Generalized ADMM And Its Convergence

Chen Feng

November 12, 2012

Alternating direction method of multipliers (ADMM) is a simple yet powerful algorithm of solving large-scale convex optimization problems. Though developed in the 1970s [1], ADMM has received renewed interest recently, and found practical use in many large-scale distributed convex optimization problems in statistics, machine learning, and related areas [2].

ADMM solves problems in the form

$$\begin{array}{ll} \min & f_1(x_1) + f_2(x_2) \\ \text{s.t.} & A_1 x_1 + A_2 x_2 = b \end{array} \tag{1}$$

with variables $x_i \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, where $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ $(i = 1, 2)$ are closed proper convex functions; $A_i \in \mathbb{R}^{l \times n_i}$ $(i = 1, 2)$ are given matrices; and $b \in \mathbb{R}^l$ is a given vector. Note that the objective function is *separable* over *two* sets of variables, which are coupled through an equality constraint.

With the recent popularity of ADMM, it is quite natural to ask the following question: **Can we extend the ADMM algorithm from two sets of variables to multiple sets of variables?** Surprisingly, little is known about such an extension, with two exceptions [3, 4] very recently. [3] establishes the convergence of m-block ($m \geq 3$) ADMM for strongly convex objective functions, but not linear convergence; [4] shows the linear convergence of m-block ADMM, but under the assumption that the relation matrices (i.e., $A_1, A_2, ...$) are full column rank.

In this article, I will show that, by replacing the full-rank assumption on the relation matrices with the strong convexity assumption on the objective function, one can obtain the same convergence and rate of convergence result. The motivation of this result is as follows: It is difficult, if not impossible, to satisfy the full-rank assumption in practice, especially when l is smaller than n_i's. The proof techniques for this result are mainly based on [3, 4].

1 Algorithm

We consider a convex optimization problem in the form

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} f_i(x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{m} A_i x_i = b \end{aligned} \tag{2}$$

with variables $x_i \in \mathbb{R}^{n_i}$ $(i = 1, \ldots, m)$, where $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ $(i = 1, \ldots, m)$ are closed proper convex functions; $A_i \in \mathbb{R}^{l \times n_i}$ $(i = 1, \ldots, m)$ are given matrices; and $b \in \mathbb{R}^l$ is a given vector.

We form the augmented Lagrangian

$$L_\rho(x_1, \ldots, x_m; y) = \sum_{i=1}^{m} f_i(x_i) + y^T(\sum_{i=1}^{m} A_i x_i - b) + (\rho/2)\| \sum_{i=1}^{m} A_i x_i - b\|_2^2. \tag{3}$$

A generalized ADMM algorithm consists of the iterations

$$\begin{aligned} x_i^{k+1} &= \arg\min_{x_i} L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k; y^k), \ i = 1, \ldots, m, \\ y^{k+1} &= y^k + \alpha(\sum_{i=1}^{m} A_i x_i^{k+1} - b), \end{aligned}$$

where $\alpha > 0$ is the step size for the dual update. Note that when $m = 2$ and the step size $\alpha = \rho$, the above algorithm is reduced to the standard ADMM algorithm presented in [2].

For comparison, the method of multipliers for (2) has the form

$$\begin{aligned} (\bar{x}_1^{k+1}, \ldots, \bar{x}_m^{k+1}) &= \arg\min_{x_1, \ldots, x_m} L_\rho(x_1, \ldots, x_m; y^k), \\ y^{k+1} &= y^k + \alpha(\sum_{i=1}^{m} A_i \bar{x}_i^{k+1} - b). \end{aligned} \tag{4}$$

Here, the augmented Lagrangian is minimized jointly rather than sequentially.

