\boldsymbol{\zeta}$ remain the same as our stochastic ADMM. Comparing the $\mathbf{x}$ updates of online and stochastic ADMM, it is clear that the linearization of our algorithm results in a much simpler inner product calculation, while a rank-1 matrix inversion is required for the online ADMM. Even with the Sherman-Morrison formula, this inversion process is still slower than the stochastic ADMM.

In the following experiments, we investigate two real-world datasets to examine the efficiency of our algorithm. Table 1 shows the statistics of these datasets and parameters we used for lasso. The first dataset, Abalone, obtained from the UCI Machine Learning Repository[1], is used to predict the age of abalones from physical measurements. The second dataset, E2006-tf-idf, a part of the 10K-Corpus[2], is used to predict the volatility of stock returns, an empirical measure of the financial risk of a company. The features are tf-idf of unigrams extracted from the financial reports of companies during the years 1996-2006 (Kogan et al., 2009).

The prediction results are shown in Fig.1 and 2. One can observe that all algorithms converge reasonably well, as expected from our discussions above. The stochastic ADMM performs slightly better than the other two in Abalone. For E2006-tf-idf, an acceptable accuracy is achieved with a fast sweep of merely 2,000 samples, less than 25% of the entire dataset.

[1] http://archive.ics.uci.edu/ml/datasets/Abalone
[2] http://www.ark.cs.cmu.edu/10K/

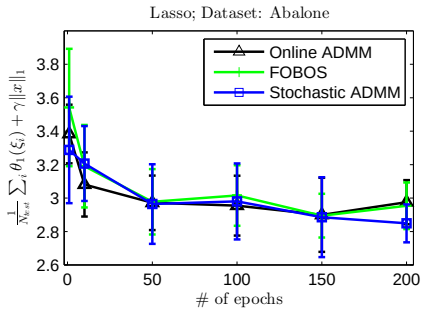

Figure 1. Lasso for Abalone Dataset.

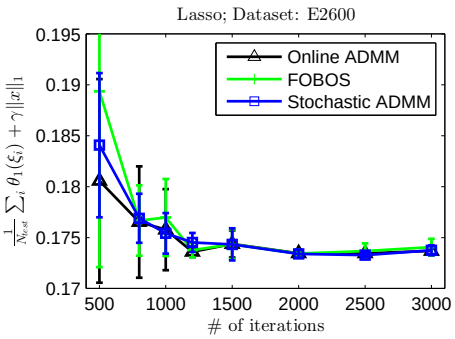

Figure 2. Lasso for E2600-tf-idf Dataset.

5.2. Graph-Guided SVM

Stochastic ADMM is more powerful for problems with complex equality constraints, for which proximal splitting methods such as FOBOS are no longer applicable, since there will be no closed-form for it. An important class of these problems is called the *generalized lasso* (Tibshirani & Taylor, 2011):

$$\min \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{2}\left(l - \mathbf{x}^T\mathbf{s}\right)^2\right] + \|F\mathbf{x}\|_1, \tag{17}$$

where the linear transformation $F \in \mathbb{R}^{f\times d_1}$ encodes the structural prior of a specific problem. When $F = I$, one recovers lasso. We can write (17) in our canonical form (1) with

$$\theta_1(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{2}\left(l - \mathbf{x}^T\mathbf{s}\right)^2, \quad \theta_2(\mathbf{y}) = \|\mathbf{y}\|_1, \\ A = F,\ B = -I,\ \mathbf{b} = \mathbf{0}. \tag{18}$$

where $\boldsymbol{\xi} = \{\mathbf{s}, l\}$ is a feature-label pair.

As a concrete example of the generalized lasso, we evaluate our algorithm based on the *graph-guided fused lasso* (GFlasso) framework (Kim et al., 2009), a graphical extension of the well-known *fused lasso* (Tibshirani et al., 2004). Denote graph $\mathcal{G} \equiv \{\mathcal{V}, \mathcal{E}\}$, in which $\mathcal{V} = \{x_1, \ldots, x_d\}$ is a set of the d variables of $\mathbf{x}$ and $\mathcal{E}$ is

Table 1. Two real-world datasets and parameters for lasso.

Name	# of Training Samples	# of Testing Samples	# of Dim. (d)	γ	β	$\frac{D_{\mathcal{X}}}{M\sqrt{2}}$
Abalone	3,342	835	8	0.01	1	1
E2006-tf-idf	16,087	3,308	150,360	0.1	1	1

the set of edges among $\mathcal{V}$. Each edge $\{i, j\}$ is assigned with a weight w_{ij}. The optimization of GFlasso can thus be formulated as:

$$\min \mathbb{E}_{\boldsymbol{\xi}}\big[\frac{1}{2}\left(l - \mathbf{x}^T\mathbf{s}\right)^2\big] + \gamma\|\mathbf{x}\|_1 + \nu \sum_{\{i,j\}\in\mathcal{E}} w_{ij}|x_i - x_j|. \tag{19}$$

The only difference between GFlasso and lasso is the last term, referred as the fusion penalty (Tibshirani et al., 2004), which penalizes the differences among variables connected in $\mathcal{G}$. A carefully designed fusion penalty helps in further reducing the risk of overfitting of our model over the training data. To implement Alg.2 to this problem, we only need to formulate the linear transformation F, which is very simple for GFlasso: $F_{ij} = w_{ij}$ and $F_{ji} = -w_{ij}$ for any edge $\{i, j\} \in \mathcal{E}$.

