The online supplement for the manuscript titled "A first-order optimization algorithm for statistical learning with hierarchical sparsity structure" by Zhang, Liu, and Davanloo Tajbakhsh

Appendix A: Proof of Lemma 1

The framework of the proof was first proposed in Luo and Tseng (1993) and also applied in Hong and Luo (2017) and requires "locally upper Lipschitzian" property of *polyhedral* multifunction for the map induced by KKT conditions– see also Stephen M. (1981), Walkup and Wets (1969), Stephen M. (1973), Mangasarian and Shiau (1987), Hoffman (1952). However, due to the presence of the conic constraints in (28), the resulting multifunction is *not polyhedral* anymore. Indeed, Walkup and Wets (1969) showed that having a polyhedral graph is a necessary condition for the upper Lipschitzian property of the multifunction. The following proof uses the specific structure of this problem to establish the dual error bound condition.

For any $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{y}^* \in Y^*$, considering the KKT conditions (29), we have

$$\begin{aligned}\|\mathbf{y}-\mathbf{y}^*\|_2^2 &= \|M^\top M(\mathbf{x}^2(\mathbf{y})-\mathbf{x}^2(\mathbf{y}^*))+\rho(\mathbf{x}^2(\mathbf{y})-\mathbf{x}^1(\mathbf{y}))-\rho(\mathbf{x}^2(\mathbf{y}^*)-\mathbf{x}^1(\mathbf{y}^*))\|_2^2\\ &= \|\nabla_{\mathbf{x}^2}\phi(M\mathbf{x}^2(\mathbf{y}))-\nabla_{\mathbf{x}^2}\phi(M\mathbf{x}^2(\mathbf{y}^*))+\nabla_{\mathbf{x}^2}\psi(E\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}^2}\psi(E\mathbf{x}(\mathbf{y}^*))\|_2^2\\ &\le \|\nabla_{\mathbf{x}^2}\phi(M\mathbf{x}^2(\mathbf{y}))-\nabla_{\mathbf{x}^2}\phi(M\mathbf{x}^2(\mathbf{y}^*))\|_2^2+\|\nabla_{\mathbf{x}}\psi(E\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}}\psi(E\mathbf{x}(\mathbf{y}^*))\|_2^2,\end{aligned}$$

where the first equality follows from (29b), the second equality follows from the definition of $\phi(\cdot)$ and $\psi(\cdot)$ (defined below (25)), the third inequality follows from the triangle inequality. Hence, using (26) and (27), we have

$$\|\mathbf{y}-\mathbf{y}^*\|_2^2 \le L_\phi^2\|M\mathbf{x}^2(\mathbf{y})-M\mathbf{x}^2(\mathbf{y}^*)\|_2^2+L_\psi^2\|E\mathbf{x}(\mathbf{y})-E\mathbf{x}(\mathbf{y}^*)\|_2^2. \tag{34}$$