2 Assumptions

In this section, we discuss several assumptions and their implications, which are needed for the convergence analysis. For convenience, we write

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \ f(x) = \sum_{i=1}^{m} f_i(x_i), \text{ and } A = [A_1 \ \ldots \ A_m].$$

Clearly, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{l \times n}$, where $n = \sum_{i=1}^{m} n_i$. Now, the problem (2) can be rewritten as

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

with the optimal value denoted by

$$p^* = \inf\{f(x) \mid Ax = b\}.$$

Similarly, the augmented Lagrangian can be rewritten as

$$L_\rho(x; y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|_2^2,$$

with the associated dual function defined by

$$d(y) = \inf_x L_\rho(x; y).$$

The dual problem is

$$\max \quad d(y)$$

with the optimal value

$$d^* = \sup\{d(y)\}.$$

Assumption 1. *The unaugmented Lagrangian L_0 has a saddle point.*

Explicitly, there exist (x^*, y^*), not necessarily unique, for which

$$L_0(x^*; y) \le L_0(x^*; y^*) \le L_0(x; y^*)$$

holds for all x, y.

Assumption 1 implies that x^* is primal optimal, y^* is dual optimal, and the optimal duality gap is zero, *i.e.*, $p^* = d^*$.

When Assumption 1 fails to hold, some subproblems in the generalized ADMM algorithm are either unsolvable or unbounded, or the sequence $\{y^k\}$ in the algorithm diverges.

Assumption 2. *The functions f_i $(i = 1, \ldots, m)$ are strongly convex.*

A function $f : \mathbb{R}^n \to R$ is *strongly convex* with constant $\nu > 0$, if for all $x_1, x_2 \in \mathbb{R}^n$, and all $\theta \in [0, 1]$,

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) - \frac{1}{2}\nu\theta(1-\theta)\|x_1 - x_2\|_2^2.$$

When f is differentiable, f is strongly convex with constant ν if and only if

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^T(x_1 - x_2) + \frac{\nu}{2}\|x_1 - x_2\|_2^2 \tag{5}$$

holds for all $x_1, x_2 \in \mathbb{R}^n$. Thus, (5) can be viewed as a first-order condition for strong convexity.

When f is twice continuously differentiable, a second-order condition for strong convexity is the following

$$\nabla^2 f(x) \succeq \nu I_n, \ \forall x \in \mathbb{R}^n.$$

Here, $\nabla^2 f$ denotes the Hessian matrix of f, I_n denotes the $n \times n$ identify matrix, and the inequality $\succeq$ means that $\nabla^2 f(x) - \nu I_n$ is positive semi-definite.

Assumption 3. *The gradients ∇f_i $(i = 1, \ldots, m)$ are Lipschitz continuous.*

That is, for each i, there exists some constant $\kappa_i > 0$ such that for all $x_1, x_2 \in \mathbb{R}^{n_i}$,

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\|_2 \leq \kappa_i \|x_1 - x_2\|_2.$$

3 Convergence

In this section, we prove the convergence of the generalized ADMM algorithm. Define the primal and dual optimality gaps as

$$\begin{aligned}\Delta_p^k &= L_\rho(x^{k+1}; y^k) - d(y^k),\\ \Delta_d^k &= d^* - d(y^k),\end{aligned}$$

respectively. By the definition of $d(y)$, $\Delta_p^k \geq 0$. Similarly, $\Delta_d^k \geq 0$. Define

$$V^k = \Delta_p^k + \Delta_d^k.$$

We will see that V^k is a *Lyapunov function* for the algorithm, *i.e.*, a nonnegative quantity that decreases in each iteration.

We first outline the main idea of the convergence analysis. The proof relies on three key inequalities. The first inequality is

$$V^k \leq V^{k-1} - \alpha \|A\bar{x}^{k+1} - b\|_2^2 - \beta \|x^{k+1} - x^k\|_2^2, \tag{6}$$

for some constant $\beta > 0$, where $\bar{x}^{k+1}$ is defined in (4).

