Chapter 6
ADMM for Distributed Optimization

In this chapter, we introduce the application of ADMM to distributed optimization. We first introduce how to use ADMM, linearized ADMM, and accelerated linearized ADMM to centralized distributed optimization, and give the corresponding convergence rates. Then, we focus on decentralized distributed optimization and show that the corresponding ADMM is equivalent to the linearized augmented Lagrangian method, and give its accelerated version. Next, we introduce the asynchronous ADMM. At last, we end this chapter by the nonconvex and the generally linearly constrained distributed ADMM.

Consider the following problem in a distributed environment:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \equiv \sum_{i=1}^{m} f_i(\mathbf{x}), \tag{6.1}$$

where m agents form a connected and undirected network and the local function f_i is only accessible by agent i due to storage or privacy reasons. We consider two kinds of networks. The first one is the centralized network with one centralized master agent and m worker agents. Each worker agent is connected to the master agent. We will introduce this kind of network in Sect. 6.1. The second one is the decentralized network, which does not have the centralized agent and each agent only communicates with its neighbors. This kind of network will be introduced in Sect. 6.2. All the agents cooperate to solve Problem (6.1).

Z. Lin et al., *Alternating Direction Method of Multipliers for Machine Learning*,
https://doi.org/10.1007/978-981-16-9840-8_6

6.1 Centralized Optimization

In the centralized network, we reformulate Problem (6.1) as the following linearly constrained one:

$$\begin{aligned}
&\min_{\{\mathbf{x}_i\},\mathbf{z}} \sum_{i=1}^{m} f_i(\mathbf{x}_i),\\
&s.t. \ \ \mathbf{x}_i = \mathbf{z}, \quad i \in [m],
\end{aligned} \tag{6.2}$$

so that we can use the ADMM type methods to solve it.

6.1.1 ADMM

Introduce the augmented Lagrangian function

$$L(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda}) = \sum_{i=1}^{m} \left(f_i(\mathbf{x}_i) + \langle \boldsymbol{\lambda}_i, \mathbf{x}_i - \mathbf{z} \rangle + \frac{\beta}{2} \|\mathbf{x}_i - \mathbf{z}\|^2 \right). \tag{6.3}$$

ADMM can be used to solve problem (6.2) with the following iterations (for example, see [2, 3]):

$$\begin{aligned}
\mathbf{z}^{k+1} =& \operatorname*{argmin}_{\mathbf{z}} \sum_{i=1}^{m} \left(\left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i^k - \mathbf{z} \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i^k - \mathbf{z} \right\|^2 \right)\\
=& \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{x}_i^k + \frac{1}{\beta} \boldsymbol{\lambda}_i^k \right),
\end{aligned} \tag{6.4a}$$

$$\begin{aligned}
\mathbf{x}_i^{k+1} =& \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i - \mathbf{z}^{k+1} \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i - \mathbf{z}^{k+1} \right\|^2 \right)\\
=& \operatorname{Prox}_{\beta^{-1} f_i} \left(\mathbf{z}^{k+1} - \frac{1}{\beta} \boldsymbol{\lambda}_i^k \right), \quad i \in [m],
\end{aligned} \tag{6.4b}$$

$$\boldsymbol{\lambda}_i^{k+1} = \boldsymbol{\lambda}_i^k + \beta \left(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1} \right), \quad i \in [m]. \tag{6.4c}$$

In the above method, the master agent is responsible for updating $\mathbf{z}$ while each worker agent is responsible for $\mathbf{x}_i$ and $\boldsymbol{\lambda}_i$. Steps (6.4b) and (6.4c) are carried out independently at each worker agent, while step (6.4a) is performed at the master agent. At each iteration, the master agent collects $\mathbf{x}_i^k$ and $\boldsymbol{\lambda}_i^k$ from each worker agent, computes the average, and sends $\mathbf{z}^{k+1}$ back to each worker agent. Then each worker agent computes $\mathbf{x}_i^{k+1}$ and $\boldsymbol{\lambda}_i^{k+1}$ in parallel. We present the above method in Algorithms 6.1 and 6.2.

Algorithm 6.1 Centralized ADMM of the master

for $k = 0, 1, 2, \cdots$ **do**
  Wait until receiving $\mathbf{x}_i^k$ and $\boldsymbol{\lambda}_i^k$ from all the workers $i \in [m]$.
  Update $\mathbf{z}^{k+1}$ by (6.4a).
  Send $\mathbf{z}^{k+1}$ to all the workers.
end for

Algorithm 6.2 Centralized ADMM of the ith worker

Initialize: $\mathbf{x}_i^0, \boldsymbol{\lambda}_i^0, i \in [m]$.
for $k = 0, 1, 2, \cdots$ **do**
  Send $(\mathbf{x}_i^k, \boldsymbol{\lambda}_i^k)$ to the master.
  Wait until receiving $\mathbf{z}^{k+1}$ from the master.
  Update $\mathbf{x}_i^{k+1}$ and $\boldsymbol{\lambda}_i^{k+1}$ by (6.4b) and (6.4c), respectively.
end for

Now we discuss the convergence of Algorithms 6.1–6.2. Denote

$$(\mathbf{x}_1^*, \cdots, \mathbf{x}_m^*, \mathbf{z}^*, \boldsymbol{\lambda}_1^*, \cdots, \boldsymbol{\lambda}_m^*)$$

as a KKT point of Problem (6.2). From Theorem 3.3 we have the following convergence result.

Theorem 6.1 *Suppose that each $f_i(\mathbf{x}_i)$ is convex, $i \in [m]$. Then for Algorithms 6.1–6.2, we have*

$$\left| \sum_{i=1}^m f_i(\hat{\mathbf{x}}_i^{K+1}) - \sum_{i=1}^m f_i(\mathbf{x}_i^*) \right| \le \frac{C}{2(K+1)} + \frac{2\sqrt{C}\sqrt{\sum_{i=1}^m \|\boldsymbol{\lambda}_i^*\|^2}}{\sqrt{\beta}(K+1)},$$

$$\sqrt{\sum_{i=1}^m \|\hat{\mathbf{x}}_i^{K+1} - \hat{\mathbf{z}}^{K+1}\|^2} \le \frac{2\sqrt{C}}{\sqrt{\beta}(K+1)},$$

where

$$\hat{\mathbf{x}}_i^{K+1} = \frac{1}{K+1} \sum_{k=1}^{K+1} \mathbf{x}_i^k, \quad i \in [m], \qquad \hat{\mathbf{z}}^{K+1} = \frac{1}{K+1} \sum_{k=1}^{K+1} \mathbf{z}^k, \textit{ and}$$

$$C = \frac{1}{\beta} \sum_{i=1}^m \|\boldsymbol{\lambda}_i^0 - \boldsymbol{\lambda}_i^*\|^2 + \beta \sum_{i=1}^m \|\mathbf{x}_i^0 - \mathbf{x}_i^*\|^2.$$

Proof Algorithms 6.1–6.2 are a direct application of the original ADMM (Algorithm 2.1) to Problem (6.2) by setting

$$\mathbf{x} = \mathbf{z}, \quad \mathbf{y} = (\mathbf{x}_1^T, \cdots, \mathbf{x}_m^T)^T, \quad \mathbf{A} = \mathbf{1}_m \otimes \mathbf{I}_d, \quad \mathbf{B} = -\mathbf{I}_{md},$$

$$\mathbf{b} = \mathbf{0}, \quad f(\mathbf{x}) = 0, \quad \text{and} \quad g(\mathbf{y}) = \sum_i f_i(\mathbf{x}_i)$$

in (2.13), where d is the dimension of $\mathbf{x}_i$, $\mathbf{1}_m$ is the vector of m ones, and $\otimes$ is the Kronecker product. □

Similarly, from Theorem 3.4 we have the following linear convergence result.

Theorem 6.2 *Suppose that each $f_i(\mathbf{x}_i)$ is μ-strongly convex and L-smooth, $i \in [m]$. Let $\beta = \sqrt{\mu L}$. Then for Algorithms 6.1–6.2, we have*

$$\begin{aligned}&\sum_{i=1}^m \left(\frac{1}{2\beta}\|\boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^*\|^2 + \frac{\beta}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2\right)\\ &\quad \le \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{-1} \sum_{i=1}^m \left(\frac{1}{2\beta}\|\boldsymbol{\lambda}_i^{k} - \boldsymbol{\lambda}_i^*\|^2 + \frac{\beta}{2}\|\mathbf{x}_i^{k} - \mathbf{x}_i^*\|^2\right).\end{aligned}$$

6.1.2 Linearized ADMM

When each f_i is L-smooth, we can also linearize f_i in step (6.4b) to simplify the computation, if the proximal mapping of f_i is not easily computable. The iterations of resulting linearized ADMM are as follows:

$$\begin{aligned}\mathbf{z}^{k+1} &= \operatorname*{argmin}_{\mathbf{z}} \sum_{i=1}^m \left(\left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i^k - \mathbf{z}\right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i^k - \mathbf{z}\right\|^2\right)\\ &= \frac{1}{m}\sum_{i=1}^m \left(\mathbf{x}_i^k + \frac{1}{\beta}\boldsymbol{\lambda}_i^k\right), \end{aligned} \tag{6.5a}$$

$$\begin{aligned}\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i - \mathbf{z}^{k+1}\right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i - \mathbf{z}^{k+1}\right\|^2 + D_{\psi_i}(\mathbf{x}_i, \mathbf{x}_i^k)\right)\\ &= \operatorname*{argmin}_{\mathbf{x}_i} \left(\left\langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i - \mathbf{x}_i^k\right\rangle + \frac{L}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2\right.\\ &\qquad \left. + \left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i - \mathbf{z}^{k+1}\right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i - \mathbf{z}^{k+1}\right\|^2\right)\end{aligned}$$

$$= \frac{1}{L+\beta}\left(L\mathbf{x}_i^k + \beta\mathbf{z}^{k+1} - \nabla f_i(\mathbf{x}_i^k) - \boldsymbol{\lambda}_i^k\right), \quad i \in [m], \tag{6.5b}$$

$$\boldsymbol{\lambda}_i^{k+1} = \boldsymbol{\lambda}_i^k + \beta\left(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1}\right), \quad i \in [m], \tag{6.5c}$$

by choosing

$$\psi_i(\mathbf{x}_i) = \frac{L}{2}\|\mathbf{x}_i\|^2 - f_i(\mathbf{x}_i).$$

We summarize the method in Algorithms 6.3 and 6.4.

Algorithm 6.3 Centralized linearized ADMM of the master

for $k = 0, 1, 2, \cdots$ **do**
  Wait until receiving $\mathbf{x}_i^k$ and $\boldsymbol{\lambda}_i^k$ from all the workers $i \in [m]$.
  Update $\mathbf{z}^{k+1}$ by (6.5a).
  Send $\mathbf{z}^{k+1}$ to all the workers.
end for

Algorithm 6.4 Centralized linearized ADMM of the ith worker

Initialize: $\mathbf{x}_i^0, \boldsymbol{\lambda}_i^0, i \in [m]$.
for $k = 0, 1, 2, \cdots$ **do**
  Send $(\mathbf{x}_i^k, \boldsymbol{\lambda}_i^k)$ to the master.
  Wait until receiving $\mathbf{z}^{k+1}$ from the master.
  Update $\mathbf{x}_i^{k+1}$ and $\boldsymbol{\lambda}_i^{k+1}$ by (6.5b) and (6.5c), respectively.
end for

Similar to Theorem 6.1, from Theorem 3.6 we can also have the $O(1/K)$ convergence rate. We omit the details and mainly discuss the linear convergence rate under stronger conditions. From Theorem 3.8 and using $L_\psi \le L - \mu$, where $\psi(\mathbf{x}) = \sum_{i=1}^m \psi_i(\mathbf{x}_i)$, we have the following linear convergence result.