Next, consider

$$\begin{aligned}&\|M\mathbf{x}^2(\mathbf{y})-M\mathbf{x}^2(\mathbf{y}^*)\|_2^2+\rho\|E\mathbf{x}(\mathbf{y})-E\mathbf{x}(\mathbf{y}^*)\|_2^2\\ &=\left\langle M^\top M\mathbf{x}^2(\mathbf{y})-M^\top M\mathbf{x}^2(\mathbf{y}^*),\mathbf{x}^2(\mathbf{y})-\mathbf{x}^2(\mathbf{y}^*)\right\rangle+\rho\left\langle E^\top E\mathbf{x}(\mathbf{y})-M^\top M\mathbf{x}(\mathbf{y}^*),\mathbf{x}(\mathbf{y})-\mathbf{x}(\mathbf{y}^*)\right\rangle\\ &=\left\langle\nabla_{\mathbf{x}}\phi(M\mathbf{x}^2(\mathbf{y}))-\nabla_{\mathbf{x}}\phi(M\mathbf{x}^2(\mathbf{y}^*)),\mathbf{x}(\mathbf{y})-\mathbf{x}(\mathbf{y}^*)\right\rangle+\left\langle\nabla_{\mathbf{x}}\psi(E\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}}\psi(E\mathbf{x}(\mathbf{y}^*)),\mathbf{x}(\mathbf{y})-\mathbf{x}(\mathbf{y}^*)\right\rangle\\ &=\left\langle\nabla_{\mathbf{x}}\ell(\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}}\ell(\mathbf{x}(\mathbf{y}^*)),\mathbf{x}(\mathbf{y})-\mathbf{x}(\mathbf{y}^*)\right\rangle\\ &=\left\langle\nabla_{\mathbf{x}^1}\ell(\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}^1}\ell(\mathbf{x}(\mathbf{y}^*)),\mathbf{x}^1(\mathbf{y})-\mathbf{x}^1(\mathbf{y}^*)\right\rangle+\left\langle\nabla_{\mathbf{x}^2}\ell(\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}^2}\ell(\mathbf{x}(\mathbf{y}^*)),\mathbf{x}^2(\mathbf{y})-\mathbf{x}^2(\mathbf{y}^*)\right\rangle\\ &=\sum_{g\in\mathcal{G}}\left\langle\nabla_{\mathbf{x}^1_{j(g)}}\ell(\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}^1_{j(g)}}\ell(\mathbf{x}(\mathbf{y}^*)),\mathbf{x}^1_{j(g)}(\mathbf{y})-\mathbf{x}^1_{j(g)}(\mathbf{y}^*)\right\rangle\\ &\quad+\sum_{g\in\mathcal{G}}\left\langle\nabla_{\mathbf{x}^2_{j(g)}}\ell(\mathbf{x}(\mathbf{y}))-\nabla_{\mathbf{x}^2_{j(g)}}\ell(\mathbf{x}(\mathbf{y}^*)),\mathbf{x}^2_{j(g)}(\mathbf{y})-\mathbf{x}^2_{j(g)}(\mathbf{y}^*)\right\rangle\\ &=\sum_{g\in\mathcal{G}}\left\langle w_g\boldsymbol{\mu}_g(\mathbf{y})-\mathbf{y}_{j(g)}-w_g\boldsymbol{\mu}_g(\mathbf{y}^*)+\mathbf{y}^*_{j(g)},\mathbf{x}^1_{j(g)}(\mathbf{y})-\mathbf{x}^1_{j(g)}(\mathbf{y}^*)\right\rangle\\ &\quad+\sum_{g\in\mathcal{G}}\left\langle\mathbf{y}_{j(g)}-\mathbf{y}^*_{j(g)},\mathbf{x}^2_{j(g)}(\mathbf{y})-\mathbf{x}^2_{j(g)}(\mathbf{y}^*)\right\rangle\end{aligned}$$

where the second and third equalities follow from the definitions of ϕ, ψ, and ℓ, in the forth and fifth equalities the gradient is expanded over each $\mathbf{x}^1$ and $\mathbf{x}^2$, and the last equality follows from (29a),(29b). Rearranging the terms in the last line above and using $\mathbf{x}^1_{j(g)}(\mathbf{y}^*) = \mathbf{x}^2_{j(g)}(\mathbf{y}^*)\ \forall g \in \mathcal{G}$, we get

$$\begin{aligned}&\|M\mathbf{x}^2(\mathbf{y}) - M\mathbf{x}^2(\mathbf{y}^*)\|_2^2 + \rho\|E\mathbf{x}(\mathbf{y}) - E\mathbf{x}(\mathbf{y}^*)\|_2^2 \\ &\quad= \sum_{g\in\mathcal{G}} \left\langle w_g\boldsymbol{\mu}_g(\mathbf{y}) - w_g\boldsymbol{\mu}_g(\mathbf{y}^*), \mathbf{x}^1_{j(g)}(\mathbf{y}) - \mathbf{x}^1_{j(g)}(\mathbf{y}^*)\right\rangle + \sum_{g\in\mathcal{G}} \left\langle \mathbf{y}_{j(g)} - \mathbf{y}^*_{j(g)}, \mathbf{x}^2_{j(g)}(\mathbf{y}) - \mathbf{x}^1_{j(g)}(\mathbf{y})\right\rangle,\end{aligned} \tag{35}$$