This states that V^k decreases in each iteration by an amount that depends on the norm of $A\bar{x}^{k+1} - b$ and on the change in each x_i over one iteration. Iterating the inequalities above gives that

$$\sum_{k=0}^{\infty} \left(\alpha \|A\bar{x}^{k+1} - b\|_2^2 + \beta \|x^{k+1} - x^k\|_2^2\right) \leq V^0,$$

which implies that

$$\|A\bar{x}^{k+1} - b\|_2^2 \to 0, \text{ and } \|x^{k+1} - x^k\|_2^2 \to 0,$$

as $k \to \infty$.

Suppose that the level set of $\Delta_p + \Delta_d$ is bounded, *i.e.*,

$$\delta = \sup\{\|x\| + \|y\| \mid (L_\rho(x; y) - d(y)) + (d^* - d(y)) \leq V^0\} < \infty. \tag{7}$$

Since $V^k \leq V^0$ for all k, it follows that the sequence $\{x^{k+1}, y^k\}$ in the generalized ADMM algorithm is bounded by $\|x^{k+1}\| + \|y^k\| \leq \delta$, for all k. In particular, this implies that $\|x^k\|, \|y^k\| \leq \delta$ for all k. By the Bolzano-Weierstrass theorem, the sequence $\{x^k, y^k\}$ has a convergent subsequence, *i.e.*,

$$\lim_{k \in \mathcal{R}, k \to \infty} (x^k, y^k) = (\tilde{x}, \tilde{y}),$$

for some subsequence $\mathcal{R}$, where $(\tilde{x}, \tilde{y})$ denotes the limit point. It is natural to expect that $\tilde{x}$ is primal optimal and $\tilde{y}$ is dual optimal. We will see that this is indeed the case.

The second key inequality says that there exists a constant $\tau > 0$ such that for any (x, y) satisfying $\|x\| + \|y\| \leq 2\delta$, the following Hoffman-like error bound holds

$$\|x - \bar{x}(y)\| \leq \tau \|\nabla_x L_\rho(x; y)\|, \tag{8}$$

where $\bar{x}(y) = \arg\min_x L_\rho(x; y)$.

In particular, inequality (8) implies that

$$\|x^k - \bar{x}^{k+1}\|_2^2 \leq \tau^2 \|\nabla_x L_\rho(x^k; y^k)\|_2^2. \tag{9}$$

This is because $\|x^k\|, \|y^k\| \leq \delta$ for all k, and $\bar{x}^{k+1}$ minimizes $L_\rho(x; y^k)$.

The third inequality is

$$\|\nabla_x L_\rho(x^k; y^k)\|_2 \leq \eta \|x^k - x^{k+1}\|_2. \tag{10}$$

Thus, we have

$$\|x^k - \bar{x}^{k+1}\|_2^2 \leq \tau^2 \eta^2 \|x^k - x^{k+1}\|_2^2.$$

It follows that

$$\lim_{k \in \mathcal{R}, k \to \infty} \|x^k - \bar{x}^{k+1}\| \to 0.$$

This further implies that the subsequence $\{\bar{x}^{k+1}\}_{k \in \mathcal{R}}$ converges to $\tilde{x}$, as $k \to \infty$. Since $\|A\bar{x}^{k+1} - b\|_2 \to 0$, we have $\|A\tilde{x} - b\| = 0$, or equivalently, $A\tilde{x} - b = 0$.

Now we are ready to show that $\tilde{x}$ is primal optimal and $\tilde{y}$ is dual optimal. Note that

$$\begin{aligned} d^* - d(y^k) &= p^* - L_\rho(\bar{x}^{k+1}; y^k) && (11) \\ &= p^* - f(\bar{x}^{k+1}) - (y^k)^T (A\bar{x}^{k+1} - b) - (\rho/2)\|A\bar{x}^{k+1} - b\|_2^2 && (12) \end{aligned}$$

where (11) follows from Assumption 1 and the fact that $\bar{x}^{k+1}$ minimizes $L_\rho(x; y^k)$. By taking limit in (12) along the subsequence $\mathcal{R}$, we obtain

$$\begin{aligned} d^* - d(\tilde{y}) &= p^* - f(\tilde{x}) - \tilde{y}^T (A\tilde{x} - b) - (\rho/2)\|A\tilde{x} - b\|_2^2 \\ &= p^* - f(\tilde{x}), \end{aligned}$$

where the last equality follows from the fact that $A\tilde{x} - b = 0$.