According to our convergence analysis, the loss θ_1 and regularizator θ_2 are allowed to be any convex functions. To meet the goal of classification, we replace the least squares loss in (19) by a nonsmooth *hinge loss* $L(\mathbf{x}, \boldsymbol{\xi}) \equiv \max\{0, 1 - l\mathbf{s}^T\mathbf{x}\}$ and the $l1$-norm by an Euclidean norm to enforce the maximum margin. The resulting combination is also known as support vector machine (SVM). With the additional graph-guided fusion penalty, we name our formulation *Graph-Guided SVM (GGSVM)*:

$$\begin{aligned} &\min \mathbb{E}_{\boldsymbol{\xi}} L(\mathbf{x}, \boldsymbol{\xi}) + \frac{\gamma}{2}\|\mathbf{x}\|_2^2 + \nu\|\mathbf{y}\|_1 \\ &\text{s.t. } F\mathbf{x} - \mathbf{y} = \mathbf{0}. \end{aligned} \tag{20}$$

Before presenting the penalty term, we first give an algorithmic solution of (20). Applying our stochastic ADMM to GGSVM, we obtain the following updates:

$$\begin{aligned} \mathbf{x}_{k+1} &\leftarrow \arg\min \mathbf{x}^T L'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}) + \gamma\mathbf{x}^T\mathbf{x}_k + \\ &\quad \frac{\beta}{2}\|F\mathbf{x} - \mathbf{y}_k - \boldsymbol{\lambda}_k/\beta\|_2^2 + \frac{\|\mathbf{x} - \mathbf{x}_k\|_2^2}{2\eta_{k+1}}, \\ \mathbf{y}_{k+1} &\leftarrow \arg\min \nu\|\mathbf{y}\|_1 + \frac{\beta}{2}\|F\mathbf{x}_{k+1} - \mathbf{y} - \boldsymbol{\lambda}_k/\beta\|_2^2, \\ \boldsymbol{\lambda}_{k+1} &\leftarrow \boldsymbol{\lambda}_k - \beta(F\mathbf{x}_{k+1} - \mathbf{y}_{k+1}). \end{aligned} \tag{21}$$

Without the graph-guided regularization, the stochastic ADMM becomes exactly the same as the classic stochastic gradient descent (SGD): $\mathbf{x}_{k+1} \leftarrow \arg\min \mathbf{x}^T L'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}) + \gamma\mathbf{x}^T\mathbf{x}_k + \frac{\|\mathbf{x}-\mathbf{x}_k\|_2^2}{2\eta_{k+1}}$.

The first two updates of (21) have close-forms:

$$\begin{aligned} \mathbf{x}_{k+1} &\leftarrow \left(\frac{I}{\eta_{k+1}} + \beta F^T F\right)^{-1} \Big[F^T(\beta\mathbf{y}_k + \boldsymbol{\lambda}_k) \\ &\quad + (1/\eta_{k+1} - \gamma)\mathbf{x}_k - L'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\Big], \\ \mathbf{y}_{k+1} &\leftarrow S_{\frac{\nu}{\beta}}\left(F\mathbf{x}_{k+1} - \frac{\boldsymbol{\lambda}_k}{\beta}\right). \end{aligned} \tag{22}$$

Note that this simple $\mathbf{x}$-update is exactly the benefit that stochastic ADMM brings. In contrast, neither the classic ADMM nor its variants have closed-forms due to the nonseparable form of the hinge loss.

In each $\mathbf{x}$-update of (22), due to the time-varying η_{k+1}, one has to solve a symmetric linear system with a different system matrix. This can be carried out using standard methods, e.g. conjugate gradient, where the sparsity of $F^T F$ can help in reducing the time complexity. However, for large-scale problems we can remove this computational burden completely by replacing η_{k+1} with a fixed η_t, if we want to run t iterations. This indeed leads to a convergent algorithm, although the proof is not shown in Section 3 due to limited space. By this means we only need to solve the linear system *once*, and save the result for successive iterations.

The data is the publicly available 20newsgroups dataset[3], which contains binary occurrences of 100 popular words counted from 16,242 newsgroup postings. On the top level of these postings are 4 main categories: computer, recreation, science and talks. We are interested in a multi-class classification task: to predict the category that a posting belongs to. We split the original data into a training set and a testing set. In each posting category, 80% postings are used for training and the rest 20% for testing. We use the one-vs-rest scheme for the multi-class classification.