Theorem 6.3 *Suppose that each $f_i(\mathbf{x}_i)$ is μ-strongly convex and L-smooth, $i \in [m]$. Let $\beta = \sqrt{\mu(2L-\mu)}$. Then for Algorithm 6.3–6.4, we have*

$$\begin{aligned}
&\sum_{i=1}^m \left(\frac{1}{2\beta}\|\boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^*\|^2 + \frac{\beta}{2}\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2 + D_{\psi_i}(\mathbf{x}_i^*, \mathbf{x}_i^{k+1})\right) \\
&\le \left[1 + \frac{1}{3}\min\left(\sqrt{\frac{\mu}{2L-\mu}}, \frac{\mu}{L-\mu}\right)\right]^{-1} \\
&\quad \times \sum_{i=1}^m \left(\frac{1}{2\beta}\|\boldsymbol{\lambda}_i^k - \boldsymbol{\lambda}_i^*\|^2 + \frac{\beta}{2}\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2 + D_{\psi_i}(\mathbf{x}_i^*, \mathbf{x}_i^k)\right).
\end{aligned}$$

6.1.3 Accelerated Linearized ADMM

Motivated by the results in Sect. 3.3.2, we can also use the accelerated linearized ADMM to solve Problem (6.2) to further improve the convergence rate of the linearized ADMM. From Algorithm 3.6 given in Sect. 3.3.2, we have the following iterations:

$$\mathbf{w}_i^k = \theta\mathbf{x}_i^k + (1-\theta)\tilde{\mathbf{x}}_i^k, \tag{6.6a}$$

$$\begin{aligned}\mathbf{z}^{k+1} &= \operatorname*{argmin}_{\mathbf{z}} \sum_{i=1}^m \left(\left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i^k - \mathbf{z}\right\rangle + \frac{\beta\theta}{2}\|\mathbf{x}_i^k - \mathbf{z}\|^2\right)\\ &= \frac{1}{m}\sum_{i=1}^m \left(\mathbf{x}_i^k + \frac{1}{\beta\theta}\boldsymbol{\lambda}_i^k\right),\end{aligned} \tag{6.6b}$$

$$\mathbf{x}_i^{k+1} = \frac{1}{\frac{\theta}{\alpha}+\mu}\left\{\mu\mathbf{w}_i^k + \frac{\theta}{\alpha}\mathbf{x}_i^k - \left[\nabla f_i(\mathbf{x}_i^k) + \boldsymbol{\lambda}_i^k + \beta\theta\left(\mathbf{x}_i^k - \mathbf{z}^{k+1}\right)\right]\right\}, \tag{6.6c}$$

$$\tilde{\mathbf{z}}^{k+1} = \theta\mathbf{z}^{k+1} + (1-\theta)\tilde{\mathbf{z}}^k, \tag{6.6d}$$

$$\tilde{\mathbf{x}}_i^{k+1} = \theta\mathbf{x}_i^{k+1} + (1-\theta)\tilde{\mathbf{x}}_i^k, \tag{6.6e}$$

$$\boldsymbol{\lambda}_i^{k+1} = \boldsymbol{\lambda}_i^k + \beta\theta\left(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1}\right). \tag{6.6f}$$

We summarize the method in Algorithms 6.5 and 6.6.

Algorithm 6.5 Accelerated centralized linearized ADMM of the master

---

Initialize: $\tilde{\mathbf{z}}^0$.
for $k = 0, 1, 2, \cdots$ **do**
  Wait until receiving $\mathbf{x}_i^k$ and $\boldsymbol{\lambda}_i^k$ from all the workers, $i \in [m]$.
  Update $\mathbf{z}^{k+1}$ and $\tilde{\mathbf{z}}^{k+1}$ by (6.6b) and (6.6d), respectively.
  Send $\mathbf{z}^{k+1}$ to all the workers.
end for

Algorithm 6.6 Accelerated centralized linearized ADMM of the ith worker

---

Initialize: $\mathbf{x}_i^0$, $\boldsymbol{\lambda}_i^0$, $i \in [m]$, and $\tilde{\mathbf{x}}_i^0$.
for $k = 0, 1, 2, \cdots$ **do**
  Send $(\mathbf{x}_i^k, \boldsymbol{\lambda}_i^k)$ to the master.
  Wait until receiving $\mathbf{z}^{k+1}$ from the master.
  Update $\mathbf{x}_i^{k+1}$, $\tilde{\mathbf{x}}_i^{k+1}$, $\boldsymbol{\lambda}_i^{k+1}$, and $\mathbf{w}_i^{k+1}$ by (6.6c), (6.6e), (6.6f), and (6.6a), respectively.
end for

Table 6.1 Complexity comparisons between centralized ADMM, centralized linearized ADMM (LADMM), and its accelerated version

Centralized ADMM	Centralized LADMM	Accelerated centralized LADMM
$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

Denote

$$\ell_k = (1-\theta)\sum_{i=1}^{m}\left(f_i(\tilde{\mathbf{x}}_i^k) - f_i(\mathbf{x}_i^*) + \left\langle \boldsymbol{\lambda}_i^*, \tilde{\mathbf{x}}_i^k - \tilde{\mathbf{z}}^k \right\rangle\right)$$
$$+\frac{\theta^2}{2\alpha}\sum_{i=1}^{m}\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2 + \frac{1}{2\beta}\sum_{i=1}^{m}\|\boldsymbol{\lambda}_i^k - \boldsymbol{\lambda}_i^*\|^2.$$

From Theorem 3.12 we have the following linear convergence result.

Theorem 6.4 *Suppose that each $f_i(\mathbf{x}_i)$ is μ-strongly convex and L-smooth, $i \in [m]$. Let*

$$\alpha = \frac{1}{4L}, \quad \beta = L, \quad \textit{and} \quad \theta = \sqrt{\frac{\mu}{L}}.$$

Then for the accelerated linearized ADMM (Algorithms 6.5–6.6), we have

$$\ell_{k+1} \le \left(1 - \sqrt{\frac{\mu}{L}}\right)\ell_k.$$

We list the convergence rate comparisons of different centralized ADMM methods in Table 6.1. Similar to the comparisons in Table 3.2, we see that the accelerated linearized ADMM is faster than the linearized ADMM with a better dependence on the condition number L/μ. The original ADMM has the same convergence rate as the accelerated linearized ADMM. However, the original ADMM may need to solve a subproblem iteratively at each iteration, while the accelerated linearized ADMM only performs a gradient descent type update.

6.2 Decentralized Optimization

In this section we consider the decentralized topology. In this case, we cannot use the constraints in (6.2) since there is no central node to compute $\mathbf{z}$. Denote $\mathcal{E}$ as the set of edges. Assume that all the nodes are ordered from 1 to m. For any two nodes i and j, if i and j are directly connected in the network and $i < j$, we say $(i, j) \in \mathcal{E}$. To simplify the presentation, we order the edges from 1 to $|\mathcal{E}|$. For each node i, we denote $\mathcal{N}_i$ as its neighborhood:

$$\mathcal{N}_i = \{j|(i, j) \in \mathcal{E} \text{ or } (j, i) \in \mathcal{E}\},$$

and $d_i = |\mathcal{N}_i|$ as its degree.

Introduce auxiliary variables $\mathbf{z}_{ij}$ if $(i, j) \in \mathcal{E}$. Then we can reformulate Problem (6.1) as follows (for example, see [1, 10, 12, 14]):

$$\begin{aligned} &\min_{\mathbf{x}_i, \mathbf{z}_{ij}} \sum_{i=1}^{m} f_i(\mathbf{x}_i), \\ &s.t. \;\; \mathbf{x}_i = \mathbf{z}_{ij}, \quad \mathbf{x}_j = \mathbf{z}_{ij}, \quad \forall (i, j) \in \mathcal{E}. \end{aligned} \tag{6.7}$$

That is to say, each variable $\mathbf{x}_i$ corresponds to one node, while each variable $\mathbf{z}_{ij}$ $(i < j)$ corresponds to one edge. The augmented Lagrangian function of Problem (6.7)

$$\begin{aligned} L(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda}) = \sum_{i=1}^{m} f_i(\mathbf{x}_i) + \sum_{(i,j)\in\mathcal{E}} \Big(&\langle \boldsymbol{\lambda}_{ij}, \mathbf{x}_i - \mathbf{z}_{ij} \rangle + \langle \boldsymbol{\gamma}_{ij}, \mathbf{x}_j - \mathbf{z}_{ij} \rangle \\ &+ \frac{\beta}{2} \|\mathbf{x}_i - \mathbf{z}_{ij}\|^2 + \frac{\beta}{2} \|\mathbf{x}_j - \mathbf{z}_{ij}\|^2 \Big). \end{aligned}$$

6.2.1 ADMM

We can use ADMM to solve Problem (6.7), which consists of the following iterations:

$$\begin{aligned} \mathbf{x}_i^{k+1} = \operatorname*{argmin}_{\mathbf{x}_i} \Bigg[f_i(\mathbf{x}_i) + &\sum_{j:(i,j)\in\mathcal{E}} \left(\left\langle \boldsymbol{\lambda}_{ij}^k, \mathbf{x}_i - \mathbf{z}_{ij}^k \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i - \mathbf{z}_{ij}^k \right\|^2 \right) \\ &+ \sum_{j:(j,i)\in\mathcal{E}} \left(\left\langle \boldsymbol{\gamma}_{ji}^k, \mathbf{x}_i - \mathbf{z}_{ji}^k \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i - \mathbf{z}_{ji}^k \right\|^2 \right) \Bigg], \end{aligned} \tag{6.8a}$$

$$\begin{aligned} \mathbf{z}_{ij}^{k+1} &= \operatorname*{argmin}_{\mathbf{z}_{ij}} \left(-\left\langle \boldsymbol{\lambda}_{ij}^k + \boldsymbol{\gamma}_{ij}^k, \mathbf{z}_{ij} \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i^{k+1} - \mathbf{z}_{ij} \right\|^2 + \frac{\beta}{2} \left\| \mathbf{x}_j^{k+1} - \mathbf{z}_{ij} \right\|^2 \right) \\ &= \frac{1}{2\beta} \left(\boldsymbol{\lambda}_{ij}^k + \boldsymbol{\gamma}_{ij}^k \right) + \frac{1}{2} \left(\mathbf{x}_i^{k+1} + \mathbf{x}_j^{k+1} \right), \end{aligned} \tag{6.8b}$$

$$\boldsymbol{\lambda}_{ij}^{k+1} = \boldsymbol{\lambda}_{ij}^k + \beta \left(\mathbf{x}_i^{k+1} - \mathbf{z}_{ij}^{k+1} \right), \tag{6.8c}$$

$$\boldsymbol{\gamma}_{ij}^{k+1} = \boldsymbol{\gamma}_{ij}^k + \beta \left(\mathbf{x}_j^{k+1} - \mathbf{z}_{ij}^{k+1} \right). \tag{6.8d}$$

Next, we introduce the result in [10] to simplify the above method by eliminating variables $\mathbf{z}_{ij}$, $\boldsymbol{\lambda}_{ij}$, and $\boldsymbol{\gamma}_{ij}$.

Summing (6.8c) and (6.8d) and using (6.8b), we have

$$\boldsymbol{\lambda}_{ij}^{k+1} + \boldsymbol{\gamma}_{ij}^{k+1} = \mathbf{0}, \quad \forall k \geq 0.$$

Initialize $\boldsymbol{\lambda}_{ij}^{0} = \boldsymbol{\gamma}_{ij}^{0} = \mathbf{0}$, we have

$$\boldsymbol{\lambda}_{ij}^{k} + \boldsymbol{\gamma}_{ij}^{k} = \mathbf{0}, \quad \forall k \geq 0.$$

Plugging it into (6.8b), we have

$$\mathbf{z}_{ij}^{k+1} = \frac{1}{2}\left(\mathbf{x}_i^{k+1} + \mathbf{x}_j^{k+1}\right), \quad \forall k \geq 0. \tag{6.9}$$

We may initialize

$$\mathbf{z}_{ij}^{0} = \frac{1}{2}\left(\mathbf{x}_i^{0} + \mathbf{x}_j^{0}\right).$$

From (6.9) and (6.8c), we have

$$\boldsymbol{\lambda}_{ij}^{k+1} = \boldsymbol{\lambda}_{ij}^{k} + \frac{\beta}{2}\left(\mathbf{x}_i^{k+1} - \mathbf{x}_j^{k+1}\right). \tag{6.10}$$

So we have

$$\boldsymbol{\lambda}_{ij}^{k+1} = \beta \sum_{t=1}^{k+1} \frac{1}{2}\left(\mathbf{x}_i^{t} - \mathbf{x}_j^{t}\right).$$

Similarly, we can have

$$\boldsymbol{\gamma}_{ij}^{k+1} = \beta \sum_{t=1}^{k+1} \frac{1}{2}\left(\mathbf{x}_j^{t} - \mathbf{x}_i^{t}\right).$$

Note that we only define $\boldsymbol{\lambda}_{ij}$, $\boldsymbol{\gamma}_{ij}$, and $\mathbf{z}_{ij}$ for $i < j$. Now we define

$$\boldsymbol{\lambda}_{ij} \equiv \boldsymbol{\gamma}_{ji} \quad \text{and} \quad \mathbf{z}_{ij} \equiv \mathbf{z}_{ji} \quad \text{for } i > j.$$

Then

$$\boldsymbol{\lambda}_{ij}^{k+1} = \beta \sum_{t=1}^{k+1} \frac{1}{2}\left(\mathbf{x}_i^{t} - \mathbf{x}_j^{t}\right) \quad \text{and} \quad \mathbf{z}_{ij}^{k+1} = \frac{1}{2}\left(\mathbf{x}_i^{k+1} + \mathbf{x}_j^{k+1}\right)$$

for both $i < j$ and $i > j$. So is (6.10). Thus (6.8a) can be simplified to

$$\begin{aligned}
\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i}\Bigg[f_i(\mathbf{x}_i) + \sum_{j:(i,j)\in\mathcal{E}}\left(\left\langle\boldsymbol{\lambda}_{ij}^k - \beta\mathbf{z}_{ij}^k, \mathbf{x}_i\right\rangle + \frac{\beta}{2}\|\mathbf{x}_i\|^2\right) \\
&\qquad + \sum_{j:(j,i)\in\mathcal{E}}\left(\left\langle\boldsymbol{\gamma}_{ji}^k - \beta\mathbf{z}_{ji}^k, \mathbf{x}_i\right\rangle + \frac{\beta}{2}\|\mathbf{x}_i\|^2\right)\Bigg] \\
&= \operatorname*{argmin}_{\mathbf{x}_i}\left[f_i(\mathbf{x}_i) + \sum_{j\in\mathcal{N}_i}\left(\left\langle\boldsymbol{\lambda}_{ij}^k - \beta\mathbf{z}_{ij}^k, \mathbf{x}_i\right\rangle + \frac{\beta}{2}\|\mathbf{x}_i\|^2\right)\right] \\
&= \operatorname*{argmin}_{\mathbf{x}_i}\left[f_i(\mathbf{x}_i) + \sum_{j\in\mathcal{N}_i}\left(\left\langle\boldsymbol{\lambda}_{ij}^k - \beta\mathbf{z}_{ij}^k + \beta\mathbf{x}_i^k, \mathbf{x}_i\right\rangle + \frac{\beta}{2}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2\right)\right] \\
&= \operatorname*{argmin}_{\mathbf{x}_i}\left[f_i(\mathbf{x}_i) + \sum_{j\in\mathcal{N}_i}\left(\left\langle\boldsymbol{\lambda}_{ij}^k + \frac{\beta}{2}\left(\mathbf{x}_i^k - \mathbf{x}_j^k\right), \mathbf{x}_i\right\rangle + \frac{\beta}{2}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2\right)\right].
\end{aligned} \tag{6.11}$$