Recall that $d^* \geq d(\tilde{y})$, and $p^* \leq f(\tilde{x})$ when $A\tilde{x} - b = 0$. Thus, we have $d^* = d(\tilde{y})$ and $p^* = f(\tilde{x})$. That is, for each convergent subsequence $\{x^k, y^k\}_{k \in \mathcal{R}}$, the associate limit point $(\tilde{x}, \tilde{y})$ is an optimal primal-dual solution.

Next, we show that $V^k = \Delta_p^k + \Delta_d^k \to 0$, as $k \to \infty$. On the one hand, we have

$$\lim_{k \in \mathcal{R}, k \to \infty} \Delta_d^k = d^* - d(\tilde{y}) = 0.$$

On the other hand, we have

$$\begin{aligned} \lim_{k \in \mathcal{R}, k \to \infty} \Delta_p^k &\leq \lim_{k \in \mathcal{R}, k \to \infty} L_\rho(x^k; y^k) - d(y^k) \\ &= L_\rho(\tilde{x}; \tilde{y}) - d(\tilde{y}) \\ &= p^* - d^* \\ &= 0. \end{aligned}$$

Since $\Delta_p^k \geq 0$ for all k, we conclude that

$$\lim_{k\in\mathcal{R},k\to\infty} \Delta_p^k = 0.$$

Thus, we have

$$\lim_{k\in\mathcal{R},k\to\infty} \Delta_p^k + \Delta_d^k = 0.$$

Recall that V^k decreases in each iteration, *i.e.*, $V^{k+1} \leq V^k$ for all k. Thus, the convergence of a subsequence of V^k implies the convergence of V^k, and we have

$$\lim_{k\to\infty} \Delta_p^k + \Delta_d^k = 0.$$

This further implies that both Δ_p^k and Δ_d^k converge to 0, *i.e.*,

$$\lim_{k\to\infty} \Delta_p^k = \lim_{k\to\infty} \Delta_d^k = 0.$$

To sum up, we have the following theorem that describes the convergence of the generalized ADMM algorithm.

Theorem 1 *Suppose Assumptions 1, 2, 3 hold and that the level set of* $\Delta_p + \Delta_d$ *is bounded. Then both the primal gap* Δ_p^k *and the dual gap* Δ_d^k *converge to* 0. *Moreover, the sequence* $\{x^k, y^k\}$ *has a convergent subsequence, and any convergent subsequence of* $\{x^k, y^k\}$ *converges to an optimal primal-dual solution for the problem (2).*

Remark 1: Although $\Delta_p^k + \Delta_d^k$ decreases in each iteration, there is no guarantee that the primal gap Δ_p^k is reduced in each iteration. The same applies to the dual gap Δ_d^k as well.

Remark 2: By some additional argument, we can show that the sequence $\{\Delta_p^k + \Delta_d^k\}$ converges to zero Q-linearly, and that both Δ_p^k and Δ_d^k converge to zero R-linearly[1]. The key step is to show that $\{\Delta_p^k + \Delta_d^k\}$ contracts geometrically, *i.e.*,

$$\Delta_p^{k+1} + \Delta_d^{k+1} \leq \mu(\Delta_p^k + \Delta_d^k)$$

[1]Suppose a sequence $\{u^k\}$ converges to $\tilde{u}$. We say the convergence is (in some norm $\|\cdot\|$)

- Q-linear, if there exists $\mu \in (0,1)$ such that $\|u^{k+1} - \tilde{u}\| \leq \mu\|u^k - \tilde{u}\|$;
- R-linear, if there exists a sequence $\{\sigma^k\}$ such that $\|u^k - \tilde{u}\| \leq \sigma^k$ and $\sigma^k \to 0$ Q-linearly.

for some $\mu \in (0,1)$. Since similar arguments can be found in [4], we omit the proof for the linear convergences here.