The graphical structures we want to explore are the dependencies among these 100 words. Specifically, if two words i and j are strongly dependent, the difference between x_i and x_j in the linear predictor $\mathbf{x} \in \mathbb{R}^{100}$ should be penalized. In order to obtain F, we use the

[3] http://www.cs.nyu.edu/~roweis/data.html

sparse inverse covariance selection (Banerjee et al., 2008) (also known as *graphical lasso* (Friedman & Tibshirani, 2007; Boyd et al., 2010)) and determine the sparsity pattern of the inverse covariance matrix Σ^{-1}. By properly thresholding the components of Σ^{-1} to 0 and 1, we obtain the affinity matrix of $\mathcal{G}$ and plot the relations of these 100 words accordingly in Fig. 3. For simplicity, we take all the weights in F to be 1 and -1 whenever there is an edge.

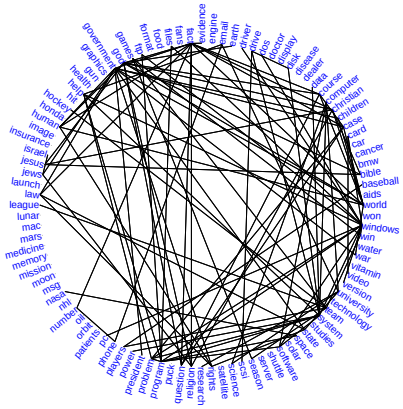

Figure 3. Graph of relations among 100 popular words in 20newsgroups dataset.

We compare the prediction accuracies with and without graphical regularization. Fig. 4 shows the experimental results. The x-axis stands for the number of epochs for stochastic algorithms. For this dataset, each epoch means 12,994 iterations. We calculate the mean and the standard deviation of all the accuracies based on 10 runs of experiments under the same setting. This figure clearly indicates that GGSVM outperforms the classical SVM consistently in every setting. After a single epoch, which corresponds to 1 iteration for the deterministic ADMM, the prediction accuracy is already very close to the best. This is a further evidence for the efficiency of our stochastic ADMM.

6. Summary and Future Work

In this paper, we have proposed the stochastic setting for ADMM along with our stochastic linearized ADMM algorithm. As a benefit of the first-order approximation on the stochastic function, our algorithm is applicable to a very broad class of problems even with functions that have no closed-form solution to the subproblem of minimizing the augmented θ_1. We have also established convergence rates under various structural assumptions of θ_1: $O(1/\sqrt{t})$ for convex functions and $O(\log t/t)$ for strongly convex functions. We are working on integrating Nesterov's optimal first-order methods (Nesterov, 2004) to our algorithm, which will help in achieving optimal convergence rates. More interesting and challenging applications will be carried out in our future work.

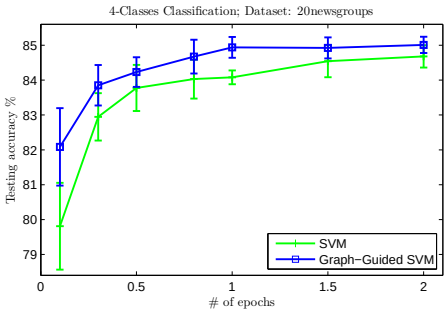

Figure 4. Accuracies for multi-class classification.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):193–202, 2009.

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

Combettes, P. L. and Pesquet, J. Proximal splitting methods in signal processing. In Bauschke, H. H. et. al. (ed.), *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter 10, pp. 185–212. Springer, New York, 2011.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):11681200, 2005.

Daubechies, I., Defrise, M., and Mol, C. De. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):14131457, 2004.

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

Duchi, J. and Singer, Y. Efficient online and batch learning using forward backward splitting. *JMLR*, 10:2899–2934, 2009.

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

Friedman, J. and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3): 432–441, 2007.

Gabay, D. Applications of the method of multipliers to variational inequalities. In Fortin, M. and Glowinski, R. (eds.), *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*. North-Holland: Amsterdam, 1983.

Gabay, D. and Mercier, B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1), 1976.

Glowinski, R. and Marroco, A. Sur lapproximation, par elements nis dordre un, et la resolution, par penalisation-dualite, dune classe de problems de dirichlet non lineares. *Revue Francaise dAutomatique, Informatique, et Recherche Operationelle*, 9(2), 1975.

Glowinski, R. and Tallec, P. L. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Studies in Applied and Numerical Mathematics. SIAM, 1989.

Goldfarb, D., Ma, S., and Scheinberg, K. Fast alternating linearization methods for minimizing the sum of two convex functions, 2010. URL http://arxiv.org/abs/0912.4571.

Goldstein, T. and Osher, S. The split bregman method for l1-regularized problems. *SIAM J. Imaging Sci.*, 2 (2):323–343, 2009.

He, B. and Yuan, X. On the $o(1/n)$ convergence rate of the douglas-rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012a.

He, B. and Yuan, X. On non-ergodic convergence rate of douglas-rachford alternating direction method of multipliers. 2012b.

Hong, Mingyi and Luo, Zhi-Quan. On the linear convergence of the alternating direction method of multipliers. http://arxiv.org/abs/1208.3922, 2012.

