

# Supplement:

## Bregman Alternating Direction Method of Multipliers

### 1 Convergence Analysis of BADMM

Bregman ADMM (BADMM) has the following iterates:

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c} \rangle + \rho B_\phi(\mathbf{c} - \mathbf{Ax}, \mathbf{Bz}_t) + \rho_{\mathbf{x}} B_{\varphi_{\mathbf{x}}}(\mathbf{x}, \mathbf{x}_t), \quad (1)$$

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c} \rangle + \rho B_\phi(\mathbf{Bz}, \mathbf{c} - \mathbf{Ax}_{t+1}) + \rho_{\mathbf{z}} B_{\varphi_{\mathbf{z}}}(\mathbf{z}, \mathbf{z}_t), \quad (2)$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \tau(\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}). \quad (3)$$

We need the following assumption in establishing the convergence of BADMM:

#### Assumption 1

(a)  $f : \mathbb{R}^{n_1} \rightarrow \mathbb{R} \cup \{+\infty\}$  and  $g : \mathbb{R}^{n_2} \rightarrow \mathbb{R} \cup \{+\infty\}$  are closed, proper and convex.

(b) An optimal solution exists.

(c) The Bregman divergence  $B_\phi$  is defined on an  $\alpha$ -strongly convex function  $\phi$  with respect to a  $p$ -norm  $\|\cdot\|_p^2$ , i.e.,  $B_\phi(\mathbf{u}, \mathbf{v}) \geq \frac{\alpha}{2} \|\mathbf{u} - \mathbf{v}\|_p^2$ , where  $\alpha > 0$ .

We start with the partial Lagrangian, which is defined as follows:

$$L(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{Ax} + \mathbf{Bz} - \mathbf{c} \rangle. \quad (4)$$

Assume that  $\mathbf{x}^* \in \mathcal{X}$ ,  $\mathbf{z}^* \in \mathcal{Z}$  and  $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$  satisfies the KKT conditions of (4), i.e.,

$$-\mathbf{A}^T \mathbf{y}^* \in \partial f(\mathbf{x}^*), \quad (5)$$

$$-\mathbf{B}^T \mathbf{y}^* \in \partial g(\mathbf{z}^*), \quad (6)$$

$$\mathbf{Ax}^* + \mathbf{Bz}^* - \mathbf{c} = \mathbf{0}. \quad (7)$$

$\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$  is an optimal solution. Considering (1) and (2),  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{z} \in \mathcal{Z}$  are always satisfied in BADMM. Let  $f'(\mathbf{x}_{t+1}) \in \partial f(\mathbf{x}_{t+1})$  and  $g'(\mathbf{z}_{t+1}) \in \partial g(\mathbf{z}_{t+1})$ . For  $\mathbf{x}^* \in \mathcal{X}$  and  $\mathbf{z}^* \in \mathcal{Z}$ , the optimality conditions of (1) and (2) are

$$\langle f'(\mathbf{x}_{t+1}) + \mathbf{A}^T \{ \mathbf{y}_t + \rho(-\nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1}) + \nabla\phi(\mathbf{Bz}_t)) \} + \rho_{\mathbf{x}}(\nabla\varphi_{\mathbf{x}}(\mathbf{x}_{t+1}) - \nabla\varphi_{\mathbf{x}}(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle \leq 0, \quad (8)$$

$$\langle g'(\mathbf{z}_{t+1}) + \mathbf{B}^T \{ \mathbf{y}_t + \rho(\nabla\phi(\mathbf{Bz}_{t+1}) - \nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1})) \} + \rho_{\mathbf{z}}(\nabla\varphi_{\mathbf{z}}(\mathbf{z}_{t+1}) - \nabla\varphi_{\mathbf{z}}(\mathbf{z}_t)), \mathbf{z}_{t+1} - \mathbf{z}^* \rangle \leq 0. \quad (9)$$