Denote $\mathbf{L} \in \mathbb{R}^{m\times m}$ as the Laplacian matrix (Definition A.2) and $\mathbf{D}$ as the diagonal degree matrix with $\mathbf{D}_{ii} = d_i$. It is well known that $\mathbf{L}$ is symmetric and satisfies $\mathbf{0} \preceq \mathbf{L} \preceq 2\mathbf{D}$.[1]

Define

$$\mathbf{X} = \begin{pmatrix}\mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T\end{pmatrix} \in \mathbb{R}^{m\times d}, \quad f(\mathbf{X}) = \sum_{i=1}^m f_i(\mathbf{x}_i),$$

$$\boldsymbol{\upsilon}_i = \sum_{j\in\mathcal{N}_i}\boldsymbol{\lambda}_{ij}, \quad \text{and} \quad \boldsymbol{\Upsilon} = \begin{pmatrix}\boldsymbol{\upsilon}_1^T \\ \vdots \\ \boldsymbol{\upsilon}_m^T\end{pmatrix} \in \mathbb{R}^{m\times d}.$$

Then we have

$$\mathbf{L}_i^T\mathbf{X} = d_i\mathbf{x}_i^T - \sum_{j\in\mathcal{N}_i}\mathbf{x}_j^T,$$

where $\mathbf{L}_i$ is the i-th column of $\mathbf{L}$.

[1] $0 \le \boldsymbol{\alpha}^T\mathbf{L}\boldsymbol{\alpha} = \frac{1}{2}\sum_{(i,j)\in\mathcal{E}}(\alpha_i - \alpha_j)^2 \le \sum_{(i,j)\in\mathcal{E}}\left(\alpha_i^2 + \alpha_j^2\right) = 2\boldsymbol{\alpha}^T\mathbf{D}\boldsymbol{\alpha}$.

With the Laplacian matrix $\mathbf{L}$ and $\boldsymbol{v}_i$ introduced, (6.11) can be written as

$$\begin{aligned}\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \left[f_i(\mathbf{x}_i) + \left\langle \boldsymbol{v}_i^k, \mathbf{x}_i \right\rangle + \frac{\beta}{2} \left\langle \sum_{j \in \mathcal{N}_i} \mathbf{L}_{ij} \mathbf{x}_j^k, \mathbf{x}_i \right\rangle + \frac{\beta d_i}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^k \right\|^2 \right] \\ &= \operatorname{Prox}_{(\beta d_i)^{-1} f_i} \left(\mathbf{x}_i^k - \frac{1}{\beta d_i} \left(\boldsymbol{v}_i^k + \frac{\beta}{2} \sum_{j \in \mathcal{N}_i} \mathbf{L}_{ij} \mathbf{x}_j^k \right) \right), \quad i \in [m]. \end{aligned} \tag{6.12}$$

Summing (6.10) over $j \in \mathcal{N}_i$, we have that (6.10) gives

$$\boldsymbol{v}_i^{k+1} = \boldsymbol{v}_i^k + \frac{\beta}{2} \sum_{j \in \mathcal{N}_i} \mathbf{L}_{ij} \mathbf{x}_j^{k+1}, \quad i \in [m]. \tag{6.13}$$

(6.12)–(6.13) can be written in a compact form:

$$\mathbf{X}^{k+1} = \operatorname*{argmin}_{\mathbf{X}} \left(f(\mathbf{X}) + \left\langle \boldsymbol{\Upsilon}^k + \frac{\beta}{2} \mathbf{L} \mathbf{X}^k, \mathbf{X} \right\rangle + \frac{\beta}{2} \left\| \sqrt{\mathbf{D}} (\mathbf{X} - \mathbf{X}^k) \right\|^2 \right), \tag{6.14}$$

$$\boldsymbol{\Upsilon}^{k+1} = \boldsymbol{\Upsilon}^k + \frac{\beta}{2} \mathbf{L} \mathbf{X}^{k+1}. \tag{6.15}$$

Denoting $\mathbf{W} = \sqrt{\mathbf{L}/2}$, (6.15) can be rewritten as

$$\boldsymbol{\Upsilon}^{k+1} = \boldsymbol{\Upsilon}^k + \beta \mathbf{W}^2 \mathbf{X}^{k+1}.$$

Letting $\boldsymbol{\Upsilon}^0 \in \operatorname{Span}(\mathbf{W}^2)$, we know that

$$\boldsymbol{\Upsilon}^k \in \operatorname{Span}(\mathbf{W}^2), \quad \forall k \geq 0,$$

and there exists $\boldsymbol{\Omega}^k$ such that $\boldsymbol{\Upsilon}^k = \mathbf{W}\boldsymbol{\Omega}^k$.[2] Then (6.14) and (6.15) can be rewritten as[3]

$$\begin{aligned}\mathbf{X}^{k+1} &= \operatorname*{argmin}_{\mathbf{X}} \left(f(\mathbf{X}) + \left\langle \boldsymbol{\Omega}^k, \mathbf{W}\mathbf{X} \right\rangle + \beta \left\langle \mathbf{W}^2 \mathbf{X}^k, \mathbf{X} \right\rangle + \frac{\beta}{2} \left\| \sqrt{\mathbf{D}} (\mathbf{X} - \mathbf{X}^k) \right\|^2 \right) \\ &= \operatorname*{argmin}_{\mathbf{X}} \left(f(\mathbf{X}) + \left\langle \boldsymbol{\Omega}^k, \mathbf{W}\mathbf{X} \right\rangle + \frac{\beta}{2} \|\mathbf{W}\mathbf{X}\|^2 + D_\psi(\mathbf{X}, \mathbf{X}^k) \right), \end{aligned} \tag{6.16a}$$

$$\boldsymbol{\Omega}^{k+1} = \boldsymbol{\Omega}^k + \beta \mathbf{W}\mathbf{X}^{k+1}, \tag{6.16b}$$

---

[2] Denote $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ to be the eigen-decomposition of $\mathbf{W}$ with $\mathbf{U} \in \mathbb{R}^{m \times (m-1)}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{(m-1)\times(m-1)}$, then $\boldsymbol{\Lambda}$ is invertible. Since there exists $\mathbf{R}^k$ such that $\boldsymbol{\Upsilon}^k = \mathbf{U}\mathbf{R}^k$, we can choose $\boldsymbol{\Omega}^k = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{R}^k$ such that $\boldsymbol{\Upsilon}^k = \mathbf{W}\boldsymbol{\Omega}^k$.

[3] From (6.15), we have $\mathbf{R}^{k+1} = \mathbf{R}^k + \beta \boldsymbol{\Lambda}^2 \mathbf{U}^T \mathbf{X}^{k+1}$. Multiplying both sides by $\mathbf{U}\boldsymbol{\Lambda}^{-1}$, we get (6.16b).

with

$$\psi(\mathbf{X}) = \frac{\beta}{2}\left\|\sqrt{\mathbf{D}}\mathbf{X}\right\|^2 - \frac{\beta}{2}\|\mathbf{W}\mathbf{X}\|^2.$$

Thus, algorithm (6.8a)–(6.8d) is equivalent to using the linearized augmented Lagrangian method to solve problem

$$\min_{\mathbf{X}} f(\mathbf{X}), \quad s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{0}.$$

Algorithm (6.16a)–(6.16b) is not implementable in the distributed manner due to $\mathbf{W} = \sqrt{\mathbf{L}/2}$, which is only used for analysis. In practice, we implement the original (6.12)–(6.13) instead. We present algorithm (6.12)–(6.13) in Algorithm 6.7.

Algorithm 6.7 Decentralized ADMM of the ith node

Initialize: $\mathbf{x}_i^0$ and $\boldsymbol{v}_i^0 = \mathbf{0}$, $i \in [m]$.
Send $\mathbf{x}_i^0$ to its neighbors.
Wait until receiving $\mathbf{x}_j^0$ from all its neighbors, $j \in \mathcal{N}_i$.
for $k = 0, 1, 2, \cdots$ **do**
  Update $\mathbf{x}_i^{k+1}$ by (6.12).
  Send $\mathbf{x}_i^{k+1}$ to its neighbors.
  Wait until receiving $\mathbf{x}_j^{k+1}$ from all its neighbors, $j \in \mathcal{N}_i$.
  Update $\boldsymbol{v}_i^{k+1}$ by (6.13).
end for

6.2.1.1 Convergence Analysis

We consider the linearized augmented Lagrangian method (6.16a)–(6.16b) with a general ψ. From Theorem 3.14 or 3.8, we have the following convergence result.

Theorem 6.5 *Assume that each f_i is μ-strongly convex and L-smooth, $i \in [m]$, and $\psi(\mathbf{y})$ is convex and L_ψ-smooth. Initialize $\boldsymbol{\Omega}^0 = \mathbf{0}$. Then for algorithm (6.16a)–(6.16b) we have*

$$\begin{aligned}
&\frac{1}{2\beta}\|\boldsymbol{\Omega}^{k+1} - \boldsymbol{\Omega}^*\|^2 + \frac{\beta}{2}\|\mathbf{W}\mathbf{X}^{k+1} - \mathbf{W}\mathbf{X}^*\|^2 + D_\psi(\mathbf{X}^*, \mathbf{X}^{k+1}) \\
&\leq \left(1 + \frac{1}{3}\min\left\{\frac{\beta\sigma_{\mathbf{L}}}{2(L + L_\psi)}, \frac{\mu}{\beta\|\mathbf{W}\|_2^2}, \frac{\mu}{L_\psi}\right\}\right)^{-1} \\
&\quad \times \left(\frac{1}{2\beta}\|\boldsymbol{\Omega}^k - \boldsymbol{\Omega}^*\|^2 + \frac{\beta}{2}\|\mathbf{W}\mathbf{X}^k - \mathbf{W}\mathbf{X}^*\|^2 + D_\psi(\mathbf{X}^*, \mathbf{X}^k)\right),
\end{aligned}$$

where $\sigma_{\mathbf{L}}$ is the smallest positive eigenvalue of $\mathbf{L}$.

Proof From the proof of Theorem 3.8, to prove this theorem we only need to check

$$\|\mathbf{W}(\boldsymbol{\Omega}^k - \boldsymbol{\Omega}^*)\| \geq \sqrt{\sigma_{\mathbf{L}}/2}\|\boldsymbol{\Omega}^k - \boldsymbol{\Omega}^*\|.$$

Note that $\mathbf{B} = \mathbf{W}$ and $\sigma^2 = \frac{\sigma_{\mathbf{L}}}{2}$ in Theorem 3.8.

Since the network has to be connected, the rank of the Laplacian matrix $\mathbf{L}$ is $m-1$ (Proposition A.2). Let $\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{L}$ be its economical SVD with $\mathbf{V} \in \mathbb{R}^{m\times(m-1)}$. For any $\boldsymbol{\Omega}$ belonging to the column space of $\mathbf{W}$, we have

$$\begin{aligned}
\|\mathbf{W}\boldsymbol{\Omega}\|^2 &= \sum_{i=1}^{d} \boldsymbol{\Omega}_i^T \mathbf{W}^2 \boldsymbol{\Omega}_i \\
&= \frac{1}{2}\sum_{i=1}^{d} \boldsymbol{\Omega}_i^T \mathbf{L} \boldsymbol{\Omega}_i \\
&= \frac{1}{2}\sum_{i=1}^{d} (\mathbf{V}^T\boldsymbol{\Omega}_i)^T \boldsymbol{\Sigma} (\mathbf{V}^T\boldsymbol{\Omega}_i) \\
&\geq \frac{\sigma_{\mathbf{L}}}{2}\sum_{i=1}^{d} \|\mathbf{V}^T\boldsymbol{\Omega}_i\|^2 = \frac{\sigma_{\mathbf{L}}}{2}\|\mathbf{V}^T\boldsymbol{\Omega}\|^2 \stackrel{a}{=} \frac{\sigma_{\mathbf{L}}}{2}\|\boldsymbol{\Omega}\|^2,
\end{aligned}$$

where we denote $\boldsymbol{\Omega}_i$ to be the ith column of $\boldsymbol{\Omega}$, and $\stackrel{a}{=}$ follows from the fact that $\boldsymbol{\Omega}$ belongs to the column space of $\mathbf{W}$, i.e., there exists $\boldsymbol{\alpha} \in \mathbb{R}^{(m-1)\times d}$ such that $\boldsymbol{\Omega} = \mathbf{V}\boldsymbol{\alpha}$.