3.1 Proof of inequality (10)

Taking ∇_{x_i} on both sides of (3), we get

$$\nabla_{x_i} L_\rho(x^k; y^k) = \nabla f_i(x_i^k) + A_i^T y^k + \rho A_i^T \left(\sum_{i=1}^{m} A_i x_i^k - b \right).$$

Recall that x_i^{k+1} minimizes $L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k; y^k)$, so we have

$$0 = \nabla f_i(x_i^{k+1}) + A_i^T y^k + \rho A_i^T \left(\sum_{j=1}^{i} A_j x_j^{k+1} + \sum_{j=i+1}^{m} A_j x_j^k - b \right).$$

Combining the two equalities above, we obtain

$$\nabla_{x_i} L_\rho(x^k; y^k) = \nabla f_i(x_i^k) - \nabla f_i(x_i^{k+1}) + \rho A_i^T \left(\sum_{j=1}^{i} A_j (x_j^k - x_j^{k+1}) \right).$$

This implies that

$$\begin{aligned} \|\nabla_{x_i} L_\rho(x^k; y^k)\|_2 &\leq \|\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)\|_2 + \sum_{j=1}^{i} \|\rho A_i^T A_j (x_j^k - x_j^{k+1})\|_2 \\ &\leq \|\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)\|_2 + \sum_{j=1}^{i} \|\rho A_i^T A_j\|_2 \|x_j^k - x_j^{k+1}\|_2, \end{aligned} \tag{13}$$

where the last inequality follows from the definition of the matrix norm.

Recall that, by Assumption 3, we have

$$\|\nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k)\|_2 \leq \kappa_i \|x_i^k - x_i^{k+1}\|_2.$$

Substituting this in (13) gives

$$\|\nabla_{x_i} L_\rho(x^k; y^k)\|_2 \leq \kappa_i \|x_i^k - x_i^{k+1}\|_2 + \sum_{j=1}^{i} \|\rho A_i^T A_j\|_2 \|x_j^k - x_j^{k+1}\|_2.$$

Since $\|x_j^k - x_j^{k+1}\|_2 \le \|x^k - x^{k+1}\|_2$ for all j, we have

$$\|\nabla_{x_i} L_\rho(x^k; y^k)\|_2 \le \theta \|x^k - x^{k+1}\|_2$$

for some $\theta > 0$. In particular, if we choose

$$\theta = \max_i \left\{ \kappa_i + \sum_{j=1}^{i} \|\rho A_i^T A_j\|_2 \right\},$$

then $\|\nabla_{x_i} L_\rho(x^k; y^k)\|_2 \le \theta \|x^k - x^{k+1}\|_2$ for all i, which implies that

$$\|\nabla_x L_\rho(x^k; y^k)\|_2^2 \le \sum_{i=1}^{m} \|\nabla_{x_i} L_\rho(x^k; y^k)\|_2^2 \le m\theta^2 \|x^k - x^{k+1}\|_2^2.$$

3.2 Proof of inequality (8)

This inequality is proved in Lemma 2.2 under three assumptions in pp. 5 of [4]. Since these assumptions are valid in our setup, we omit the detailed proof here.

3.3 Proof of inequality (6)

We first introduce two lemmas that bound the changes in Δ_d^k and Δ_p^k over one iteration.