If  $\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} = \mathbf{c}$ , then  $\mathbf{y}_{t+1} = \mathbf{y}_t$ . Further, if  $\mathbf{Ax}_{t+1} + \mathbf{Bz}_t = \mathbf{c}$ ,  $\mathbf{x}_{t+1} = \mathbf{x}_t$ , (8) reduces to

$$\langle f'(\mathbf{x}_{t+1}) + \mathbf{A}^T \mathbf{y}_{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \rangle \leq 0 \quad (10)$$

Therefore,  $\mathbf{A}^T \mathbf{y}_{t+1} = f'(\mathbf{x}_{t+1})$  is a sufficient condition and (5) is satisfied. Similarly, (6) is satisfied if  $\mathbf{z}_{t+1} = \mathbf{z}_t$  in (9). Overall, we have the following sufficient conditions for the KKT conditions (5)-(7) to be satisfied:

$$B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t) = 0, B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t) = 0, \quad (11a)$$

$$\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c} = 0, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} = 0. \quad (11b)$$

For the exact BADMM,  $\rho_{\mathbf{x}} = \rho_{\mathbf{z}} = 0$  in (1) and (2), the optimality conditions are (11b), which is equivalent to the optimality conditions used in the proof of ADMM in [2], i.e.,

$$\mathbf{Bz}_{t+1} - \mathbf{Bz}_t = 0, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} = 0. \quad (12)$$

Define the residuals of optimality conditions (11) at  $(t+1)$  as:

$$R(t+1) = \frac{\rho_{\mathbf{x}}}{\rho} B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t) + \frac{\rho_{\mathbf{z}}}{\rho} B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t) + B_{\phi}(\mathbf{c} - \mathbf{Ax}_{t+1}, \mathbf{Bz}_t) + \gamma \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2, \quad (13)$$

where  $\gamma > 0$ . If  $R(t+1) = 0$ , the optimality conditions (11) and (11b) are satisfied. It is sufficient to show the convergence of BADMM by showing  $R(t+1)$  converges to zero. We need the following lemma.

**Lemma 1** *Let the sequence  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1)-(3). For any  $\mathbf{x}^* \in \mathcal{X}, \mathbf{z}^* \in \mathcal{Z}$  satisfying  $\mathbf{Ax}^* + \mathbf{Bz}^* = \mathbf{c}$ , we have*

$$\begin{aligned} & f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq -\langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle - \rho(B_{\phi}(\mathbf{c} - \mathbf{Ax}_{t+1}, \mathbf{Bz}_t) + B_{\phi}(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1})) \\ & + \rho(B_{\phi}(\mathbf{Bz}^*, \mathbf{Bz}_t) - B_{\phi}(\mathbf{Bz}^*, \mathbf{Bz}_{t+1})) + \rho_{\mathbf{x}}(B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t)) \\ & + \rho_{\mathbf{z}}(B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_t) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_{t+1}) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t)). \end{aligned} \quad (14)$$

*Proof:* and its subgradient given in (8) at  $\mathbf{x}^{t+1} \in \mathcal{X}$ , For any  $\mathbf{x} \in \mathcal{X}$ , the optimality condition of (8) is

$$\langle f'(\mathbf{x}_{t+1}) + \mathbf{A}^T \{\mathbf{y}_t + \rho(-\nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1}) + \nabla\phi(\mathbf{Bz}_t))\} + \rho_{\mathbf{x}}(\nabla\varphi_{\mathbf{x}}(\mathbf{x}_{t+1}) - \nabla\varphi_{\mathbf{x}}(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x} \rangle \leq 0. \quad (15)$$

Rearranging the terms using the convexity of  $f$ , we have

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}) \leq \langle f'(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x} \rangle \quad (16)$$

$$\begin{aligned} & \leq -\langle \mathbf{A}^T \{\mathbf{y}_t + \rho(-\nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1}) + \nabla\phi(\mathbf{Bz}_t))\} + \rho_{\mathbf{x}}(\nabla\varphi_{\mathbf{x}}(\mathbf{x}_{t+1}) - \nabla\varphi_{\mathbf{x}}(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x} \rangle \\ & = -\langle \mathbf{y}_t, \mathbf{A}(\mathbf{x}_{t+1} - \mathbf{x}) \rangle + \rho \langle \nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1}) - \nabla\phi(\mathbf{Bz}_t), \mathbf{A}(\mathbf{x}_{t+1} - \mathbf{x}) \rangle \\ & - \rho_{\mathbf{x}} \langle \nabla\varphi_{\mathbf{x}}(\mathbf{x}_{t+1}) - \nabla\varphi_{\mathbf{x}}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x} \rangle. \end{aligned} \quad (17)$$