Hu, Chonghai, Kwok, James T., and Pan, Weike. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS 22*, 2009.

Kim, S., Sohn, K., and Xing, E. P. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):204–212, 2009.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. Predicting risk from financial reports with regression. In *NAACL-HLT 2009, Boulder, CO*, 2009.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010. doi: DOI10.1007/s10107-010-0434-y.

Lan, G. and Ghadimi, S. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, i: a generic algorithmic framework. *SIAM J. on Optimization*, 2011.

Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, pp. 777–801, 2009.

Monteiro, Renato D. C. and Svaiter, B. F. Iteration-complexity of block-decomposition algorithms and the alternating minimization augmented lagrangian method. Technical report, Georgia Institute of Technology, 2010.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. John Wiley and Sons, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization, A Basic Course*. Kluwer Academic Publishers, 2004.

Nesterov, Y. Gradient methods for minimizing composite objective function. Technical Report CORE DISCUSSION PAPER 2007/76, 2007.

Suzuki, Taiji. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of ICML*, 2013.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1): 91–108, 2004.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *SIAM J. Optim.*, 2008.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York Incorporated, 2000.

Wang, H. and Banerjee, A. Online alternating direction method. In *Proceedings of ICML*, 2012.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7): 2479–2493, 2009.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *JMLR*, 11:2543–2596, 2010.

Yang, J. and Yuan, X. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 2012. doi: http://dx.doi.org/10.1090/S0025-5718-2012-02598-1.

Yang, J. and Zhang, Y. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM J. on Scientific Computing*, 33(1):250–278, 2011.

Zhang, X., Burger, M., and Osher, S. A unified primal-dual algorithm framework based on bregman iteration. *J. of Scientific Computing*, 46(1):20–46, 2011.

7. Appendix

7.1. Proof of Lemma 1

Proof. Invoking the optimality condition for (6), we have

$$\langle \mathbf{g}(\mathbf{x}^*) + s\nabla D(\mathbf{x}^*, \mathbf{u}), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \ \forall \mathbf{x} \in \mathcal{X},$$

which is equivalent to

$$\begin{aligned} \langle \mathbf{g}(\mathbf{x}^*), \mathbf{x}^* - \mathbf{x} \rangle &\leq s \langle \nabla D(\mathbf{x}^*, \mathbf{u}), \mathbf{x} - \mathbf{x}^* \rangle \\ &= s \langle \nabla \omega(\mathbf{x}^*) - \nabla \omega(\mathbf{u}), \mathbf{x} - \mathbf{x}^* \rangle \\ &= s \left[D(\mathbf{x}, \mathbf{u}) - D(\mathbf{x}, \mathbf{x}^*) - D(\mathbf{x}^*, \mathbf{u}) \right]. \end{aligned}$$

□

7.2. Proof of Lemma 2

Proof. Due to the convexity of θ_1 and using the definition of $\boldsymbol{\delta}_k$, we have

$$\theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) \leq \left\langle \theta_1'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x} \right\rangle = \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x} \right\rangle + \left\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \right\rangle + \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle. \tag{23}$$

Applying Lemma 1 to Line 1 of Alg.2 and taking $D(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|^2$, we have

$$\begin{aligned} &\left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}) + A^T \left[\beta(A\mathbf{x}_{k+1} + B\mathbf{y}_k - \mathbf{b}) - \boldsymbol{\lambda}_k \right], \mathbf{x}_{k+1} - \mathbf{x} \right\rangle \\ &\leq \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 - \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \right) \end{aligned} \tag{24}$$

Combining (23) and (24) we have

$$\begin{aligned} &\theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \\ &\overset{(23)}{\leq} \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x} \right\rangle + \left\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \right\rangle + \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle + \\ &\quad \left\langle \mathbf{x}_{k+1} - \mathbf{x}, A^T \left[\beta(A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b}) - \boldsymbol{\lambda}_k \right] \right\rangle \\ &= \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}) + A^T \left[\beta(A\mathbf{x}_{k+1} + B\mathbf{y}_k - \mathbf{b}) - \boldsymbol{\lambda}_k \right], \mathbf{x}_{k+1} - \mathbf{x} \right\rangle + \\ &\quad \left\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \right\rangle + \left\langle \mathbf{x} - \mathbf{x}_{k+1}, \beta A^T B(\mathbf{y}_k - \mathbf{y}_{k+1}) \right\rangle + \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle \\ &\overset{(24)}{\leq} \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \right) + \left\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \right\rangle + \\ &\quad \left\langle \mathbf{x} - \mathbf{x}_{k+1}, \beta A^T B(\mathbf{y}_k - \mathbf{y}_{k+1}) \right\rangle + \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle \end{aligned} \tag{25}$$