Lemma 1

$$\Delta_d^k - \Delta_d^{k-1} \le -\alpha (Ax^k - b)^T (A\bar{x}^{k+1} - b). \tag{14}$$

Proof: By definition, $\bar{x}^{k+1}$ minimizes $L_\rho(x; y^k)$, *i.e.*,

$$L_\rho(\bar{x}^{k+1}; y^k) = d(y^k).$$

Thus, we have

$$\begin{aligned}
\Delta_d^k - \Delta_d^{k-1} &= \big(d^* - d(y^k)\big) - \big(d^* - d(y^{k-1})\big) \\
&= d(y^{k-1}) - d(y^k) \\
&= L_\rho(\bar{x}^k; y^{k-1}) - L_\rho(\bar{x}^{k+1}; y^k) \\
&= \big(L_\rho(\bar{x}^{k+1}; y^{k-1}) - L_\rho(\bar{x}^{k+1}; y^k)\big) + \big(L_\rho(\bar{x}^k; y^{k-1}) - L_\rho(\bar{x}^{k+1}; y^{k-1})\big) \\
&= (y^{k-1} - y^k)^T (A\bar{x}^{k+1} - b) + \big(L_\rho(\bar{x}^k; y^{k-1}) - L_\rho(\bar{x}^{k+1}; y^{k-1})\big) \\
&= -\alpha (Ax^k - b)^T (A\bar{x}^{k+1} - b) + \big(L_\rho(\bar{x}^k; y^{k-1}) - L_\rho(\bar{x}^{k+1}; y^{k-1})\big) \\
&\le -\alpha (Ax^k - b)^T (A\bar{x}^{k+1} - b),
\end{aligned}$$

where the last inequality follows from (4). □

Lemma 2

$$\Delta_p^k - \Delta_p^{k-1} \leq \alpha\|Ax^k - b\|_2^2 - \gamma\|x^{k+1} - x^k\|_2^2 - \alpha(Ax^k - b)^T(A\bar{x}^{k+1} - b). \quad (15)$$

Proof: First, we have

$$\begin{aligned}\Delta_p^k - \Delta_p^{k-1} &= \big(L_\rho(x^{k+1}; y^k) - d(y^k)\big) - \big(L_\rho(x^k; y^{k-1}) - d(y^{k-1})\big) \\ &= \big(L_\rho(x^{k+1}; y^k) - L_\rho(x^k; y^{k-1})\big) + \big(d(y^{k-1}) - d(y^k)\big) \\ &\leq \big(L_\rho(x^{k+1}; y^k) - L_\rho(x^k; y^{k-1})\big) - \alpha(Ax^k - b)^T(A\bar{x}^{k+1} - b),\end{aligned} \quad (16)$$

where the last inequality follows from (14).

We next bound the term $L_\rho(x^{k+1}; y^k) - L_\rho(x^k; y^{k-1})$. Note that

$$\begin{aligned}L_\rho(x^k; y^k) - L_\rho(x^k; y^{k-1}) &= (y^k - y^{k-1})^T(Ax^k - b) \\ &= (\alpha(Ax^k - b))^T(Ax^k - b) \\ &= \alpha\|Ax^k - b\|_2^2.\end{aligned} \quad (17)$$

Note also that, by Assumption 2, the augmented Lagrangian

$$L_\rho(x_1, \ldots, x_m; y) = \sum_{i=1}^m f_i(x_i) + y^T(\sum_{i=1}^m A_i x_i - b) + (\rho/2)\|\sum_{i=1}^m A_i x_i - b\|_2^2$$

is strongly convex over each variable x_i, as the sum of a strongly convex function and a convex function is strongly convex. Thus, we have

$$\begin{aligned}&L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_m^k; y^k) \\ &- L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k; y^k) \\ &\geq \nabla_{x_i} L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k; y^k) + \frac{\nu_i}{2}\|x_i^k - x_i^{k+1}\|_2^2 \\ &= \frac{\nu_i}{2}\|x_i^k - x_i^{k+1}\|_2^2,\end{aligned}$$

for $i = 1, \ldots, m$, where the last equality follows from the fact that x_i^{k+1} minimizes $L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k; y^k)$.