Setting  $\mathbf{x} = \mathbf{x}^*$  and using  $\mathbf{Ax}^* + \mathbf{Bz}^* = \mathbf{c}$ , we have

$$\begin{aligned} & f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \\ & \leq -\langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz}^* - \mathbf{c} \rangle + \rho \langle \nabla\phi(\mathbf{c} - \mathbf{Ax}_{t+1}) - \nabla\phi(\mathbf{Bz}_t), \mathbf{Bz}^* - (\mathbf{c} - \mathbf{Ax}_{t+1}) \rangle \\ & - \rho_{\mathbf{x}} \langle \nabla\varphi_{\mathbf{x}}(\mathbf{x}_{t+1}) - \nabla\varphi_{\mathbf{x}}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle \\ & = -\langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz}^* - \mathbf{c} \rangle + \rho(B_{\phi}(\mathbf{Bz}^*, \mathbf{Bz}_t) - B_{\phi}(\mathbf{Bz}^*, \mathbf{c} - \mathbf{Ax}_{t+1}) - B_{\phi}(\mathbf{c} - \mathbf{Ax}_{t+1}, \mathbf{Bz}_t)) \\ & + \rho_{\mathbf{x}}(B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t)). \end{aligned} \quad (18)$$

where the last equality uses the three point property of Bregman divergence, i.e.,

$$\langle \nabla \phi(\mathbf{u}) - \nabla \phi(\mathbf{v}), \mathbf{w} - \mathbf{u} \rangle = B_\phi(\mathbf{w}, \mathbf{v}) - B_\phi(\mathbf{w}, \mathbf{u}) - B_\phi(\mathbf{u}, \mathbf{v}). \quad (19)$$

Similarly, using the convexity of  $g$  and its subgradient given in (9) at  $\mathbf{z}^{t+1} \in \mathcal{Z}$ , for any  $\mathbf{z} \in \mathcal{Z}$ ,

$$\begin{aligned} g(\mathbf{z}_{t+1}) - g(\mathbf{z}) &\leq \langle g'(\mathbf{z}_{t+1}), \mathbf{z}_{t+1} - \mathbf{z} \rangle \\ &\leq \langle -\mathbf{B}^T \{ \mathbf{y}_t + \rho(\nabla \phi(\mathbf{B}\mathbf{z}_{t+1}) - \nabla \phi(\mathbf{c} - \mathbf{A}\mathbf{x}_{t+1})) \} - \rho_{\mathbf{z}}(\nabla \varphi_{\mathbf{z}}(\mathbf{z}_{t+1}) - \nabla \varphi_{\mathbf{z}}(\mathbf{z}_t)), \mathbf{z}_{t+1} - \mathbf{z} \rangle \\ &= -\langle \mathbf{y}_t, \mathbf{B}(\mathbf{z}_{t+1} - \mathbf{z}) \rangle + \rho \langle \nabla \phi(\mathbf{B}\mathbf{z}_{t+1}) - \nabla \phi(\mathbf{c} - \mathbf{A}\mathbf{x}_{t+1}), \mathbf{B}\mathbf{z} - \mathbf{B}\mathbf{z}_{t+1} \rangle \\ &\quad - \rho_{\mathbf{z}} \langle \nabla \varphi_{\mathbf{z}}(\mathbf{z}_{t+1}) - \nabla \varphi_{\mathbf{z}}(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z} \rangle \\ &= -\langle \mathbf{y}_t, \mathbf{B}(\mathbf{z}_{t+1} - \mathbf{z}) \rangle + \rho \{ B_\phi(\mathbf{B}\mathbf{z}, \mathbf{c} - \mathbf{A}\mathbf{x}_{t+1}) - B_\phi(\mathbf{B}\mathbf{z}, \mathbf{B}\mathbf{z}_{t+1}) - B_\phi(\mathbf{B}\mathbf{z}_{t+1}, \mathbf{c} - \mathbf{A}\mathbf{x}_{t+1}) \} \\ &\quad + \rho_{\mathbf{z}} (B_{\varphi_{\mathbf{z}}}(\mathbf{z}, \mathbf{z}_t) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}, \mathbf{z}_{t+1}) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t)). \end{aligned} \quad (20)$$