We handle the last two terms separately:

$$\begin{aligned} &\left\langle \mathbf{x} - \mathbf{x}_{k+1}, \beta A^T B(\mathbf{y}_k - \mathbf{y}_{k+1}) \right\rangle = \beta \left\langle A\mathbf{x} - A\mathbf{x}_{k+1}, B\mathbf{y}_k - B\mathbf{y}_{k+1} \right\rangle \\ &= \frac{\beta}{2} \left[\left(\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\|^2 - \|A\mathbf{x} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2 \right) + \left(\|A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2 - \|A\mathbf{x}_{k+1} + B\mathbf{y}_k - \mathbf{b}\|^2 \right) \right] \\ &\leq \frac{\beta}{2} \left(\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\|^2 - \|A\mathbf{x} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2 \right) + \frac{1}{2\beta} \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2 \end{aligned} \tag{26}$$

and

$$\left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle \leq \frac{\eta_{k+1} \|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \frac{\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2}{2\eta_{k+1}}, \tag{27}$$

where the last step is due to Young's inequality. Inserting (26) and (27) into (25), we have

$$\begin{aligned} &\theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \\ &\leq \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 \right) + \frac{\eta_{k+1} \|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \left\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \right\rangle \\ &+ \frac{\beta}{2} \left(\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\|^2 - \|A\mathbf{x} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2 \right) + \frac{1}{2\beta} \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2, \end{aligned} \tag{28}$$

Due to the optimality condition of Line 2 in Alg.2 and the convexity of θ_2, we have

$$\theta_2(\mathbf{y}_{k+1}) - \theta_2(\mathbf{y}) + \left\langle \mathbf{y}_{k+1} - \mathbf{y}, -B^T\boldsymbol{\lambda}_{k+1} \right\rangle \leq 0. \tag{29}$$

Using Line 3 in Alg.2, we have

$$\begin{aligned} &\langle \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}, A\mathbf{x}_{k+1} + B\mathbf{y}_{k+1} - \mathbf{b} \rangle \\ &= \frac{1}{\beta} \langle \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}, \boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1} \rangle \\ &= \frac{1}{2\beta} \left(\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_k\|^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_{k+1}\|^2 - \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|^2 \right) \end{aligned} \tag{30}$$

Taking the summation of inequalities (28) (29) and (30), we obtain the result as desired. □

7.3. Proof of Theorem 1

Proof. (i). Invoking convexity of $\theta_1(\cdot)$ and $\theta_2(\cdot)$ and the monotonicity of operator $F(\cdot)$, we have $\forall \mathbf{w} \in \mathcal{W}$:

$$\begin{aligned} \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}) + (\bar{\mathbf{w}}_t - \mathbf{w})^T F(\bar{\mathbf{w}}_t) &\leq \frac{1}{t} \sum_{k=1}^{t} \left[\theta_1(\mathbf{x}_{k-1}) + \theta_2(\mathbf{y}_k) - \theta(\mathbf{u}) + (\mathbf{w}_k - \mathbf{w})^T F(\mathbf{w}_k) \right] \\ &= \frac{1}{t} \sum_{k=0}^{t-1} \left[\theta_1(\mathbf{x}_k) + \theta_2(\mathbf{y}_{k+1}) - \theta(\mathbf{u}) + (\mathbf{w}_{k+1} - \mathbf{w})^T F(\mathbf{w}_{k+1}) \right] \end{aligned} \tag{31}$$

Applying Lemma 2 at the optimal solution $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}_*, \mathbf{y}_*)$, we can derive from (31) that, $\forall \boldsymbol{\lambda}$

$$\begin{aligned} &\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + (\bar{\mathbf{x}}_t - \mathbf{x}_*)^T(-A^T\bar{\boldsymbol{\lambda}}_t) + (\bar{\mathbf{y}}_t - \mathbf{y}_*)^T(-B^T\bar{\boldsymbol{\lambda}}_t) + (\bar{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda})^T(A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}) \\ &\overset{(7)}{\leq} \frac{1}{t} \sum_{k=0}^{t-1} \left[\frac{\eta_{k+1} \|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x}_k - \mathbf{x}_*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \right) + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k \rangle \right] \\ &\quad + \frac{1}{t} \left(\frac{\beta}{2} \|A\mathbf{x}_* + B\mathbf{y}_0 - \mathbf{b}\|^2 + \frac{1}{2\beta} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|^2 \right) \\ &\leq \frac{1}{t} \sum_{k=0}^{t-1} \left[\frac{\eta_{k+1} \|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k \rangle \right] + \frac{1}{t} \left(\frac{D_{\mathcal{X}}^2}{2\eta_t} + \frac{\beta}{2} D_{\mathbf{y}_*, B}^2 + \frac{1}{2\beta} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_2^2 \right) \end{aligned} \tag{32}$$