For all $g \in \mathcal{G}$, we have

$$\begin{aligned}&\left\langle w_g\boldsymbol{\mu}_g(\mathbf{y}) - w_g\boldsymbol{\mu}_g(\mathbf{y}^*), \mathbf{x}^1_{j(g)}(\mathbf{y}) - \mathbf{x}^1_{j(g)}(\mathbf{y}^*)\right\rangle \\ &\quad= w_g\boldsymbol{\mu}_g(\mathbf{y})^\top\mathbf{x}^1_{j(g)}(\mathbf{y}) - w_g\boldsymbol{\mu}_g(\mathbf{y}^*)^\top\mathbf{x}^1_{j(g)}(\mathbf{y}) - w_g\boldsymbol{\mu}_g(\mathbf{y})^\top\mathbf{x}^1_{j(g)}(\mathbf{y}^*) + w_g\boldsymbol{\mu}_g(\mathbf{y}^*)^\top\mathbf{x}^1_{j(g)}(\mathbf{y}^*) \\ &\quad= -w_g(s_g/w_g)\lambda - w_g\boldsymbol{\mu}_g(\mathbf{y}^*)^\top\mathbf{x}^1_{j(g)}(\mathbf{y}) - w_g\boldsymbol{\mu}_g(\mathbf{y})^\top\mathbf{x}^1_{j(g)}(\mathbf{y}^*) - w_g(s_g/w_g)\lambda \\ &\quad\le w_g\|\boldsymbol{\mu}_g(\mathbf{y}^*)\|_2\|\mathbf{x}^1_{j(g)}(\mathbf{y})\|_2 + w_g\|\boldsymbol{\mu}_g(\mathbf{y})\|_2\|\mathbf{x}^1_{j(g)}(\mathbf{y}^*)\|_2 - 2s_g\lambda \\ &\quad\le 0,\end{aligned}$$

where the second equality follows from (29c) and (29f), the third inequality uses Cauchy-Schwarz inequality, and the last inequality follows from (29c), (29d), and (29e). Hence, we have

$$\|M\mathbf{x}^2(\mathbf{y}) - M\mathbf{x}^2(\mathbf{y}^*)\|_2^2 + \rho\|E\mathbf{x}(\mathbf{y}) - E\mathbf{x}(\mathbf{y}^*)\|_2^2 \le \sum_{g\in\mathcal{G}} \left\langle \mathbf{y}_{j(g)} - \mathbf{y}^*_{j(g)}, \mathbf{x}^2_{j(g)}(\mathbf{y}) - \mathbf{x}^1_{j(g)}(\mathbf{y})\right\rangle \tag{36}$$

$$\le \|\mathbf{y} - \mathbf{y}^*\|\|\nabla g_\rho(\mathbf{y})\|, \tag{37}$$

where (36) uses nonpositivity of the first term in (35) (shown above), and (37) follows from Cauchy-Schwarz inequality and the fact that $\nabla g_\rho(\mathbf{y}) = \mathbf{x}^1(\mathbf{y}) - \mathbf{x}^2(\mathbf{y})$ – see e.g. Hong and Luo (2017) - Lemma 2.1. Finally, using (37) and (34), we get

$$\|\mathbf{y} - \mathbf{y}^*\|_2^2 \le \max\{L_\phi^2, L_\psi^2/\rho\}\Big(\|M\mathbf{x}^2(\mathbf{y}) - M\mathbf{x}^2(\mathbf{y}^*)\|_2^2 + \rho\|E\mathbf{x}(\mathbf{y}) - E\mathbf{x}(\mathbf{y}^*)\|_2^2\Big) \tag{38}$$

$$\le \max\{L_\phi^2, L_\psi^2/\rho\}\|\mathbf{y} - \mathbf{y}^*\|_2\|\nabla g_\rho(\mathbf{y})\|_2. \tag{39}$$

Hence, we have

$$\operatorname{dist}(\mathbf{y}, Y^*) \le \|\mathbf{y} - \mathbf{y}^*\| \le \max\{L_\phi^2, L_\psi^2/\rho\}\|\nabla g_\rho(\mathbf{y})\|_2, \tag{40}$$

and the existence of τ_d follows.