where the last equality uses the three point property of Bregman divergence (19). Set  $\mathbf{z} = \mathbf{z}^*$  in (20). Adding (18) and (20) completes the proof.  $\blacksquare$

Under Assumption 1(c), the following lemma shows that (13) is bounded by a telescoping series of  $D(\mathbf{w}^*, \mathbf{w}_t) - D(\mathbf{w}^*, \mathbf{w}_{t+1})$ , where  $D(\mathbf{w}^*, \mathbf{w}_t)$  defines the distance from the current iterate  $\mathbf{w}_t = (\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t)$  to a KKT point  $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*)$  as follows:

$$D(\mathbf{w}^*, \mathbf{w}_t) = \frac{1}{2\tau\rho} \|\mathbf{y}^* - \mathbf{y}_t\|_2^2 + B_\phi(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_t) + \frac{\rho_{\mathbf{x}}}{\rho} B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) + \frac{\rho_{\mathbf{z}}}{\rho} B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_t). \quad (21)$$

**Lemma 2** *Let the sequence  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1)-(3),  $\mathbf{x}^* \in \mathcal{X}, \mathbf{z}^* \in \mathcal{Z}$  and  $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$  satisfying (5)-(7). Let the Assumption 1 hold.  $R(t+1)$  and  $D(\mathbf{w}^*, \mathbf{w}_t)$  are defined in (13) and (21) respectively. Set  $\tau \leq (\alpha\sigma - 2\gamma)\rho$ , where  $\sigma = \min\{1, m^{\frac{2}{p}-1}\}$  and  $0 < \gamma < \frac{\alpha\sigma}{2}$ . Then*

$$R(t+1) \leq D(\mathbf{w}^*, \mathbf{w}_t) - D(\mathbf{w}^*, \mathbf{w}_{t+1}). \quad (22)$$

*Proof:* Assume  $\mathbf{x}^* \in \mathcal{X}$  and  $\{\mathbf{x}^*, \mathbf{y}^*\}$  satisfies (5). Since  $f$  is convex, then

$$f(\mathbf{x}^*) - f(\mathbf{x}_{t+1}) \leq -\langle \mathbf{A}^T \mathbf{y}^*, \mathbf{x}^* - \mathbf{x}_{t+1} \rangle = -\langle \mathbf{y}^*, \mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}_{t+1} \rangle. \quad (23)$$

Similarly, for convex function  $g$ ,  $\mathbf{z}^* \in \mathcal{Z}$  and  $\{\mathbf{z}^*, \mathbf{y}^*\}$  satisfying (6), we have

$$g(\mathbf{z}^*) - g(\mathbf{z}_{t+1}) \leq -\langle \mathbf{B}^T \mathbf{y}^*, \mathbf{z}^* - \mathbf{z}_{t+1} \rangle = -\langle \mathbf{y}^*, \mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1} \rangle. \quad (24)$$

Adding them together and using the fact that  $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$ , we have

$$f(\mathbf{x}^*) + g(\mathbf{z}^*) - (f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1})) \leq \langle \mathbf{y}^*, \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} \rangle. \quad (25)$$