From (6.16b) and the KKT condition, we know that both $\boldsymbol{\Omega}^k$ and $\boldsymbol{\Omega}^*$ belong to the column space of $\mathbf{W}$. So we have

$$\|\mathbf{W}(\boldsymbol{\Omega}^k - \boldsymbol{\Omega}^*)\| \geq \sqrt{\sigma_{\mathbf{L}}/2}\|\boldsymbol{\Omega}^k - \boldsymbol{\Omega}^*\|.$$

From Theorem 3.8, we get the conclusion. □

Now, we discuss algorithm (6.16a)–(6.16b) with the special

$$\psi(\mathbf{X}) = \frac{\beta}{2}\left\|\sqrt{\mathbf{D}}\mathbf{X}\right\|^2 - \frac{\beta}{2}\|\mathbf{W}\mathbf{X}\|^2 \quad \text{and} \quad L_\psi = \beta d_{\max},$$

where $d_{\max} = \max\{d_i\}$. Then algorithm (6.16a)–(6.16b) reduces to Algorithm 6.7. From Remark 3.4 and

$$\|\mathbf{W}\|_2^2 = \frac{1}{2}\|\mathbf{L}\|_2 \leq \|\mathbf{D}\|_2 \leq d_{\max}$$

(that is, $\|\mathbf{B}\|_2^2 \leq d_{\max}$ and $\sigma^2 = \frac{\sigma_{\mathbf{L}}}{2}$ in Remark 3.4), we have the following theorem.

Theorem 6.6 *Assume that each f_i is μ-strongly convex and L-smooth, $i \in [m]$. Initialize $\mathbf{\Omega}^0 = \mathbf{0}$ and let $\beta = O\left(\sqrt{\frac{\mu L}{\sigma_{\mathbf{L}} d_{\max}}}\right)$. Then Algorithm 6.7 needs $O\left(\left(\sqrt{\frac{L d_{\max}}{\mu \sigma_{\mathbf{L}}}} + \frac{d_{\max}}{\sigma_{\mathbf{L}}}\right) \log \frac{1}{\epsilon}\right)$ iterations to find an ϵ-approximate solution $(\mathbf{X}, \mathbf{\Omega})$, i.e.,*

$$\frac{1}{2\beta}\|\mathbf{\Omega} - \mathbf{\Omega}^*\|^2 + \frac{\beta}{2}\|\mathbf{W}\mathbf{X} - \mathbf{W}\mathbf{X}^*\|^2 + D_\psi(\mathbf{X}^*, \mathbf{X}) \le \epsilon.$$

We see that the complexity depends on the condition number $\frac{L}{\mu}$ of the objective function and $\frac{d_{\max}}{\sigma_{\mathbf{L}}}$. The latter one can be regarded as the condition number of the Laplacian matrix $\mathbf{L}$.

6.2.2 Linearized ADMM

The subproblem in (6.8a) is a proximal mapping of f_i (c.f. (6.12)). When the proximal mapping of f_i is not easily computable, as in Sect. 3.2 we may linearize the objective f_i, which leads to the following step [10]:

$$\begin{aligned}
\mathbf{x}_i^{k+1} = \operatorname*{argmin}_{\mathbf{x}_i} \Bigg[& \left\langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i - \mathbf{x}_i^k \right\rangle + \frac{L}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2 \\
& + \sum_{j:(i,j)\in\mathcal{E}} \left(\left\langle \boldsymbol{\lambda}_{ij}^k, \mathbf{x}_i - \mathbf{z}_{ij}^k \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i - \mathbf{z}_{ij}^k\right\|^2 \right) \\
& + \sum_{j:(j,i)\in\mathcal{E}} \left(\left\langle \boldsymbol{\gamma}_{ji}^k, \mathbf{x}_i - \mathbf{z}_{ji}^k \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i - \mathbf{z}_{ji}^k\right\|^2 \right) \Bigg].
\end{aligned}$$

Steps (6.8b)–(6.8d) remain unchanged. Similar to (6.11), we have

$$\begin{aligned}
\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \Bigg[\left\langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i - \mathbf{x}_i^k \right\rangle + \frac{L}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2 \\
&\quad + \sum_{j\in\mathcal{N}_i} \left(\left\langle \boldsymbol{\lambda}_{ij}^k + \frac{\beta}{2}\left(\mathbf{x}_i^k - \mathbf{x}_j^k\right), \mathbf{x}_i \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2 \right) \Bigg] \\
&= \mathbf{x}_i^k - \frac{1}{L + \beta d_i} \left\{ \nabla f_i(\mathbf{x}_i^k) + \sum_{j\in\mathcal{N}_i} \left[\boldsymbol{\lambda}_{ij}^k + \frac{\beta}{2}\left(\mathbf{x}_i^k - \mathbf{x}_j^k\right) \right] \right\}.
\end{aligned}$$

Similar to the deductions in Sect. 6.2.1, the resultant linearized ADMM can be rewritten as

$$\begin{aligned}
\mathbf{X}^{k+1} &= \operatorname*{argmin}_{\mathbf{X}} \left(\left\langle \nabla f(\mathbf{X}^k), \mathbf{X} \right\rangle + \frac{L}{2}\|\mathbf{X} - \mathbf{X}^k\|^2 \right. \\
&\quad \left. + \left\langle \boldsymbol{\Omega}^k, \mathbf{W}\mathbf{X} \right\rangle + \beta \left\langle \mathbf{W}^2\mathbf{X}^k, \mathbf{X} \right\rangle + \frac{\beta}{2} \left\| \sqrt{\mathbf{D}}(\mathbf{X} - \mathbf{X}^k) \right\|^2 \right) \\
&= \mathbf{X}^k - (L\mathbf{I} + \beta\mathbf{D})^{-1} \left(\beta\mathbf{W}^2\mathbf{X}^k + \nabla f(\mathbf{X}^k) + \mathbf{W}\boldsymbol{\Omega}^k \right), \qquad (6.17a) \\
\boldsymbol{\Omega}^{k+1} &= \boldsymbol{\Omega}^k + \beta\mathbf{W}\mathbf{X}^{k+1}, \qquad (6.17b)
\end{aligned}$$

which is also a special case of algorithm (6.16a)–(6.16b) with

$$\psi(\mathbf{X}) = \frac{L}{2}\|\mathbf{X}\|^2 - f(\mathbf{X}) + \frac{\beta}{2}\left\|\sqrt{\mathbf{D}}\mathbf{X}\right\|^2 - \frac{\beta}{2}\|\mathbf{W}\mathbf{X}\|^2 \quad \text{and} \quad L_\psi = L + \beta d_{\max}.$$

We present the method in Algorithm 6.8, which is a distributed version of (6.17a)–(6.17b).

Algorithm 6.8 Decentralized linearized ADMM of the ith node

Initialize: $\mathbf{x}_i^0$ and $\boldsymbol{v}_i^0 = \mathbf{0}$, $i \in [m]$.
Send $\mathbf{x}_i^0$ to its neighbors.
Wait until receiving $\mathbf{x}_j^0$ from all its neighbors, $j \in \mathcal{N}_i$.
for $k = 0, 1, 2, \cdots$ **do**

$$\begin{aligned}
\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \left(\left\langle \nabla f_i(\mathbf{x}_i^k), \mathbf{x}_i \right\rangle + \left\langle \boldsymbol{v}_i^k, \mathbf{x}_i \right\rangle + \frac{\beta}{2} \left\langle \sum_{j\in\mathcal{N}_i} \mathbf{L}_{ij}\mathbf{x}_j^k, \mathbf{x}_i \right\rangle + \frac{\beta d_i + L}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^k \right\|^2 \right) \\
&= \mathbf{x}_i^k - \frac{1}{\beta d_i + L} \left(\nabla f_i(\mathbf{x}_i^k) + \boldsymbol{v}_i^k + \frac{\beta}{2} \sum_{j\in\mathcal{N}_i} \mathbf{L}_{ij}\mathbf{x}_j^k \right).
\end{aligned}$$

Send $\mathbf{x}_i^{k+1}$ to its neighbors.
Wait until receiving $\mathbf{x}_j^{k+1}$ from all its neighbors, $j \in \mathcal{N}_i$.
$\boldsymbol{v}_i^{k+1} = \boldsymbol{v}_i^k + \frac{\beta}{2}\sum_{j\in\mathcal{N}_i} \mathbf{L}_{ij}\mathbf{x}_j^{k+1}$.
end for

From Remark 3.4, we have the following theorem.

Theorem 6.7 *Assume that each f_i is μ-strongly convex and L-smooth, $i \in [m]$. Initialize $\boldsymbol{\Omega}^0 = \mathbf{0}$ and let $\beta = O\left(\sqrt{\frac{\mu L}{\sigma_{\mathbf{L}} d_{\max}}}\right)$. Then Algorithm 6.8 needs*

$O\left(\left(\frac{L}{\mu}+\frac{d_{\max}}{\sigma_{\mathbf{L}}}\right)\log\frac{1}{\epsilon}\right)$ *iterations to find* $(\mathbf{X},\boldsymbol{\Omega})$ *such that*

$$\frac{1}{2\beta}\|\boldsymbol{\Omega}-\boldsymbol{\Omega}^*\|^2+\frac{\beta}{2}\|\mathbf{W}\mathbf{X}-\mathbf{W}\mathbf{X}^*\|^2+D_\psi(\mathbf{X}^*,\mathbf{X})\le\epsilon.$$

6.2.3 Accelerated Linearized ADMM

In this section, we accelerate algorithm (6.17a)–(6.17b) using Algorithm 3.6. The resultant algorithm has the following iterations [8]:

$$\mathbf{Y}^k=\theta\mathbf{X}^k+(1-\theta)\widetilde{\mathbf{X}}^k, \tag{6.18a}$$

$$\mathbf{X}^{k+1}=\frac{1}{\frac{\theta}{\alpha}+\mu}\left[\mu\mathbf{Y}^k+\frac{\theta}{\alpha}\mathbf{X}^k-\left(\nabla f(\mathbf{Y}^k)+\mathbf{W}\boldsymbol{\Omega}^k+\beta\theta\mathbf{W}^2\mathbf{X}^k\right)\right], \tag{6.18b}$$

$$\widetilde{\mathbf{X}}^{k+1}=\theta\mathbf{X}^{k+1}+(1-\theta)\widetilde{\mathbf{X}}^k, \tag{6.18c}$$

$$\boldsymbol{\Omega}^{k+1}=\boldsymbol{\Omega}^k+\beta\theta\mathbf{W}\mathbf{X}^{k+1}, \tag{6.18d}$$

and it is presented in Algorithm 6.9 in the distributed manner.

Algorithm 6.9 Accelerated decentralized linearized ADMM of the ith node

Initialize: $\mathbf{x}_i^0=\widetilde{\mathbf{x}}_i^0$ and $\boldsymbol{v}_i^0=\mathbf{0}$, $i\in[m]$.
Send $\mathbf{x}_i^0$ to its neighbors.
Wait until receiving $\mathbf{x}_j^0$ from all its neighbors, $j\in\mathcal{N}_i$.
for $k=0,1,2,\cdots$ **do**
  $\mathbf{y}_i^k=\theta\mathbf{x}_i^k+(1-\theta)\widetilde{\mathbf{x}}_i^k$.
  $\mathbf{x}_i^{k+1}=\frac{1}{\frac{\theta}{\alpha}+\mu}\left[\mu\mathbf{y}_i^k+\frac{\theta}{\alpha}\mathbf{x}_i^k-\left(\nabla f_i(\mathbf{y}_i^k)+\boldsymbol{v}_i^k+\frac{\beta\theta}{2}\sum_{j\in\mathcal{N}_i}\mathbf{L}_{ij}\mathbf{x}_j^k\right)\right]$.
  $\widetilde{\mathbf{x}}_i^{k+1}=\theta\mathbf{x}_i^{k+1}+(1-\theta)\widetilde{\mathbf{x}}_i^k$.
  Send $\mathbf{x}_i^{k+1}$ to its neighbors.
  Wait until receiving $\mathbf{x}_j^{k+1}$ from all its neighbors, $j\in\mathcal{N}_i$.
  $\boldsymbol{v}_i^{k+1}=\boldsymbol{v}_i^k+\frac{\beta\theta}{2}\sum_{j\in\mathcal{N}_i}\mathbf{L}_{ij}\mathbf{x}_j^{k+1}$.
end for

Denote

$$\ell_k=(1-\theta)\left(f(\widetilde{\mathbf{X}}^k)-f(\mathbf{X}^*)+\left\langle\boldsymbol{\Omega}^*,\mathbf{W}\widetilde{\mathbf{X}}^k\right\rangle\right)$$
$$+\frac{\theta^2}{2\alpha}\|\mathbf{X}^k-\mathbf{X}^*\|^2+\frac{1}{2\beta}\|\boldsymbol{\Omega}^k-\boldsymbol{\Omega}^*\|^2.$$

Table 6.2 Complexity comparisons between decentralized ADMM, decentralized linearized ADMM (LADMM), and its accelerated version

Decentralized ADMM	Decentralized LADMM	Accelerated decentralized LADMM
$O\left(\left(\sqrt{\frac{Ld_{\max}}{\mu\sigma_{\mathbf{L}}}}+\frac{d_{\max}}{\sigma_{\mathbf{L}}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\left(\frac{L}{\mu}+\frac{d_{\max}}{\sigma_{\mathbf{L}}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{Ld_{\max}}{\mu\sigma_{\mathbf{L}}}}\log\frac{1}{\epsilon}\right)$

From Theorem 3.15 (note that $\|\mathbf{B}\|_2^2 \le d_{\max}$ and $\sigma^2 = \frac{\sigma_{\mathbf{L}}}{2}$ in Theorem 3.15), we have the following convergence result.