The above inequality is true for all $\boldsymbol{\lambda} \in \mathbb{R}^m$, hence it also holds in the ball $\mathcal{B}_0 = \{\boldsymbol{\lambda} : \|\boldsymbol{\lambda}\|_2 \leq \rho\}$. Combing with the fact that the optimal solution must also be feasible, it follows that

$$\begin{aligned} &\max_{\boldsymbol{\lambda} \in \mathcal{B}_0} \left\{ \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + (\bar{\mathbf{x}}_t - \mathbf{x}_*)^T(-A^T\bar{\boldsymbol{\lambda}}_t) + (\bar{\mathbf{y}}_t - \mathbf{y}_*)^T(-B^T\bar{\boldsymbol{\lambda}}_t) + (\bar{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda})^T(A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}) \right\} \\ &= \max_{\boldsymbol{\lambda} \in \mathcal{B}_0} \left\{ \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \bar{\boldsymbol{\lambda}}_t^T(A\mathbf{x}_* + B\mathbf{y}_* - b) - \boldsymbol{\lambda}^T(A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}) \right\} \\ &= \max_{\boldsymbol{\lambda} \in \mathcal{B}_0} \left\{ \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) - \boldsymbol{\lambda}^T(A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}) \right\} \\ &= \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho \|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2 \end{aligned} \tag{33}$$

Taking an expectation over (33) and using (32) we have:

$$\begin{aligned}
&\mathbb{E}\left[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2\right] \\
&\leq \mathbb{E}\left[\frac{1}{t}\sum_{k=0}^{t-1}\left(\frac{\eta_{k+1}\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k\rangle\right) + \frac{1}{t}\left(\frac{D_{\mathcal{X}}^2}{2\eta_t} + \frac{\beta}{2}D_{\mathbf{y}_*,B}^2\right)\right] \\
&\quad + \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathcal{B}_0}\left\{\frac{1}{2\beta t}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_2^2\right\}\right] \\
&\leq \frac{1}{t}\left(\frac{M^2}{2}\sum_{k=1}^{t}\eta_k + \frac{D_{\mathcal{X}}^2}{2\eta_t}\right) + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k\rangle\right] \\
&= \frac{1}{t}\left(\frac{M^2}{2}\sum_{k=1}^{t}\eta_k + \frac{D_{\mathcal{X}}^2}{2\eta_t}\right) + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t} \\
&\leq \frac{\sqrt{2}D_{\mathcal{X}}M}{\sqrt{t}} + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t}
\end{aligned}$$

In the second last step, we use the fact that $\mathbf{x}_k$ is independent of $\boldsymbol{\xi}_{k+1}$, hence $\mathbb{E}_{\boldsymbol{\xi}_{k+1}|\boldsymbol{\xi}_{[1:k]}}\langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k\rangle = \left\langle \mathbb{E}_{\boldsymbol{\xi}_{k+1}|\boldsymbol{\xi}_{[1:k]}}\boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k\right\rangle = 0$.

(ii) From the steps in the proof of part (i), it follows that,

$$\begin{aligned}
&\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\| \\
&\leq \frac{1}{t}\sum_{k=0}^{t-1}\frac{\eta_{k+1}\ \|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \frac{1}{t}\sum_{k=0}^{t-1}\langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_* - \mathbf{x}_k\rangle + \frac{1}{t}\left(\frac{D_{\mathcal{X}}^2}{2\eta_t} + \frac{\beta}{2}D_{\mathbf{y}_*,B}^2 + \frac{\rho^2}{2\beta}\right) \qquad (34) \\
&\equiv A_t + B_t + C_t
\end{aligned}$$

Note that random variables A_t and B_t are dependent on $\boldsymbol{\xi}_{[t]}$.

Claim 1. *For* $\Omega_1 > 0$,

$$Prob\left(A_t \geq (1+\Omega_1)\frac{M^2}{2t}\sum_{k=1}^{t}\eta_k\right) \leq \exp\{-\Omega_1\}. \qquad (35)$$

Let $\alpha_k \equiv \frac{\eta_k}{\sum_{k=1}^{t}\eta_k}\ \forall k = 1, \ldots, t$, then $0 \leq \alpha_k \leq 1$ and $\sum_{k=1}^{t}\alpha_k = 1$. Using the fact that $\{\boldsymbol{\delta}_k, \forall k\}$ are independent and applying Assumption 2, one has

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\sum_{k=1}^{t}\alpha_k\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2/M^2\right\}\right] &= \prod_{k=1}^{t}\mathbb{E}\left[\exp\left\{\alpha_k\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2/M^2\right\}\right] \\
&\leq \prod_{k=1}^{t}\left(\mathbb{E}\left[\exp\left\{\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2/M^2\right\}\right]\right)^{\alpha_k} \qquad \text{(Jensen's Inequality)} \\
&\leq \prod_{k=1}^{t}(\exp\{1\})^{\alpha_k} = \exp\left\{\sum_{k=1}^{t}\alpha_k\right\} = \exp\{1\}
\end{aligned}$$

Hence, by Markov's Inequality, we can get

$$\text{Prob}\left(A_t \geq (1+\Omega_1)\frac{M^2}{2t}\sum_{k=1}^{t}\eta_k\right) \leq \exp\left\{-(1+\Omega_1)\right\}\mathbb{E}\left[\exp\left\{\sum_{k=1}^{t}\alpha_k\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2/M^2\right\}\right] \leq \exp\{-\Omega_1\}.$$

We have therefore proved Claim 1.