Adding (25) and (14) together yields

$$\begin{aligned} 0 &\leq \langle \mathbf{y}^* - \mathbf{y}_t, \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} \rangle - \rho(B_\phi(\mathbf{c} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t) + B_\phi(\mathbf{B}\mathbf{z}_{t+1}, \mathbf{c} - \mathbf{A}\mathbf{x}_{t+1})) \\ &\quad + \rho(B_\phi(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_t) - B_\phi(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_{t+1})) + \rho_{\mathbf{x}}(B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t)) \\ &\quad + \rho_{\mathbf{z}}(B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_t) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_{t+1}) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t)). \end{aligned} \quad (26)$$

Using  $\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} = \frac{1}{\tau}(\mathbf{y}_{t+1} - \mathbf{y}_t)$ , the first term can be rewritten as

$$\begin{aligned} \langle \mathbf{y}^* - \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle &= \frac{1}{\tau} \langle \mathbf{y}^* - \mathbf{y}_t, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle \\ &= \frac{1}{2\tau} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2 + \|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2) \\ &= \frac{1}{2\tau} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2) + \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2. \end{aligned} \quad (27)$$

Plugging into (26) and rearranging the terms, we have

$$\begin{aligned} &\frac{1}{2\tau} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2) + \rho(B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_t) - B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_{t+1})) \\ &\rho_{\mathbf{x}}(B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_{t+1})) + \rho_{\mathbf{z}}(B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_t) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_{t+1})) \\ &\geq \rho_{\mathbf{x}} B_{\varphi_{\mathbf{x}}}(\mathbf{x}_{t+1}, \mathbf{x}_t) + \rho_{\mathbf{z}} B_{\varphi_{\mathbf{z}}}(\mathbf{z}_{t+1}, \mathbf{z}_t) + \rho B_\phi(\mathbf{c} - \mathbf{Ax}_{t+1}, \mathbf{Bz}_t) \\ &+ \rho B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) - \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2. \end{aligned} \quad (28)$$

Dividing both sides by  $\rho$  and letting  $R(t+1)$  and  $D(\mathbf{w}^*, \mathbf{w}_t)$  be defined in (13) and (21) respectively, we have

$$\begin{aligned} D(\mathbf{w}^*, \mathbf{w}_t) - D(\mathbf{w}^*, \mathbf{w}_{t+1}) &\geq R(t+1) + B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) - (\frac{\tau}{2\rho} + \gamma) \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 \\ &\geq R(t+1) + \frac{\alpha}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_p^2 - (\frac{\tau}{2\rho} + \gamma) \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2, \end{aligned} \quad (29)$$

where the last inequality uses the Assumption 1(c).

If  $0 < p \leq 2$ ,  $\|\mathbf{u}\|_p \geq \|\mathbf{u}\|_2$ . Set  $\frac{\alpha}{2} \geq \frac{\tau}{2\rho} + \gamma$  in (29), i.e.,  $\tau \leq (\alpha - 2\gamma)\rho$ . We can always find a  $\gamma < \frac{\alpha}{2}$ , thus (22) follows.

If  $p > 2$ ,  $\|\mathbf{u}\|_2 \leq m^{\frac{1}{2} - \frac{1}{p}} \|\mathbf{u}\|_p$  for any  $\mathbf{u} \in \mathbb{R}^{m \times 1}$ , so  $\|\mathbf{u}\|_p^2 \geq m^{\frac{2}{p} - 1} \|\mathbf{u}\|_2^2$ . In (29), set  $\frac{\alpha}{2} m^{\frac{2}{p} - 1} \geq \frac{\tau}{2\rho} + \gamma$ , i.e.,  $\tau \leq (\alpha m^{\frac{2}{p} - 1} - 2\gamma)\rho$ . As long as  $\gamma < \frac{\alpha}{2} m^{\frac{2}{p} - 1}$ , we have (22).  $\blacksquare$

**Remark 1** (a) If  $0 < p \leq 2$ , then  $\sigma = 1$  and  $\tau \leq (\alpha - 2\gamma)\rho$ . The case that  $0 < p \leq 2$  includes two widely used Bregman divergences, i.e., Euclidean distance and KL divergence. For KL divergence in the unit simplex, we have  $\alpha = 1, p = 1$  in the Assumption 1 (c), i.e.,  $KL(\mathbf{u}, \mathbf{v}) \geq \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_1^2$  [1].