Theorem 6.8 *Suppose that each f_i is μ-strongly convex and L-smooth, $i \in [m]$. Assume that $\frac{2d_{\max}}{\sigma_{\mathbf{L}}} \le \frac{L}{\mu}$, where $\sigma_{\mathbf{L}}$ is the smallest non-zero singular value of* $\mathbf{L}$. *Let*

$$\alpha = \frac{1}{4L}, \quad \beta = \frac{L}{d_{\max}}, \quad \textit{and} \quad \theta = \sqrt{\frac{2\mu d_{\max}}{L\sigma_{\mathbf{L}}}}.$$

Then for algorithm (6.18a)–(6.18d) (Algorithm 6.9), we have

$$\ell_{k+1} \le O\left(1-\sqrt{\frac{\mu\sigma_{\mathbf{L}}}{2Ld_{\max}}}\right)\ell_k.$$

We list the convergence rates comparisons in Table 6.2.

6.3 Asynchronous Distributed ADMM

Algorithms 6.1–6.2 proceed in a synchronous manner. That is, the master needs to wait for all the workers to finish their updates before it can proceed. When the workers have different delays, the master has to wait for the slowest worker before the next iteration, i.e., the system proceeds at the pace of the slowest worker. In this section, we introduce the asynchronous ADMM proposed in [4, 5] to reduce the waiting time.

In the asynchronous ADMM, the master does not wait for all the workers, but proceeds as long as it receives information from a partial set of workers instead. We denote the partial set at iteration k as $\mathcal{A}^k$, and $\mathcal{A}_c^k$ as the complementary set of $\mathcal{A}^k$, which means the set of workers whose information does not arrive at iteration k. We use α to lower bound the size of $\mathcal{A}^k$. In the asynchronous ADMM, we often require that the master has to receive the updates from every worker at least once in every τ iterations. That is, we do not allow some workers to be absent for a long time. So we make the following bounded delay assumption.

Assumption 3 The maximum tolerable delay for all i and k is upper bounded.

Algorithm 6.10 Asynchronous ADMM of the master

Initialize: $\tilde{d}_1^1 = \cdots = \tilde{d}_m^1 = 0$.
for $k = 1, 2, \cdots$ **do**
Wait until receiving $\hat{\mathbf{x}}_i^k$ and $\hat{\boldsymbol{\lambda}}_i^k$ from workers $i \in \mathcal{A}^k$ such that $|\mathcal{A}^k| \geq \alpha$ and $\tilde{d}_j^k < \tau - 1$ for all $j \in \mathcal{A}_c^k$.

$$\mathbf{x}_i^{k+1} = \begin{cases} \hat{\mathbf{x}}_i^k, \ \forall i \in \mathcal{A}^k, \\ \mathbf{x}_i^k, \ \forall i \in \mathcal{A}_c^k. \end{cases}$$

$$\boldsymbol{\lambda}_i^{k+1} = \begin{cases} \hat{\boldsymbol{\lambda}}_i^k, \ \forall i \in \mathcal{A}^k, \\ \boldsymbol{\lambda}_i^k, \ \forall i \in \mathcal{A}_c^k. \end{cases}$$

$$\tilde{d}_i^{k+1} = \begin{cases} 0, \qquad \forall i \in \mathcal{A}^k, \\ \tilde{d}_i^k + 1, \ \forall i \in \mathcal{A}_c^k. \end{cases}$$

$$\begin{aligned} \mathbf{z}^{k+1} &= \operatorname*{argmin}_{\mathbf{z}} \left[\sum_{i=1}^m \left(\left\langle \boldsymbol{\lambda}_i^{k+1}, \mathbf{x}_i^{k+1} - \mathbf{z} \right\rangle + \frac{\beta}{2} \|\mathbf{x}_i^{k+1} - \mathbf{z}\|^2 \right) + \frac{\rho}{2} \|\mathbf{z} - \mathbf{z}^k\|^2 \right] \\ &= \frac{1}{\rho + m\beta} \left[\rho \mathbf{z}^k + \sum_{i=1}^m \left(\boldsymbol{\lambda}_i^{k+1} + \beta \mathbf{x}_i^{k+1} \right) \right]. \end{aligned}$$

Broadcast $\mathbf{z}^{k+1}$ to the workers in $\mathcal{A}^k$.
end for

Denote the upper bound as τ, then it must be that for every i,

$$i \in \mathcal{A}^k \cup \mathcal{A}^{k-1} \cdots \cup \mathcal{A}^{\max\{k-\tau+1,0\}}.$$

We describe the asynchronous ADMM in Algorithms 6.10–6.11. It has several differences from the synchronous ADMM:

1. The master only updates $(\mathbf{x}_i^{k+1}, \boldsymbol{\lambda}_i^{k+1})$ with $i \in \mathcal{A}^k$.
2. $\mathbf{z}$ is updated by solving a subproblem with an additional proximal term.
3. We introduce $\tilde{d}_i$, the amount of delay, for each worker such that the bounded delay assumption holds. The master must wait if there exists one worker with $\tilde{d}_i = \tau - 1$.
4. The master only broadcasts the up-to-date $\mathbf{z}$ to the arrived workers in $\mathcal{A}^k$.

6.3.1 Convergence

To simplify the analysis, we rewrite the method from the master's point of view:

$$\mathbf{x}_i^{k+1} = \begin{cases} \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \boldsymbol{\lambda}_i^{\bar{k}_i+1}, \mathbf{x}_i \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i - \mathbf{z}^{\bar{k}_i+1} \right\|^2 \right), & \forall i \in \mathcal{A}^k, \\ \mathbf{x}_i^k, & \forall i \in \mathcal{A}_c^k. \end{cases} \tag{6.19a}$$

Algorithm 6.11 Asynchronous ADMM of the ith worker

Initialize: $\hat{\mathbf{x}}_i^0$ and $\hat{\boldsymbol{\lambda}}_i^0$, $i \in [m]$.
for $k_i = 1, 2, \cdots$ **do**
Wait until receiving $\mathbf{z}$ from the master.

$$\begin{aligned}\hat{\mathbf{x}}_i^{k_i+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \hat{\boldsymbol{\lambda}}_i^{k_i}, \mathbf{x}_i - \mathbf{z} \right\rangle + \frac{\beta}{2}\|\mathbf{x}_i - \mathbf{z}\|^2 \right) \\ &= \operatorname{Prox}_{\beta^{-1} f_i}\left(\mathbf{z} - \frac{1}{\beta}\hat{\boldsymbol{\lambda}}_i^{k_i} \right).\end{aligned}$$

$$\hat{\boldsymbol{\lambda}}_i^{k_i+1} = \hat{\boldsymbol{\lambda}}_i^{k_i} + \beta\left(\hat{\mathbf{x}}_i^{k_i+1} - \mathbf{z} \right).$$

Send $(\hat{\mathbf{x}}_i^{k_i+1}, \hat{\boldsymbol{\lambda}}_i^{k_i+1})$ to the master.
end for

$$\boldsymbol{\lambda}_i^{k+1} = \begin{cases} \boldsymbol{\lambda}_i^{\overline{k}_i+1} + \beta\left(\mathbf{x}_i^{k+1} - \mathbf{z}^{\overline{k}_i+1} \right), & \forall i \in \mathcal{A}^k, \\ \boldsymbol{\lambda}_i^k, & \forall i \in \mathcal{A}_c^k. \end{cases} \tag{6.19b}$$

$$\mathbf{z}^{k+1} = \operatorname*{argmin}_{\mathbf{z}} \left[\sum_{i=1}^m \left(\left\langle \boldsymbol{\lambda}_i^{k+1}, \mathbf{x}_i^{k+1} - \mathbf{z} \right\rangle + \frac{\beta}{2}\left\| \mathbf{x}_i^{k+1} - \mathbf{z} \right\|^2 \right) + \frac{\rho}{2}\left\| \mathbf{z} - \mathbf{z}^k \right\|^2 \right], \tag{6.19c}$$

where we denote $\overline{k}_i$ as the last iteration before iteration k for which worker $i \in \mathcal{A}^k$ arrives, i.e., $i \in \mathcal{A}^{\overline{k}_i}$. Thus, for all workers $i \in \mathcal{A}^k$, we have

$$\begin{aligned} &\mathbf{x}_i^{\overline{k}_i+1} = \mathbf{x}_i^{\overline{k}_i+2} = \cdots = \mathbf{x}_i^k, \\ &\boldsymbol{\lambda}_i^{\overline{k}_i+1} = \boldsymbol{\lambda}_i^{\overline{k}_i+2} = \cdots = \boldsymbol{\lambda}_i^k, \text{ and} \\ &\max\{k - \tau, 0\} \le \overline{k}_i < k. \end{aligned} \tag{6.20}$$

For each $i \in \mathcal{A}_c^k$, we denote $\widetilde{k}_i$ as the last iteration before iteration k for which worker i arrives, i.e., $i \in \mathcal{A}^{\widetilde{k}_i}$. Under the bounded delay assumption, we have

$$\max\{k - \tau + 1, 0\} \le \widetilde{k}_i < k.$$

Thus, for all workers $i \in \mathcal{A}_c^k$, we have

$$\begin{aligned} &\mathbf{x}_i^{\widetilde{k}_i+1} = \mathbf{x}_i^{\widetilde{k}_i+2} = \cdots = \mathbf{x}_i^k = \mathbf{x}_i^{k+1} \text{ and} \\ &\boldsymbol{\lambda}_i^{\widetilde{k}_i+1} = \boldsymbol{\lambda}_i^{\widetilde{k}_i+2} = \cdots = \boldsymbol{\lambda}_i^k = \boldsymbol{\lambda}_i^{k+1}. \end{aligned}$$

We also denote $\hat{k}_i$ as the last iteration before $\widetilde{k}_i$ for which $i \in \mathcal{A}^{\widetilde{k}_i}$ arrives, i.e., $i \in \mathcal{A}^{\hat{k}_i}$. We also have

$$\max\{\widetilde{k}_i - \tau, 0\} \le \hat{k}_i < \widetilde{k}_i.$$

Thus, for all workers $i \in \mathcal{A}_c^k$, we have

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^{\widetilde{k}_i+1} = \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \boldsymbol{\lambda}_i^{\hat{k}_i+1}, \mathbf{x}_i \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i - \mathbf{z}^{\hat{k}_i+1} \right\|^2 \right), \tag{6.21}$$

$$\boldsymbol{\lambda}_i^{k+1} = \boldsymbol{\lambda}_i^{\widetilde{k}_i+1} = \boldsymbol{\lambda}_i^{\hat{k}_i+1} + \beta \left(\mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{z}^{\hat{k}_i+1} \right), \tag{6.22}$$

$$\begin{aligned} &\mathbf{x}_i^{\hat{k}_i+1} = \mathbf{x}_i^{\hat{k}_i+2} = \cdots = \mathbf{x}_i^{\widetilde{k}_i}, \quad \text{and} \\ &\boldsymbol{\lambda}_i^{\hat{k}_i+1} = \boldsymbol{\lambda}_i^{\hat{k}_i+2} = \cdots = \boldsymbol{\lambda}_i^{\widetilde{k}_i}. \end{aligned} \tag{6.23}$$

Denote $(\mathbf{x}_1^*, \cdots, \mathbf{x}_m^*, \mathbf{z}^*, \boldsymbol{\lambda}_1^*, \cdots, \boldsymbol{\lambda}_m^*)$ to be a KKT point. We have

$$\sum_{i=1}^m \boldsymbol{\lambda}_i^* = \mathbf{0}, \quad \mathbf{z}^* = \mathbf{x}_i^*, \quad \text{and} \quad \nabla f_i(\mathbf{x}_i^*) + \boldsymbol{\lambda}_i^* = \mathbf{0}, \quad i \in [m].$$

Also denote $f^* = \sum_{i=1}^m f_i(\mathbf{z}^*)$.