Adding all the inequalities above together, we obtain

$$L_\rho(x^k; y^k) - L_\rho(x^{k+1}; y^k) \geq \sum_{i=1}^{m} \frac{\nu_i}{2} \|x_i^k - x_i^{k+1}\|_2^2.$$

If we choose $\gamma = \min_i\{\nu_i/2\}$, then we have

$$L_\rho(x^k; y^k) - L_\rho(x^{k+1}; y^k) \geq \gamma \|x^k - x^{k+1}\|_2^2,$$

or equivalently,

$$L_\rho(x^{k+1}; y^k) - L_\rho(x^k; y^k) \leq -\gamma \|x^{k+1} - x^k\|_2^2. \tag{18}$$

Adding (17) and (18), we obtain

$$L_\rho(x^k; y^k) - L_\rho(x^k; y^{k-1}) \leq \alpha \|Ax^k - b\|_2^2 - \gamma \|x^{k+1} - x^k\|_2^2.$$

Substituting this in the first term of (16), we get

$$\Delta_p^k - \Delta_p^{k-1} \leq \alpha \|Ax^k - b\|_2^2 - \gamma \|x^{k+1} - x^k\|_2^2 - \alpha (Ax^k - b)^T (A\bar{x}^{k+1} - b). \square$$

Now we are ready to prove the inequality (6). Adding (14) and (15) gives

$$V^k - V^{k-1} \leq \alpha \|Ax^k - b\|_2^2 - \gamma \|x^{k+1} - x^k\|_2^2 - 2\alpha (Ax^k - b)^T (A\bar{x}^{k+1} - b). \tag{19}$$

Since $Ax^k - A\bar{x}^{k+1} = (Ax^k - b) - (A\bar{x}^{k+1} - b)$, we have

$$\begin{aligned} \|Ax^k - A\bar{x}^{k+1}\|_2^2 &= \|(Ax^k - b) - (A\bar{x}^{k+1} - b)\|_2^2 \\ &= \|Ax^k - b\|_2^2 - 2(Ax^k - b)^T (A\bar{x}^{k+1} - b) + \|A\bar{x}^{k+1} - b\|_2^2. \end{aligned}$$

Substituting this in (19) yields

$$V^k - V^{k-1} \leq \alpha \|Ax^k - A\bar{x}^{k+1}\|_2^2 - \alpha \|A\bar{x}^{k+1} - b\|_2^2 - \gamma \|x^{k+1} - x^k\|_2^2. \tag{20}$$

Note that

$$\begin{aligned} \|Ax^k - A\bar{x}^{k+1}\|_2^2 &\leq \|A\|_2^2 \|x^k - \bar{x}^{k+1}\|_2^2 && (21) \\ &\leq \tau^2 \|A\|_2^2 \|\nabla_x L_\rho(x^k; y^k)\|_2^2 && (22) \\ &\leq \tau^2 \eta^2 \|A\|_2^2 \|x^k - x^{k+1}\|_2^2, && (23) \end{aligned}$$

where (21) follows form the definition of the matrix norm, (22) follows form (9), and (23) follows from (10).

Substituting this in the first term of (20), we obtain

$$\begin{aligned} V^k - V^{k-1} &\leq (\alpha \tau^2 \eta^2 \|A\|_2^2 - \gamma) \|x^{k+1} - x^k\|_2^2 - \alpha \|A\bar{x}^{k+1} - b\|_2^2 \\ &= -\alpha \|A\bar{x}^{k+1} - b\|_2^2 - \beta \|x^{k+1} - x^k\|_2^2, \end{aligned}$$

where $\beta = \gamma - \alpha \tau^2 \eta^2 \|A\|_2^2$. Note that, when the stepsize α is small enough, we have $\beta > 0$, which gives (6).

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

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

[3] Deren Han and Xiaoming Yuan. A note on the alternating direction method of multipliers. *J. Optim. Theory Appl.*, 155:227–238, 2012.

[4] Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers, August 2012.