(b) Since we often set  $B_\phi$  to be a quadratic function ( $p = 2$ ), the three special cases in Section 2.1 could choose step size  $\tau = (\alpha - 2\gamma)\rho$ .

(c) If  $p > 2$ ,  $\sigma$  will be small, leading to a small step size  $\tau$  which may be not be necessary in practice. It would be interesting to see whether a large step size can be used for any  $p > 0$ .

The following theorem establishes the global convergence for BADMM.

**Theorem 1** Let the sequence  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1)-(3),  $\mathbf{x}^* \in \mathcal{X}, \mathbf{z}^* \in \mathcal{Z}$  and  $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$  satisfying (5)-(7). Let the Assumption 1 hold and  $\tau, \gamma$  satisfy the conditions in Lemma 2. Then  $R(t+1)$  converges to zero and  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  converges to a KKT point  $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$ .

*Proof:* Since  $R(t+1) \geq 0$ , (22) implies  $D(\mathbf{w}^*, \mathbf{w}_{t+1}) \leq D(\mathbf{w}^*, \mathbf{w}_t)$ . Therefore,  $D(\mathbf{w}^*, \mathbf{w}_t)$  is monotonically nonincreasing and  $\mathbf{w}_t$  converges to a KKT point  $\mathbf{w}^*$ . Summing (22) over  $t$  from 0 to  $\infty$  yields

$$\sum_{t=0}^{\infty} R(t+1) \leq D(\mathbf{w}^*, \mathbf{w}_0). \quad (30)$$

Since  $R(t+1) \geq 0$ ,  $R(t+1) \rightarrow 0$  as  $t \rightarrow \infty$ , which completes the proof.  $\blacksquare$

The following theorem establishes a  $O(1/T)$  convergence rate for the objective and residual of constraints in an ergodic sense.

**Theorem 2** *Let the sequences  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1),(2),(3) and  $\mathbf{y}_0 = \mathbf{0}$ . Let  $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$ ,  $\bar{\mathbf{z}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t$ . Set  $\tau \leq (\alpha\sigma - 2\gamma)\rho$ , where  $\sigma = \min\{1, m^{\frac{2}{p}-1}\}$  and  $0 < \gamma < \frac{\alpha\sigma}{2}$ . For any  $\mathbf{x}^* \in \mathcal{X}$ ,  $\mathbf{z}^* \in \mathcal{Z}$  and  $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*)$  satisfying KKT conditions (5)-(7), we have*

$$f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \leq \frac{D_1}{T}, \quad (31)$$

$$\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T - \mathbf{c}\|_2^2 \leq \frac{D(\mathbf{w}^*, \mathbf{w}_0)}{\gamma T}, \quad (32)$$

where  $D_1 = \rho B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_0) + \rho_x B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_0) + \rho_z B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_0)$ .

*Proof:* Using (3), we have

$$\begin{aligned} -\langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle &= -\frac{1}{\tau} \langle \mathbf{y}_t, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle \\ &= -\frac{1}{2\tau} (\|\mathbf{y}_{t+1}\|_2^2 - \|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2) \\ &= \frac{1}{2\tau} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2. \end{aligned} \quad (33)$$

Plugging into (14) and ignoring some negative terms yield

$$\begin{aligned} &f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ &\leq \frac{1}{2\tau} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \rho(B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_t) - B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_{t+1})) + \rho_x(B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_{t+1})) \\ &\quad + \rho_z(B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_t) - B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_{t+1})) - \rho B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) + \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2. \end{aligned} \quad (34)$$

Assume  $B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) \geq \frac{\alpha}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2$ . If  $0 < p \leq 2$ , using  $\|\mathbf{u}\|_p \leq \|\mathbf{u}\|_2$ ,

$$-\rho B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) + \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 \leq -\frac{\alpha\rho - \tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2.$$

Setting  $\tau \leq (\alpha - 2\gamma)\rho$ , the last two terms on the right hand side of (34) can be removed.