Theorem 6.9 *Assume that each f_i is convex and L-smooth, $i \in [m]$, and Assumption 3 holds true. Let*

$$\beta > \frac{1 + L^2 + \sqrt{(1+L^2)^2 + 8L^2}}{2} \quad \textit{and} \quad \rho > \frac{1}{2}\left[m(1+\beta^2)(\tau-1)^2 - m\beta \right].$$

Suppose that $(\mathbf{x}_1^k, \cdots, \mathbf{x}_m^k, \mathbf{z}^k, \boldsymbol{\lambda}_1^k, \cdots, \boldsymbol{\lambda}_m^k)$ generated by (6.19a)–(6.19c) are bounded, then $(\mathbf{x}_1^k, \cdots, \mathbf{x}_m^k, \mathbf{z}^k, \boldsymbol{\lambda}_1^k, \cdots, \boldsymbol{\lambda}_m^k)$ converge to the set of KKT points of Problem (6.2) in the sense of

$$\sum_{i=1}^m \boldsymbol{\lambda}_i^k \to \mathbf{0}, \quad \mathbf{x}_i^{k+1} - \mathbf{z}^{k+1} \to \mathbf{0}, \quad \textit{and} \quad \nabla f_i(\mathbf{x}_i^{k+1}) + \boldsymbol{\lambda}_i^{k+1} = \mathbf{0}, \quad i \in [m].$$

Proof Recall the augmented Lagrangian function in (6.3). Notice that

$$\begin{aligned} &L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L(\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k) \\ &= \left(L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L(\mathbf{x}^{k+1}, \mathbf{z}^k, \boldsymbol{\lambda}^{k+1}) \right) \end{aligned}$$

$$+\left(L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})-L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^k)\right)$$
$$+\left(L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^k)-L(\mathbf{x}^k,\mathbf{z}^k,\boldsymbol{\lambda}^k)\right).$$

We bound the three terms one by one.

For the first term, from the $(m\beta+\rho)$-strong convexity of $L(\mathbf{x},\mathbf{z},\boldsymbol{\lambda})+\frac{\rho}{2}\left\|\mathbf{z}-\mathbf{z}^k\right\|^2$ with respect to $\mathbf{z}$, (6.19c), and (A.7), we have

$$\begin{aligned}&L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})-\left(L(\mathbf{x}^{k+1},\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1})+\frac{\rho}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2\right)\\&\geq\frac{m\beta+\rho}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2.\end{aligned}$$

Therefore,

$$L(\mathbf{x}^{k+1},\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1})-L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})\leq-\left(\frac{m\beta}{2}+\rho\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2.$$

For the second term, from the augmented Lagrangian function in (6.3), we have

$$\begin{aligned}&L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^{k+1})-L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^k)\\&=\sum_{i=1}^m\left\langle\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k,\mathbf{x}_i^{k+1}-\mathbf{z}^k\right\rangle\\&\overset{a}{=}\sum_{i\in\mathcal{A}^k}\left\langle\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k,\mathbf{x}_i^{k+1}-\mathbf{z}^k\right\rangle\\&=\sum_{i\in\mathcal{A}^k}\left(\left\langle\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k,\mathbf{x}_i^{k+1}-\mathbf{z}^{\bar{k}_i+1}\right\rangle+\left\langle\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k,\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\rangle\right)\\&\overset{b}{=}\sum_{i\in\mathcal{A}^k}\left(\frac{1}{\beta}\left\|\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k\right\|^2+\left\langle\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^k,\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\rangle\right),\end{aligned}$$

where we use $\boldsymbol{\lambda}_i^{k+1}=\boldsymbol{\lambda}_i^k$ for $i\in\mathcal{A}_c^k$ in $\overset{a}{=}$, and (6.19b) and (6.20) in $\overset{b}{=}$.

For the third term, from the β-strong convexity of $L(\mathbf{x},\mathbf{z},\boldsymbol{\lambda})$ with respect to $\mathbf{x}_i$, we have

$$\begin{aligned}&L(\mathbf{x}^{k+1},\mathbf{z}^k,\boldsymbol{\lambda}^k)-L(\mathbf{x}^k,\mathbf{z}^k,\boldsymbol{\lambda}^k)\\&\overset{c}{=}\sum_{i\in\mathcal{A}^k}\left[\left(f_i(\mathbf{x}_i^{k+1})+\left\langle\boldsymbol{\lambda}_i^k,\mathbf{x}_i^{k+1}-\mathbf{z}^k\right\rangle+\frac{\beta}{2}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^k\right\|^2\right)\right.\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(f_i(\mathbf{x}_i^k) + \left\langle \boldsymbol{\lambda}_i^k, \mathbf{x}_i^k - \mathbf{z}^k \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}_i^k - \mathbf{z}^k \right\|^2 \right) \Big] \\
&\leq \sum_{i \in \mathcal{A}^k} \left(\left\langle \nabla f_i(\mathbf{x}_i^{k+1}) + \boldsymbol{\lambda}_i^k + \beta(\mathbf{x}_i^{k+1} - \mathbf{z}^k), \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\rangle - \frac{\beta}{2} \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2 \right) \\
&\overset{d}{=} \sum_{i \in \mathcal{A}^k} \left(\beta \left\langle \mathbf{z}^{\bar{k}_i+1} - \mathbf{z}^k, \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\rangle - \frac{\beta}{2} \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2 \right),
\end{aligned} \tag{6.24}
$$

where we use $\mathbf{x}_i^{k+1} = \mathbf{x}_i^k$ for $i \in \mathcal{A}_c^k$ in $\overset{c}{=}$, and the optimality condition of (6.19a) and (6.20) in $\overset{d}{=}$.

Thus, we have

$$
\begin{aligned}
& L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L(\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k) \\
&\quad \leq - \left(\frac{m\beta}{2} + \rho \right) \left\| \mathbf{z}^{k+1} - \mathbf{z}^k \right\|^2 - \frac{\beta}{2} \sum_{i \in \mathcal{A}^k} \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2 \\
&\qquad + \sum_{i \in \mathcal{A}^k} \left(\frac{1}{\beta} \left\| \boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^k \right\|^2 + \left\langle \boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^k, \mathbf{z}^{\bar{k}_i+1} - \mathbf{z}^k \right\rangle \right. \\
&\qquad \qquad \left. + \beta \left\langle \mathbf{z}^{\bar{k}_i+1} - \mathbf{z}^k, \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\rangle \right).
\end{aligned}
$$

From (6.19a)–(6.19b) and (6.21)–(6.22), for any i, we have

$$
\mathbf{0} = \nabla f_i(\mathbf{x}_i^{k+1}) + \boldsymbol{\lambda}_i^{k+1}. \tag{6.25}
$$

From the L-smoothness of f_i and $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{\alpha}{2}\|\mathbf{a}\|^2 + \frac{1}{2\alpha}\|\mathbf{b}\|^2$ for $\alpha > 0$, we have

$$
\|\boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^k\| \leq L\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|
$$

and

$$
\begin{aligned}
& L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - L(\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k) \\
&\quad \leq - \left(\frac{m\beta}{2} + \rho \right) \left\| \mathbf{z}^{k+1} - \mathbf{z}^k \right\|^2 - \sum_{i \in \mathcal{A}^k} \left(\frac{\beta}{2} - \frac{L^2}{\beta} - \frac{L^2}{2} - \frac{1}{2} \right) \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2 \\
&\qquad + \sum_{i \in \mathcal{A}^k} \frac{1 + \beta^2}{2} \left\| \mathbf{z}^{\bar{k}_i+1} - \mathbf{z}^k \right\|^2.
\end{aligned} \tag{6.26}
$$

Now, we bound the last term in (6.26). It is easy to show that

$$
\begin{aligned}
\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\|^2 &= \sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\left\|\sum_{t=\bar{k}_i+1}^{k-1}\left(\mathbf{z}^t-\mathbf{z}^{t+1}\right)\right\|^2\\
&\le \sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\left(k-\bar{k}_i-1\right)\sum_{t=\bar{k}_i+1}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le \sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}(\tau-1)\sum_{t=\max\{k-\tau+1,1\}}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le m(\tau-1)\sum_{k=0}^{K}\sum_{t=\max\{k-\tau+1,1\}}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le m(\tau-1)^2\sum_{k=0}^{K}\left\|\mathbf{z}^k-\mathbf{z}^{k+1}\right\|^2
\end{aligned}
\tag{6.27}
$$

due to

$$
\max\{k-\tau,0\}\le\bar{k}_i<k \quad\text{and}\quad |\mathcal{A}^k|\le m.
$$

Thus we have

$$
\begin{aligned}
&L(\mathbf{x}^{K+1},\mathbf{z}^{K+1},\boldsymbol{\lambda}^{K+1})-L(\mathbf{x}^0,\mathbf{z}^0,\boldsymbol{\lambda}^0)\\
&\quad\le -\sum_{k=0}^{K}\left[\left(\frac{m\beta}{2}+\rho\right)-\frac{(1+\beta^2)m(\tau-1)^2}{2}\right]\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2\\
&\quad\quad -\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\left(\frac{\beta}{2}-\frac{L^2}{\beta}-\frac{L^2}{2}-\frac{1}{2}\right)\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^k\right\|^2.
\end{aligned}
$$

Letting ρ and β be large enough such that

$$
\frac{\beta}{2}-\frac{L^2}{\beta}-\frac{L^2}{2}-\frac{1}{2}>0 \quad\text{and}\quad \left(\frac{m\beta}{2}+\rho\right)-\frac{(1+\beta^2)m(\tau-1)^2}{2}>0,
$$

from the assumption that $(\mathbf{x}^{K+1},\mathbf{z}^{K+1},\boldsymbol{\lambda}^{K+1})$ is bounded, we have

$$
\mathbf{z}^{k+1}-\mathbf{z}^k\to\mathbf{0} \quad\text{and}\quad \mathbf{x}_i^{k+1}-\mathbf{x}_i^k\to\mathbf{0},\quad \forall i\in\mathcal{A}^k.
$$

From (6.25) and the smoothness of f_i, we have

$$\boldsymbol{\lambda}_i^{k+1} - \boldsymbol{\lambda}_i^k \to \mathbf{0}, \quad \forall i \in \mathcal{A}^k.$$

From (6.19b), we have

$$\mathbf{x}_i^{k+1} - \mathbf{z}^{\bar{k}_i+1} \to \mathbf{0}, \quad \forall i \in \mathcal{A}^k,$$

which further gives

$$\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1} \to \mathbf{0}, \quad \forall i \in \mathcal{A}^k,$$

due to

$$\max\{k-\tau, 0\} \le \bar{k}_i < k \quad \text{and} \quad \mathbf{z}^{\bar{k}_i+1} - \mathbf{z}^{k+1} \to \mathbf{0}.$$

For any $i \in \mathcal{A}_c^k$, we have $i \in \mathcal{A}^{\tilde{k}_i}$ and

$$\begin{aligned}
\left\|\mathbf{z}^{k+1} - \mathbf{x}_i^{k+1}\right\| &= \left\|\mathbf{z}^{k+1} - \mathbf{x}_i^{\tilde{k}_i+1}\right\| \\
&\le \left\|\mathbf{z}^{k+1} - \mathbf{z}^{\hat{k}_i+1}\right\| + \left\|\mathbf{z}^{\hat{k}_i+1} - \mathbf{x}_i^{\tilde{k}_i+1}\right\| \\
&\overset{a}{=} \left\|\mathbf{z}^{k+1} - \mathbf{z}^{\hat{k}_i+1}\right\| + \frac{1}{\beta}\left\|\boldsymbol{\lambda}_i^{\tilde{k}_i} - \boldsymbol{\lambda}_i^{\tilde{k}_i+1}\right\| \to 0,
\end{aligned}$$

where $\overset{a}{=}$ uses (6.22) and (6.23). So we have

$$\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1} \to \mathbf{0}, \quad \forall i.$$

Then from the optimality condition of (6.19c), we have

$$\sum_{i=1}^{m} \boldsymbol{\lambda}_i^k \to \mathbf{0}.$$

□

6.3.2 Linear Convergence Rate

When we further assume that each f_i is strongly convex, we have the linear convergence rate.

Theorem 6.10 *Assume that each f_i is μ-strongly convex and L-smooth, $i \in [m]$, and Assumption 3 holds true. Let β and ρ be large enough such that*

$$\begin{aligned} 8m(\beta-\mu) &\le \rho, \\ \frac{m\beta+2\rho}{2} - 1 - \tau 2^{2\tau} - \left(\frac{1+\beta^2}{2} + \frac{1}{2m}\right) m\tau 2^{\tau} &> 0, \ \textit{and} \\ \frac{\beta}{2} - \frac{L^2}{\beta} - \frac{L^2}{2} - \frac{1}{2} - \frac{L^2}{4m\beta^2} - \frac{L^2}{4m\beta^2} 2^{\tau-1}\tau &> 0. \end{aligned}$$

Then we have

$$L(\mathbf{x}^{K+1}, \mathbf{z}^{K+1}, \boldsymbol{\lambda}^{K+1}) - f^* \le \left(1 + \frac{1}{\delta\rho}\right)^{-(K+1)} \left(L(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\lambda}^0) - f^*\right),$$

where $\delta \ge \max\left\{1, \frac{1}{\rho}, \frac{\rho+m\beta}{m\mu} - 1\right\}$.