Claim 2. *For* $\Omega_2 > 0$,

$$Prob\left(B_t > 2\Omega_2\frac{D_{\mathcal{X}}M}{\sqrt{t}}\right) \leq \exp\left\{-\frac{\Omega_2^2}{4}\right\}. \qquad (36)$$

In order to prove this claim, we adopt the following facts in Nemirovski's paper (Nemirovski et al., 2009).

Lemma 3. *Given that for all $k = 1, \ldots, t$, ζ_k is a deterministic function of $\boldsymbol{\xi}_{[k]}$ with $\mathbb{E}\left[\zeta_k|\boldsymbol{\xi}_{[k-1]}\right] = 0$ and $\mathbb{E}\left[\exp\{\zeta_k^2/\sigma_k^2\}|\boldsymbol{\xi}_{[k-1]}\right] \leq \exp\{1\}$, we have*

(a) For $\gamma \geq 0$, $\mathbb{E}\left[\exp\{\gamma\zeta_k\}|\boldsymbol{\xi}_{[k-1]}\right] \leq \exp\{\gamma^2\sigma_k^2\}, \forall k = 1, \ldots, t$

(b) Let $S_t = \sum_{k=1}^t \zeta_k$, then Prob$\{S_t > \Omega\sqrt{\sum_{k=1}^t \sigma_k^2}\} \leq \exp\left\{-\frac{\Omega^2}{4}\right\}$.

Using this result by setting $\zeta_k = \langle \boldsymbol{\delta}_k, \mathbf{x}_* - \mathbf{x}_{k-1}\rangle$, $S_t = \sum_{k=1}^t \zeta_k$, and $\sigma_k = 2D_{\mathcal{X}}M, \forall k$, we can verify that $\mathbb{E}\left[\zeta_k|\boldsymbol{\xi}_{[k-1]}\right] = 0$ and

$$\mathbb{E}\left[\exp\{\zeta_k^2/\sigma_k^2\}|\boldsymbol{\xi}_{[k-1]}\right] \leq \mathbb{E}\left[\exp\{D_{\mathcal{X}}^2\|\boldsymbol{\delta}_k\|^2/\sigma_k^2\}|\boldsymbol{\xi}_{[k-1]}\right] \leq \exp\{1\},$$

since $|\zeta_k|^2 \leq \|\mathbf{x}_* - \mathbf{x}_{k-1}\|^2\|\boldsymbol{\delta}_k\|^2 \leq D_{\mathcal{X}}^2\left(2\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2 + 2M^2\right)$.

Implementing the above results, it follows that

$$\text{Prob}\left(S_t > 2\Omega_2 D_{\mathcal{X}} M\sqrt{t}\right) \leq \exp\left\{-\frac{\Omega_2^2}{4}\right\}.$$

Since $S_t = tB_t$, we have

$$\text{Prob}\left(B_t > 2\Omega_2\frac{D_{\mathcal{X}}M}{\sqrt{t}}\right) \leq \exp\left\{-\frac{\Omega_2^2}{4}\right\}$$

as desired.

Combining (34), (35) and (36), we obtain

$$\text{Prob}\left(\text{Err}_\rho(\bar{\mathbf{u}}_t) > (1+\Omega_1)\frac{M^2}{2t}\sum_{k=1}^t \eta_k + 2\Omega_2\frac{D_{\mathcal{X}}M}{\sqrt{t}} + C_t\right) \leq \exp\{-\Omega_1\} + \exp\left\{-\frac{\Omega_2}{4}\right\},$$

where $\text{Err}_\rho(\bar{\mathbf{u}}_t) \equiv \theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2$. Substituting $\Omega_1 = \Omega, \Omega_2 = 2\sqrt{\Omega}$ and plugging in $\eta_k = \frac{D_{\mathcal{X}}}{M\sqrt{2k}}$, we obtain (10) as desired. □

7.4. Proof of Theorem 2

Proof. By the strong-convexity of θ_1 we have $\forall \mathbf{x}$:

$$\begin{aligned}\theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) &\leq \left\langle \theta_1'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}\right\rangle - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \\ &= \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}\right\rangle + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k\rangle + \left\langle \theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1}\right\rangle - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_k\|^2.\end{aligned}$$