If  $p > 2$ ,  $\|\mathbf{u}\|_2 \leq m^{\frac{1}{2} - \frac{1}{p}} \|\mathbf{u}\|_p$  for any  $\mathbf{u} \in \mathbb{R}^{m \times 1}$ , so  $\|\mathbf{u}\|_p^2 \geq m^{\frac{2}{p}-1} \|\mathbf{u}\|_2^2$ . Then

$$-\rho B_\phi(\mathbf{Bz}_{t+1}, \mathbf{c} - \mathbf{Ax}_{t+1}) + \frac{\tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 \leq -\frac{\alpha\rho m^{\frac{2}{p}-1} - \tau}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2.$$

Setting  $\tau \leq (\alpha m^{\frac{2}{p}-1} - 2\gamma)\rho$ , the last two terms on the right hand side of (34) can be removed. Summing over  $t$  from 0 to  $T-1$ , we have the following telescoping sum

$$\begin{aligned} &\sum_{t=0}^{T-1} [f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*))] \\ &\leq \frac{1}{2\tau} \|\mathbf{y}_0\|_2^2 + \rho B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_0) + \rho_x B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_0) + \rho_z (B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_0)) \\ &= \rho B_\phi(\mathbf{Bz}^*, \mathbf{Bz}_0) + \rho_x B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_0) + \rho_z (B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_0)). \end{aligned} \quad (35)$$

Dividing both sides by  $T$  and applying the Jensen's inequality gives (31).

Dividing both sides of (30) by  $T$  and applying the Jensen's inequality yield (32).  $\blacksquare$

## 2 Convergence of BADMM with Time Varying Step Size

Under the assumption that  $\mathbf{y}_t$  is bounded, the following theorem requires a large step size to establish the convergence of BADMM.

**Theorem 3** *Let the sequences  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1)-(3),  $\mathbf{x}^* \in \mathcal{X}, \mathbf{z}^* \in \mathcal{Z}$  and  $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$  satisfying (5)-(7). Let the Assumption 1 hold and  $\|\mathbf{y}_t\|_2 \leq D_y$ . Setting  $\rho_x = \rho_z = c_1\sqrt{T}, \tau = c_2\sqrt{T}$  and  $\rho = \sqrt{T}$  for some positive constant  $c_1, c_2$ , then  $R(t+1)$  converges to zero.*

*Proof:* Assuming  $\|\mathbf{y}_t\|_2 \leq D_y$  and using (3), we have

$$\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 = \frac{1}{\tau^2} \|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 \leq \frac{2}{\tau^2} (\|\mathbf{y}_{t+1}\|_2^2 + \|\mathbf{y}_t\|_2^2) \leq \frac{4D_y^2}{\tau^2}. \quad (36)$$

Plugging into (29) and rearranging the terms yields

$$R(t+1) \leq D(\mathbf{w}^*, \mathbf{w}_t) - D(\mathbf{w}^*, \mathbf{w}_{t+1}) + \left(\frac{\tau}{2\rho} + \gamma\right) \frac{4D_y^2}{\tau^2}. \quad (37)$$

Setting  $\rho_x = \rho_z = c_1\sqrt{T}, \tau = c_2\sqrt{T}$  and  $\rho = \sqrt{T}$  for some positive constant  $c_1, c_2$ , we have

$$R(t+1) = c_1 B_{\varphi_x}(\mathbf{x}_{t+1}, \mathbf{x}_t) + c_1 B_{\varphi_z}(\mathbf{z}_{t+1}, \mathbf{z}_t) + B_\phi(\mathbf{c} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t) + \gamma \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2, \quad (38)$$

Summing (37) over  $t$  from 0 to  $T-1$ , we have the following telescoping sum

$$\sum_{t=0}^{T-1} R(t+1) \leq D(\mathbf{w}^*, \mathbf{w}_0) + \sum_{t=0}^{T-1} \left(\frac{\tau}{2\rho} + \gamma\right) \frac{4D_y^2}{\tau^2} = D(\mathbf{w}^*, \mathbf{w}_0) + \frac{4(c_2/2 + \gamma)D_y^2}{c_2^2}. \quad (39)$$

Therefore,  $R(t+1) \rightarrow 0$  as  $t \rightarrow \infty$ .  $\blacksquare$

The following theorem establishes the convergence rate for the objective and residual of constraints in an ergodic sense.