Proof From the strong convexity of f_i and (6.25), we have

$$f_i(\mathbf{z}^*) - f_i(\mathbf{x}_i^{k+1}) \ge -\left\langle \boldsymbol{\lambda}_i^{k+1}, \mathbf{z}^* - \mathbf{x}_i^{k+1} \right\rangle + \frac{\mu}{2} \left\| \mathbf{z}^* - \mathbf{x}_i^{k+1} \right\|^2.$$

From the optimality condition of (6.19c), we have

$$-\sum_{i=1}^m \left[\boldsymbol{\lambda}_i^{k+1} + \beta(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1})\right] + \rho(\mathbf{z}^{k+1} - \mathbf{z}^k) = \mathbf{0}.$$

So we have

$$\sum_{i=1}^m \left\langle \boldsymbol{\lambda}_i^{k+1} + \beta(\mathbf{x}_i^{k+1} - \mathbf{z}^{k+1}), \mathbf{z}^{k+1} - \mathbf{z}^* \right\rangle = \rho \left\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^{k+1} - \mathbf{z}^* \right\rangle$$

and

$$\begin{aligned} &\sum_{i=1}^m f_i(\mathbf{z}^*) - \sum_{i=1}^m f_i(\mathbf{x}_i^{k+1}) \\ &\ge -\sum_{i=1}^m \left\langle \boldsymbol{\lambda}_i^{k+1}, \mathbf{z}^{k+1} - \mathbf{x}_i^{k+1} \right\rangle + \frac{\mu}{2} \sum_{i=1}^m \left\| \mathbf{z}^* - \mathbf{x}_i^{k+1} \right\|^2 \\ &\quad + \rho \left\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^{k+1} - \mathbf{z}^* \right\rangle - \beta \sum_{i=1}^m \left\langle \mathbf{x}_i^{k+1} - \mathbf{z}^{k+1}, \mathbf{z}^{k+1} - \mathbf{z}^* \right\rangle \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{m}\left\langle \boldsymbol{\lambda}_i^{k+1}, \mathbf{z}^{k+1}-\mathbf{x}_i^{k+1}\right\rangle + \frac{\mu}{2}\sum_{i=1}^{m}\left\|\mathbf{z}^{*}-\mathbf{x}_i^{k+1}\right\|^2 \\
&\quad + \frac{\rho+m\beta}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^2 - \frac{\rho}{2}\left\|\mathbf{z}^{k}-\mathbf{z}^{*}\right\|^2 + \frac{\rho}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2 \\
&\quad - \frac{\beta}{2}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{*}\right\|^2 + \frac{\beta}{2}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right\|^2.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - f^{*} \\
&\quad \le \frac{\beta-\mu}{2}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{*}\right\|^2 + \frac{\rho}{2}\left\|\mathbf{z}^{k}-\mathbf{z}^{*}\right\|^2 \\
&\qquad - \frac{\rho+m\beta}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^2 - \frac{\rho}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2.
\end{aligned}
$$

We want to eliminate the first three terms. Since

$$
\begin{aligned}
\frac{\beta-\mu}{2}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{*}\right\|^2 \le& \frac{(\beta-\mu)(1+\delta)}{2}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right\|^2 \\
&+ \frac{(\beta-\mu)m}{2}\left(1+\frac{1}{\delta}\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^2 \quad \text{and} \\
\frac{\rho}{2}\left\|\mathbf{z}^{k}-\mathbf{z}^{*}\right\|^2 \le& \frac{\rho}{2}(1+\delta)\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2 + \frac{\rho}{2}\left(1+\frac{1}{\delta}\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^2,
\end{aligned}
$$

we have

$$
\begin{aligned}
&L(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}, \boldsymbol{\lambda}^{k+1}) - f^{*} \\
&\le \frac{\rho\delta}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2 + \left[\frac{\rho+m(\beta-\mu)}{2\delta} - \frac{m\mu}{2}\right]\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^2 \\
&\quad + (\beta-\mu)\delta\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right\|^2 \\
&\le \frac{\rho\delta}{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2 + (\beta-\mu)\delta\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right\|^2
\end{aligned}
$$

by letting $\delta > 1$ be large enough such that

$$
\frac{\rho+m(\beta-\mu)}{2\delta} - \frac{m\mu}{2} \le 0.
$$

Since

$$
\begin{aligned}
&\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{k+1}\right\|^2\\
&=\sum_{i\in\mathcal{A}^k}\left(\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{\bar{k}_i+1}+\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\quad+\sum_{i\in\mathcal{A}_c^k}\left(\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{\hat{k}_i+1}+\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\le\sum_{i\in\mathcal{A}^k}\left(2\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{\bar{k}_i+1}\right\|^2+2\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\quad+\sum_{i\in\mathcal{A}_c^k}\left(2\left\|\mathbf{x}_i^{k+1}-\mathbf{z}^{\hat{k}_i+1}\right\|^2+2\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\overset{a}{=}\sum_{i\in\mathcal{A}^k}\left(\frac{2}{\beta^2}\left\|\boldsymbol{\lambda}_i^{k+1}-\boldsymbol{\lambda}_i^{k}\right\|^2+2\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\quad+\sum_{i\in\mathcal{A}_c^k}\left(\frac{2}{\beta^2}\left\|\boldsymbol{\lambda}_i^{\tilde{k}_i+1}-\boldsymbol{\lambda}_i^{\tilde{k}_i}\right\|^2+2\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^{k+1}\right\|^2\right)\\
&\overset{b}{\le}\sum_{i\in\mathcal{A}^k}\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^{k}\right\|^2+4\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^{k}\right\|^2\right)\\
&\quad+\sum_{i\in\mathcal{A}_c^k}\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{\tilde{k}_i+1}-\mathbf{x}_i^{\tilde{k}_i}\right\|^2+4\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^{k}\right\|^2\right)+4m\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2,
\end{aligned}
$$

where $\overset{a}{=}$ uses (6.19b), (6.20), (6.22), and (6.23) and $\overset{b}{\le}$ uses (6.25) to replace $\boldsymbol{\lambda}_i^k$ with $-\nabla f_i(\mathbf{x}_i^k)$ and then apply the L-smoothness of f_i. $\overset{b}{\le}$ also uses the inequality $\|\mathbf{a}+\mathbf{b}\|^2\le 2(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)$.

By letting ρ be large enough such that $8m(\beta-\mu)\le\rho$, we have

$$
\begin{aligned}
&L(\mathbf{x}^{k+1},\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1})-f^*\\
&\le\left[\frac{\rho\delta}{2}+4m(\beta-\mu)\delta\right]\left\|\mathbf{z}^{k+1}-\mathbf{z}^{k}\right\|^2\\
&\quad+\sum_{i\in\mathcal{A}^k}(\beta-\mu)\delta\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^{k}\right\|^2+4\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^{k}\right\|^2\right)
\end{aligned}
$$

$$+\sum_{i\in\mathcal{A}_c^k}(\beta-\mu)\delta\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{\widetilde{k}_i+1}-\mathbf{x}_i^{\widetilde{k}_i}\right\|^2+4\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^k\right\|^2\right)$$

$$\leq\rho\delta\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2+\sum_{i\in\mathcal{A}^k}\frac{\rho\delta}{8m}\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^k\right\|^2+4\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\|^2\right)$$

$$+\sum_{i\in\mathcal{A}_c^k}\frac{\rho\delta}{8m}\left(\frac{2L^2}{\beta^2}\left\|\mathbf{x}_i^{\widetilde{k}_i+1}-\mathbf{x}_i^{\widetilde{k}_i}\right\|^2+4\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^k\right\|^2\right).$$

Dividing both sides of the above inequality by $\rho\delta$ and adding it with (6.26), we have

$$\left(L(\mathbf{x}^{k+1},\mathbf{z}^{k+1},\boldsymbol{\lambda}^{k+1})-f^*\right)-\frac{1}{\eta}\left(L(\mathbf{x}^k,\mathbf{z}^k,\boldsymbol{\lambda}^k)-f^*\right)$$

$$\leq\frac{1}{\eta}\left[\sum_{i\in\mathcal{A}_c^k}\frac{L^2}{4m\beta^2}\left\|\mathbf{x}_i^{\widetilde{k}_i+1}-\mathbf{x}_i^{\widetilde{k}_i}\right\|^2+\sum_{i\in\mathcal{A}_c^k}\frac{1}{2m}\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^k\right\|^2\right.$$

$$-\left(\frac{m\beta+\rho}{2}-1\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2+\sum_{i\in\mathcal{A}^k}\left(\frac{1+\beta^2}{2}+\frac{1}{2m}\right)\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\|^2$$

$$\left.-\sum_{i\in\mathcal{A}^k}\left(\frac{\beta}{2}-\frac{L^2}{\beta}-\frac{L^2}{2}-\frac{1}{2}-\frac{L^2}{4m\beta^2}\right)\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^k\right\|^2\right]$$

$$=\frac{1}{\eta}\left[\sum_{i\in\mathcal{A}_c^k}\frac{L^2}{4m\beta^2}\left\|\mathbf{x}_i^{\widetilde{k}_i+1}-\mathbf{x}_i^{\widetilde{k}_i}\right\|^2+\sum_{i\in\mathcal{A}_c^k}\frac{1}{2m}\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^k\right\|^2\right.$$

$$-\left(\frac{m\beta+\rho}{2}-1\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2+\sum_{i\in\mathcal{A}^k}\left(\frac{1+\beta^2}{2}+\frac{1}{2m}\right)\left\|\mathbf{z}^{\bar{k}_i+1}-\mathbf{z}^k\right\|^2$$

$$\left.-\sum_{i=1}^{m}\left(\frac{\beta}{2}-\frac{L^2}{\beta}-\frac{L^2}{2}-\frac{1}{2}-\frac{L^2}{4m\beta^2}\right)\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^k\right\|^2\right],$$

where we denote $\eta=1+\frac{1}{\rho\delta}$ and use $\mathbf{x}_i^{k+1}=\mathbf{x}_i^k$ for all $i\in\mathcal{A}_c^k$ in the last line.

Telescoping the above inequality from $k=0$ to K, we have

$$\left(L(\mathbf{x}^{K+1},\mathbf{z}^{K+1},\boldsymbol{\lambda}^{K+1})-f^*\right)-\frac{1}{\eta^{K+1}}\left(L(\mathbf{x}^0,\mathbf{z}^0,\boldsymbol{\lambda}^0)-f^*\right)$$

$$\leq\frac{L^2}{4m\beta^2}\sum_{k=0}^{K}\frac{1}{\eta^{K+1-k}}\sum_{i\in\mathcal{A}_c^k}\left\|\mathbf{x}_i^{\widetilde{k}_i+1}-\mathbf{x}_i^{\widetilde{k}_i}\right\|^2$$

$$
\begin{aligned}
&+\frac{1}{2m}\sum_{k=0}^{K}\frac{1}{\eta^{K+1-k}}\sum_{i\in\mathcal{A}_c^k}\left\|\mathbf{z}^{\hat{k}_i+1}-\mathbf{z}^k\right\|^2\\
&-\left(\frac{m\beta+\rho}{2}-1\right)\sum_{k=0}^{K}\frac{1}{\eta^{K+1-k}}\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2\\
&+\left(\frac{1+\beta^2}{2}+\frac{1}{2m}\right)\sum_{k=0}^{K}\frac{1}{\eta^{K+1-k}}\sum_{i\in\mathcal{A}^k}\left\|\mathbf{z}^{\overline{k}_i+1}-\mathbf{z}^k\right\|^2\\
&-\left(\frac{\beta}{2}-\frac{L^2}{\beta}-\frac{L^2}{2}-\frac{1}{2}-\frac{L^2}{4m\beta^2}\right)\sum_{k=0}^{K}\frac{1}{\eta^{K+1-k}}\sum_{i=1}^{m}\left\|\mathbf{x}_i^{k+1}-\mathbf{x}_i^k\right\|^2.
\end{aligned}
$$

We want to choose β and ρ large enough such that the right hand side is negative. Similar to (6.27), we have

$$
\begin{aligned}
&\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\eta^k\left\|\mathbf{z}^{\overline{k}_i+1}-\mathbf{z}^k\right\|^2\\
&=\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\eta^k\left\|\sum_{t=\overline{k}_i+1}^{k-1}\left(\mathbf{z}^t-\mathbf{z}^{t+1}\right)\right\|^2\\
&\le\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}\left(k-\overline{k}_i-1\right)\eta^k\sum_{t=\overline{k}_i+1}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le\sum_{k=0}^{K}\sum_{i\in\mathcal{A}^k}(\tau-1)\eta^k\sum_{t=\max\{k-\tau+1,1\}}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le m(\tau-1)\sum_{k=0}^{K}\eta^k\sum_{t=\max\{k-\tau+1,1\}}^{k-1}\left\|\mathbf{z}^t-\mathbf{z}^{t+1}\right\|^2\\
&\le m(\tau-1)\sum_{k=0}^{K}\left(\eta^{k+1}+\eta^{k+2}+\cdots+\eta^{k+\tau-1}\right)\left\|\mathbf{z}^k-\mathbf{z}^{k+1}\right\|^2\\
&\le m(\tau-1)\frac{\eta^\tau-\eta}{\eta-1}\sum_{k=0}^{K}\eta^k\left\|\mathbf{z}^k-\mathbf{z}^{k+1}\right\|^2.
\end{aligned}
$$