Following the same derivations as in Lemma 2 and Theorem 1 (i), we have

$$\begin{aligned}&\mathbb{E}\left[\theta(\bar{\mathbf{u}}_t) - \theta(\mathbf{u}_*) + \rho\|A\bar{\mathbf{x}}_t + B\bar{\mathbf{y}}_t - \mathbf{b}\|_2\right] \\ &\leq \mathbb{E}\left\{\frac{1}{t}\sum_{k=0}^{t-1}\left[\frac{\eta_{k+1}\|\theta_1'(\mathbf{x}_k, \boldsymbol{\xi}_{k+1})\|^2}{2} + \left(\frac{1}{2\eta_{k+1}} - \frac{\mu}{2}\right)\|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2}{2\eta_{k+1}}\right]\right\} \\ &+ \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \mathbb{E}\Big[\max_{\boldsymbol{\lambda}\in\mathcal{B}_0}\left\{\frac{1}{2\beta t}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0\|_0^2\right\}\Big] \\ &\leq \frac{M^2}{2t}\sum_{k=1}^t\frac{1}{\mu k} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\frac{\mu k}{2}\|\mathbf{x}_k - \mathbf{x}_*\|^2 - \frac{\mu(k+1)}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2\right] + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t} \\ &\leq \frac{M^2\log t}{\mu t} + \frac{\mu D_{\mathcal{X}}^2}{2t} + \frac{\beta D_{\mathbf{y}_*,B}^2}{2t} + \frac{\rho^2}{2\beta t}.\end{aligned}$$

□

7.5. Proof of Theorem 3

Proof. The Lipschitz smoothness of θ_1 implies that $\forall k \geq 0$:

$$\begin{aligned}\theta_1(\mathbf{x}_{k+1}) &\leq \theta_1(\mathbf{x}_k) + \langle\nabla\theta_1(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\overset{(3)}{=} \theta_1(\mathbf{x}_k) + \langle\nabla\theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle - \langle\boldsymbol{\delta}_{k+1}, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.\end{aligned}$$

It follows that $\forall \mathbf{x} \in \mathcal{X}$:

$$
\begin{aligned}
&\theta_1(\mathbf{x}_{k+1}) - \theta_1(\mathbf{x}) + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \\
&\leq \theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) + \langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle - \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \\
&= \theta_1(\mathbf{x}_k) - \theta_1(\mathbf{x}) + \langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x} - \mathbf{x}_k \rangle - \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&\quad + \left[\langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x} \rangle + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \right] \\
&\leq \langle \nabla \theta_1(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x} \rangle + \langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x} - \mathbf{x}_k \rangle - \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&\quad + \left[\langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x} \rangle + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \right] \\
&= \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_{k+1} \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \left[\langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x} \rangle + \left\langle \mathbf{x}_{k+1} - \mathbf{x}, -A^T \boldsymbol{\lambda}_{k+1} \right\rangle \right] \\
&= \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_{k+1} \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \left\langle \mathbf{x} - \mathbf{x}_{k+1}, \beta A^T B(\mathbf{y}_k - \mathbf{y}_{k+1}) \right\rangle \\
&\quad + \left\langle \nabla \theta_1(\mathbf{x}_k, \boldsymbol{\xi}_{k+1}) + A^T \left[\beta(A\mathbf{x}_{\mathbf{k+1}} + B\mathbf{y}_k - \mathbf{b}) - \boldsymbol{\lambda}_k \right], \mathbf{x}_{k+1} - \mathbf{x} \right\rangle \\
&\overset{(24)}{\leq} \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x} - \mathbf{x}_k\|^2 - \|\mathbf{x} - \mathbf{x}_{k+1}\|^2 \right) - \frac{1/\eta_{k+1} - L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&\quad + \left\langle \mathbf{x} - \mathbf{x}_{k+1}, \beta A^T B(\mathbf{y}_k - \mathbf{y}_{k+1}) \right\rangle + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_{k+1} \rangle .
\end{aligned}
$$

The last inner product can be bounded as below using Young's inequality, given that $\eta_{k+1} \leq \frac{1}{L}$:

$$
\begin{aligned}
\langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_{k+1} \rangle &= \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \rangle + \langle \boldsymbol{\delta}_{k+1}, \mathbf{x}_k - \mathbf{x}_{k+1} \rangle \\
&\leq \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\left(1/\eta_{k+1} - L\right)} \|\boldsymbol{\delta}_{k+1}\|^2 + \frac{1/\eta_{k+1} - L}{2} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 .
\end{aligned}
$$

Combining this with inequalities (26,29) and (30), we can get a similar statement as that of Lemma 2:

$$
\begin{aligned}
&\theta(\mathbf{u}_{k+1}) - \theta(\mathbf{u}) + (\mathbf{w}_{k+1} - \mathbf{w})^T F(\mathbf{w}_{k+1}) \leq \frac{\|\boldsymbol{\delta}_{k+1}\|^2}{2(1/\eta_{k+1} - L)} \\
&+ \frac{1}{2\eta_{k+1}} \left(\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 \right) + \frac{\beta}{2} \left(\|A\mathbf{x} + B\mathbf{y}_k - \mathbf{b}\|^2 - \|A\mathbf{x} + B\mathbf{y}_{k+1} - \mathbf{b}\|^2 \right) \\
&+ \langle \boldsymbol{\delta}_{k+1}, \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2\beta} \left(\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_k\|_2^2 - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_{k+1}\|_2^2 \right) .
\end{aligned}
$$

The rest of the proof are essentially the same as Theorem 1 (i), except that we use the new definition of $\bar{\mathbf{u}}_k$ in (12). □