Appendix B: Proof of Lemma 2

Before proceeding with the proof of the boundedness of the iterates, we show the existence of a finite saddle point by the following argument. Consider

$$\min_{\mathbf{x}^1,\mathbf{x}^2\in\mathbb{R}^n} \left\{ \tilde{F}(\mathbf{x}^1,\mathbf{x}^2) = \lambda\sum_{g\in\mathcal{G}} w_g\|\mathbf{x}^1_{j(g)}\|_2 + \frac{1}{2}\|M\mathbf{x}^2 - \mathbf{b}\|_2^2 + \frac{\rho}{2}\|\mathbf{x}^1 - \mathbf{x}^2\|_2^2, \ \text{ s.t. } \ \mathbf{x}^1 = \mathbf{x}^2 \right\}. \tag{41}$$

The above problem is equivalent to (8) whose objective function is coercive and continuous. By Weierstrass's Theorem (see e.g. Bertsekas (1999)), we have certain finite optimal solution to (41) $(\mathbf{x}^{1,*}, \mathbf{x}^{2,*})$, i.e. $\tilde{F}^* = \inf_{\mathbf{x}^1=\mathbf{x}^2\in R^n} \tilde{F}(\mathbf{x}^1,\mathbf{x}^2) = \tilde{F}(\mathbf{x}^{1,*},\mathbf{x}^{2,*})$. Especially, $\mathbf{x}^{1,*} = \mathbf{x}^{2,*}$. Consider $L_\rho(\mathbf{x}^1,\mathbf{x}^2;\mathbf{y}) = \tilde{F}(\mathbf{x}^1,\mathbf{x}^2) + \langle \mathbf{y}, \mathbf{x}^1 - \mathbf{x}^2\rangle$ and the corresponding dual function $g_\rho(\mathbf{y})$. First, we have

$L_\rho(\mathbf{x}^{1,*},\mathbf{x}^{2,*};\mathbf{y}) = \tilde{F}(\mathbf{x}^{1,*},\mathbf{x}^{2,*})$, for $\forall \mathbf{y}$. By strong duality (see e.g. Prop. 5.2.1 in Bertsekas (1999)), we know there is no duality gap. Furthermore, there exists at least one Lagrange multiplier $\mathbf{y}^*$, i.e. $\tilde{F}^* = \inf_{\mathbf{x}^1,\mathbf{x}^2\in\mathbb{R}} L_\rho(\mathbf{x}^1,\mathbf{x}^2;\mathbf{y}^*) = g_\rho(\mathbf{y}^*)$. We conclude $(\mathbf{x}^{1,*},\mathbf{x}^{2,*};\mathbf{y})$ is a finite saddle point.

The idea for the proof of boundedness of the iterates is similar to Theorem 5.1 in Glowinski (1984); however, we do not have the strong convexity assumption. Given a finite saddle point $((\mathbf{x}^{1,*},\mathbf{x}^{2,*});\mathbf{y}^*)$, define $\tilde{\mathbf{x}}^{1,k} \triangleq \mathbf{x}^{1,k} - \mathbf{x}^{1,*}$, $\tilde{\mathbf{x}}^{2,k} \triangleq \mathbf{x}^{2,k} - \mathbf{x}^{2,*}$, $\tilde{\mathbf{y}}^k \triangleq \mathbf{y}^k - \mathbf{y}^*$ where $\mathbf{x}^{1,*} = \mathbf{x}^{2,*}$. Establishing the boundedness of the sequence is equivalent to showing that the sequence $\{\|\tilde{\mathbf{y}}^k\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\|_2^2\}_{k=1}^\infty$ is non-increasing. From the convexity of the augmented Lagrangian function (10) in $\mathbf{x}^1$, we have

$$\left\langle \mathbf{y}^* + \rho(\mathbf{x}^{1,*} - \mathbf{x}^{2,*}), \mathbf{x}^1 - \mathbf{x}^{1,*}\right\rangle + \lambda\sum_{g\in\mathcal{G}} w_g\|\mathbf{x}^1_{j(g)}\|_2 - \lambda\sum_{g\in\mathcal{G}} w_g\|\mathbf{x}^{1,*}_{j(g)}\|_2 \geq 0, \forall \mathbf{x}^1. \tag{42}$$

Furthermore, from the convexity of the augmented Lagrangian (10) in $\mathbf{x}^2$, we have

$$\left\langle M^\top(M\mathbf{x}^{2,*} - b) - \mathbf{y}^* + \rho(\mathbf{x}^{2,*} - \mathbf{x}^{1,*}), \mathbf{x}^2 - \mathbf{x}^{2,*}\right\rangle \geq 0, \forall \mathbf{x}^2. \tag{43}$$