Analogously, we have

$$\begin{aligned}
&\sum_{k=0}^{K} \sum_{i \in \mathcal{A}_c^k} \eta^k \left\| \mathbf{z}^{\hat{k}_i+1} - \mathbf{z}^k \right\|^2 \\
&\quad \leq m(2\tau - 1) \frac{\eta^{2\tau} - \eta}{\eta - 1} \sum_{k=0}^{K} \eta^k \left\| \mathbf{z}^k - \mathbf{z}^{k+1} \right\|^2,
\end{aligned}$$

due to

$$\max\{k - \tau + 1, 0\} \leq \widetilde{k}_i < k, \quad \max\{\widetilde{k}_i - \tau, 0\} \leq \hat{k}_i < \widetilde{k}_i,$$

and thus

$$\max\{k - 2\tau + 1, 0\} \leq \hat{k}_i < k.$$

We also have

$$\begin{aligned}
&\sum_{k=0}^{K} \sum_{i \in \mathcal{A}_c^k} \eta^k \left\| \mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{x}_i^{\widetilde{k}_i} \right\|^2 \\
&= \sum_{k=0}^{K} \sum_{i \in \mathcal{A}_c^k} \eta^{k-\widetilde{k}_i} \eta^{\widetilde{k}_i} \left\| \mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{x}_i^{\widetilde{k}_i} \right\|^2 \\
&\leq \eta^{\tau-1} \sum_{k=0}^{K} \sum_{i \in \mathcal{A}_c^k} \eta^{\widetilde{k}_i} \left\| \mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{x}_i^{\widetilde{k}_i} \right\|^2 \\
&\overset{a}{\leq} \eta^{\tau-1}(\tau - 1) \sum_{k=0}^{K} \sum_{i=1}^{m} \eta^k \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2,
\end{aligned}$$

where in $\overset{a}{\leq}$ we use the fact that each $\eta^{\widetilde{k}_i} \left\| \mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{x}_i^{\widetilde{k}_i} \right\|^2$ appears no more than $\tau - 1$ times in the summation $\sum_{k=0}^{K} \sum_{i \in \mathcal{A}_c^k} \eta^{\widetilde{k}_i} \left\| \mathbf{x}_i^{\widetilde{k}_i+1} - \mathbf{x}_i^{\widetilde{k}_i} \right\|^2$.

Thus, we have

$$\begin{aligned}
&\left(L(\mathbf{x}^{K+1}, \mathbf{z}^{K+1}, \boldsymbol{\lambda}^{K+1}) - f^* \right) - \frac{1}{\eta^{K+1}} \left(L(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\lambda}^0) - f^* \right) \\
&\leq - \Bigg[\frac{m\beta + \rho}{2} - 1 - \frac{1}{2m} m(2\tau - 1) \frac{\eta^{2\tau} - \eta}{\eta - 1}
\end{aligned}$$

$$
\begin{aligned}
&- \left(\frac{1+\beta^2}{2} + \frac{1}{2m}\right) m(\tau-1)\frac{\eta^\tau - \eta}{\eta - 1}\Bigg] \sum_{k=0}^{K} \frac{1}{\eta^{K+1-k}} \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 \\
&- \left[\frac{\beta}{2} - \frac{L^2}{\beta} - \frac{L^2}{2} - \frac{1}{2} - \frac{L^2}{4m\beta^2} - \frac{L^2}{4m\beta^2}\eta^{\tau-1}(\tau-1)\right] \\
&\times \sum_{i=1}^{m}\sum_{k=0}^{K} \frac{1}{\eta^{K+1-k}} \left\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\right\|^2 \\
\le & - \left[\frac{m\beta+\rho}{2} - 1 - \tau 2^{2\tau} - \left(\frac{1+\beta^2}{2} + \frac{1}{2m}\right) m\tau 2^\tau\right] \\
&\times \sum_{k=0}^{K} \frac{1}{\eta^{K+1-k}} \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 \\
&- \left(\frac{\beta}{2} - \frac{L^2}{\beta} - \frac{L^2}{2} - \frac{1}{2} - \frac{L^2}{4m\beta^2} - \frac{L^2}{4m\beta^2}2^{\tau-1}\tau\right) \\
&\times \sum_{i=1}^{m}\sum_{k=0}^{K} \frac{1}{\eta^{K+1-k}} \left\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\right\|^2 \\
\le & \, 0,
\end{aligned}
$$

where we use

$$
\eta \le 2, \quad \tfrac{\eta^\tau - \eta}{\eta - 1} = \eta + \cdots + \eta^{\tau-1} \le 2 + \cdots + 2^{\tau-1} \le 2^\tau, \quad \text{and} \quad \tfrac{\eta^{2\tau} - \eta}{\eta - 1} \le 2^{2\tau}.
$$

□

From Theorem 6.2, we see that the synchronous ADMM needs $O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ iterations to find an ϵ-optimal solution, which has the optimal dependence on $\frac{L}{\mu}$. For the asynchronous ADMM, Theorem 6.10 only proves the linear convergence without a complexity explicitly dependent on $\frac{L}{\mu}$. We believe that in general the asynchronous ADMM needs more iterations than synchronous ADMM. It is unclear whether the time saved per iteration of the asynchronous ADMM can offset the cost of more iterations in theory, although it shows great advantages in practice.

There are some other ways to analyze asynchronous ADMM. For example, [6, 7, 13, 15] studied randomized asynchronous ADMM, which requires more assumptions than Algorithms 6.10–6.11 do, and it is also unclear whether it needs less running time than the synchronous ADMM in theory.

6.4 Nonconvex Distributed ADMM

Next, we introduce the nonconvex distributed ADMM. In fact, the asynchronous ADMM (Algorithm 6.11) can also be used to solve nonconvex problems. In this case, $L(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda})$ is $(\beta - L)$-strongly convex with respect to $\mathbf{x}$, and (6.24) should be replaced by the following one:

$$\begin{aligned} & L(\mathbf{x}^{k+1}, \mathbf{z}^k, \boldsymbol{\lambda}^k) - L(\mathbf{x}^k, \mathbf{z}^k, \boldsymbol{\lambda}^k) \\ & \leq \sum_{i \in \mathcal{A}^k} \left(\beta \left\langle \mathbf{z}^{\overline{k}_i+1} - \mathbf{z}^k, \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\rangle - \frac{\beta - L}{2} \left\| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \right\|^2 \right). \end{aligned}$$

Accordingly, we have the following convergence guarantee [4].

Theorem 6.11 *Assume that each f_i is L-smooth, $i \in [m]$, and Assumption 3 holds true. Let*

$$\beta > \frac{1 + L + L^2 + \sqrt{(1 + L + L^2)^2 + 8L^2}}{2} \quad \textit{and} \quad \rho > \frac{m(1 + \beta^2)(\tau - 1)^2 - m\beta}{2}.$$

Suppose that $(\mathbf{x}_1^k, \cdots, \mathbf{x}_m^k, \mathbf{z}^k, \boldsymbol{\lambda}_1^k, \cdots, \boldsymbol{\lambda}_m^k)$ generated by (6.19a)–(6.19c) are bounded, then $(\mathbf{x}_1^k, \cdots, \mathbf{x}_m^k, \mathbf{z}^k, \boldsymbol{\lambda}_1^k, \cdots, \boldsymbol{\lambda}_m^k)$ converge to the set of KKT points of Problem (6.2) in the sense of

$$\sum_{i=1}^m \boldsymbol{\lambda}_i^k \to \mathbf{0}, \quad \mathbf{x}_i^{k+1} - \mathbf{z}^{k+1} \to \mathbf{0}, \quad \textit{and} \quad \nabla f_i(\mathbf{x}_i^{k+1}) + \boldsymbol{\lambda}_i^{k+1} = \mathbf{0}, \quad i \in [m].$$

The synchronous ADMM is a special case of the asynchronous ADMM with $\mathcal{A}_c^k = \varnothing$ and $\overline{k}_i + 1 = k$. Thus, the above theorem also holds for the synchronous ADMM with a much simpler proof.

6.5 ADMM with Generally Linear Constraints

We end this chapter by non-consensus-based distributed ADMM. Namely, the problem is the generally linearly constrained one (3.71). The linearized ADMM with parallel splitting [9, 11] given in Algorithm 3.11 can be used to solve the problem directly. We present it in Algorithms 6.12–6.13 in the distributed manner. If the proximal mapping of f_i is not easily computable, we may linearize f_i as well, but since this is a straightforward modification over Algorithms 6.12–6.13, we omit the details.

Algorithm 6.12 Distributed linearized ADMM with parallel splitting for the master

for $k = 0, 1, 2, \cdots$ **do**
  Wait until receiving $\mathbf{y}_i^{k+1}$ from all the workers $i \in [m]$.
  $\mathbf{s}^{k+1} = \sum_{i=1}^m \mathbf{y}_i^{k+1}$.
  $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta\left(\mathbf{s}^{k+1} - \mathbf{b}\right)$.
  Send $\mathbf{s}^{k+1}$ and $\boldsymbol{\lambda}^{k+1}$ to all the workers.
end for

Algorithm 6.13 Distributed linearized ADMM with parallel splitting for the ith worker

Initialize: $\mathbf{x}_i^0$ and $\boldsymbol{\lambda}_i^0$, $i \in [m]$.
$\mathbf{y}_i^0 = \mathbf{A}_i\mathbf{x}_i^0$.
Send $\mathbf{y}_i^0$ to the master.
Wait until receiving $\mathbf{s}^0$ and $\boldsymbol{\lambda}^0$ from the master.
for $k = 0, 1, 2, \cdots$ **do**

$$\begin{aligned}
\mathbf{x}_i^{k+1} &= \operatorname*{argmin}_{\mathbf{x}_i} \left(f_i(\mathbf{x}_i) + \left\langle \boldsymbol{\lambda}^k, \mathbf{A}_i\mathbf{x}_i \right\rangle + \beta \left\langle \mathbf{A}_i^T \left(\mathbf{s}^k - \mathbf{b}\right), \mathbf{x}_i - \mathbf{x}_i^k \right\rangle \right. \\
&\qquad \left. + \frac{m\beta\|\mathbf{A}_i\|_2^2}{2} \left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2 \right) \\
&= \operatorname{Prox}_{\left(m\beta\|\mathbf{A}_i\|_2^2\right)^{-1} f_i} \left(\mathbf{x}_i^k - \frac{1}{m\beta\|\mathbf{A}_i\|_2^2} \mathbf{A}_i^T \left[\boldsymbol{\lambda}^k + \beta\left(\mathbf{s}^k - \mathbf{b}\right)\right]\right).
\end{aligned}$$

  $\mathbf{y}_i^{k+1} = \mathbf{A}_i\mathbf{x}_i^{k+1}$.
  Send $\mathbf{y}_i^{k+1}$ to the master.
  Wait until receiving $\mathbf{s}^{k+1}$ and $\boldsymbol{\lambda}^{k+1}$ from the master.
end for

References

1. N.S. Aybat, Z. Wang, T. Lin, S. Ma, Distributed linearized alternating direction mehod of multipliers. IEEE Trans. Automat. Contr. **63**(1), 5–20 (2018)
2. D.P. Bertsekas, J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods* (Prentice Hall, Hoboken, 1989)
3. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
4. T.-H. Chang, M. Hong, X. Wang, Asynchronous distributed ADMM for large-scale optimization – part I: algorithm and convergence analysis. IEEE Trans. Signal Process. **64**(12), 3118–3130 (2016)
5. T.-H. Chang, W.-C. Liao, M. Hong, X. Wang, Asynchronous distributed ADMM for large-scale optimization – part II: linear convergence analysis and numerical performance. IEEE Trans. Signal Process. **64**(12), 3131–3144 (2016)
6. F. Iutzeler, P. Bianchi, P. Ciblat, W. Hachem, Asynchronous distributed optimization using a randomized alternating direction method of multipliers, in *IEEE Conference on Decision and Control* (2013), pp. 3671–3676

7. S. Kumar, R. Jain, K. Rajawat, Asynchronous optimization over heterogeneous networks via consensus ADMM. IEEE Trans. Signal Inf. Process. Netw. **3**(1), 114–129 (2017)
8. H. Li, Z. Lin, Y. Fang, Variance reduced EXTRA and DIGing and their optimal acceleration for strongly convex decentralized optimization (2020). Arxiv:2009.04373
9. Z. Lin, R. Liu, H. Li, Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. Mach. Learn. **99**(2), 287–325 (2015)
10. Q. Ling, W. Shi, G. Wu, A. Ribeiro, DLM: Decentralized linearized alternating direction method of multipliers. IEEE Trans. Signal Process. **63**(15), 4051–4064 (2015)
11. R. Liu, Z. Lin, Z. Su, Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning, in *Asian Conference on Machine Learning* (2013), pp. 116–132
12. M. Maros, J. Jalden, On the Q-linear convergence of distributed generalized ADMM under non-strongly convex function components. IEEE Trans. Signal Inf. Process. Netw. **5**(3), 442–453 (2019)
13. Z. Peng, Y. Xu, M. Yan, W. Yin, ARock: an algorithmic framework for asynchronous parallel coordinate updates. SIAM J. Sci. Comput. **38**(5), 2851–2879 (2016)
14. W. Shi, Q. Ling, K. Yuan, G. Wu, W. Yin, On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. **62**(7), 1750–1761 (2014)
15. E. Wei, A. Ozdaglar, On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers, in *IEEE Global Conference on Signal and Information Processing* (2013), pp. 551–554