**Theorem 4** *Let the sequences  $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$  be generated by Bregman ADMM (1)-(3). Let  $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t, \bar{\mathbf{z}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t$ . Let the Assumption 1 hold and  $\|\mathbf{y}_t\|_2 \leq D_y$ . Set  $\rho_x = \rho_z = c_1\sqrt{T}, \tau = c_2\sqrt{T}, \rho = \sqrt{T}$  for some positive constants  $c_1, c_2$ . For any  $\mathbf{x}^* \in \mathcal{X}, \mathbf{z}^* \in \mathcal{Z}$  and  $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*)$  satisfying KKT conditions (5)-(7), we have*

$$f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \leq \frac{2D_y^2}{c_2\sqrt{T}} + \frac{\|\mathbf{y}_0\|_2^2}{2c_2T\sqrt{T}} + \frac{D_2}{\sqrt{T}}, \quad (40)$$

$$\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T - \mathbf{c}\|_2^2 \leq \frac{D(\mathbf{w}^*, \mathbf{w}_0)}{\gamma T} + \frac{4(c_2/2 + \gamma)D_y^2}{\gamma c_2^2 T}, \quad (41)$$

where  $D_2 = B_\phi(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_0) + c_1(B_{\varphi_x}(\mathbf{x}^*, \mathbf{x}_0) + B_{\varphi_z}(\mathbf{z}^*, \mathbf{z}_0))$ .

*Proof:* Assuming  $\|\mathbf{y}_t\|_2 \leq D_{\mathbf{y}}^2$  and using (3), we have

$$-\langle \mathbf{y}_t, \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} \rangle = -\frac{1}{\tau} \langle \mathbf{y}_t, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle \leq \frac{1}{\tau} (\|\mathbf{y}_t\|_2^2 + \|\mathbf{y}_t\|_2 * \|\mathbf{y}_{t+1}\|_2) \leq \frac{2D_{\mathbf{y}}^2}{\tau}. \quad (42)$$

Plugging into (14) and ignoring some negative terms yield

$$\begin{aligned} & f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{2D_{\mathbf{y}}^2}{\tau} + \rho(B_{\phi}(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_t) - B_{\phi}(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_{t+1})) + \rho_{\mathbf{x}}(B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_t) - B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_{t+1})) \\ & + \rho_{\mathbf{z}}(B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_t) - B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_{t+1})). \end{aligned} \quad (43)$$

Summing over  $t$  from 0 to  $T - 1$ , we have the following telescoping sum

$$\begin{aligned} & \sum_{t=0}^{T-1} [f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*))] \\ & \leq \sum_{t=0}^{T-1} \frac{2D_{\mathbf{y}}^2}{\tau} + \frac{1}{2\tau} \|\mathbf{y}_0\|_2^2 + \rho B_{\phi}(\mathbf{B}\mathbf{z}^*, \mathbf{B}\mathbf{z}_0) + \rho_{\mathbf{x}} B_{\varphi_{\mathbf{x}}}(\mathbf{x}^*, \mathbf{x}_0) + \rho_{\mathbf{z}} B_{\varphi_{\mathbf{z}}}(\mathbf{z}^*, \mathbf{z}_0). \end{aligned}$$

Setting  $\rho_{\mathbf{x}} = \rho_{\mathbf{z}} = c_1\sqrt{T}$ ,  $\tau = c_2\sqrt{T}$ ,  $\rho = \sqrt{T}$ , dividing both sides by  $T$  and applying the Jensen's inequality yield (40).

Dividing both sides of (39) by  $T$  and applying the Jensen's inequality yield (41). ■

## References

- [1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.
- [2] S. Boyd, E. Chu N. Parikh, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends Machine Learning*, 3(1):1–122, 2011.