From the fact that $L_\rho(\mathbf{x}^{1,*},\mathbf{x}^{2,*};\mathbf{y}^*) \leqslant L_\rho(\mathbf{x}^1,\mathbf{x}^2;\mathbf{y}^*), \forall \mathbf{x}^1,\mathbf{x}^2$, we have

$$\mathbf{y}^* = \mathbf{y}^* + \alpha(\mathbf{x}^{1,*} - \mathbf{x}^{2,*}). \tag{44}$$

Similar to the arguments for (42)-(44), from (13)-(15), we have

$$\left\langle \rho(\mathbf{x}^{1,k+1}_{j(g)} - \mathbf{x}^{2,k}_{j(g)}) + \mathbf{y}^k_{j(g)}, \mathbf{x}^1_{j(g)} - \mathbf{x}^{1,k+1}_{j(g)}\right\rangle + \lambda w_g\|\mathbf{x}^1_{j(g)}\|_2 - \lambda w_g\|\mathbf{x}^{1,k+1}_{j(g)}\|_2 \geq 0, \quad \forall g\in\mathcal{G}, \forall \mathbf{x}^1_{j(g)}, \tag{45}$$

$$\left\langle M^\top(M\mathbf{x}^{2,k+1} - b) + \rho(\mathbf{x}^{2,k+1} - \mathbf{x}^{1,k+1}) - \mathbf{y}^k, \mathbf{x}^2 - \mathbf{x}^{2,k+1}\right\rangle \geq 0, \quad \forall \mathbf{x}^2, \tag{46}$$

$$\mathbf{y}^{k+1}_{j(g)} = \mathbf{y}^k_{j(g)} + \alpha(\mathbf{x}^{1,k+1}_{j(g)} - \mathbf{x}^{2,k+1}_{j(g)}), \quad \forall g\in\mathcal{G}. \tag{47}$$

Since $j(g)\cap j(\bar{g}) = \emptyset$ for all $g,\bar{g}\in\mathcal{G}$ such that $g\neq\bar{g}$, from (45) and (47), we have:

$$\left\langle \rho(\mathbf{x}^{1,k+1} - \mathbf{x}^{2,k}) + \mathbf{y}^k, \mathbf{x}^1 - \mathbf{x}^{1,k+1}\right\rangle + \lambda\sum_{g\in\mathcal{G}} w_g\|\mathbf{x}^1_{j(g)}\|_2 - \lambda\sum_{g\in\mathcal{G}} w_g\|\mathbf{x}^{1,k+1}_{j(g)}\|_2 \geq 0, \quad \forall \mathbf{x}^1, \tag{48}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \alpha(\mathbf{x}^{1,k+1} - \mathbf{x}^{2,k+1}). \tag{49}$$

Setting $\mathbf{x}^1 = \mathbf{x}^{1,k+1}$ in (42), and $\mathbf{x}^1 = \mathbf{x}^{1,*}$ in (48) and adding them, we get

$$\left\langle -\tilde{\mathbf{y}}^k + \rho(\tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{1,k+1}), \tilde{\mathbf{x}}^{1,k+1}\right\rangle \geq 0 \tag{50}$$

Similarly, setting $\mathbf{x}^2 = \mathbf{x}^{2,k+1}$ in (43), and $\mathbf{x}^2 = \mathbf{x}^{2,*}$ in (46) and adding them, we get

$$\left\langle M^\top M\tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{y}}^k - \rho(\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}), -\tilde{\mathbf{x}}^{2,k+1}\right\rangle \geq 0 \tag{51}$$

Adding the left-hand-sides of (50) to (51) and rearranging the terms, we have,

$$\begin{aligned}
&\left\langle -\tilde{\mathbf{y}}^k, \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\right\rangle + \rho\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{1,k+1}, \tilde{\mathbf{x}}^{1,k+1}\right\rangle + \rho\left\langle \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{2,k+1}\right\rangle - \|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2\\
&= \left\langle -\tilde{\mathbf{y}}^k, \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\right\rangle + \rho\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{1,k+1} + \tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\right\rangle\\
&+ \rho\left\langle \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}), \tilde{\mathbf{x}}^{2,k+1}\right\rangle - \|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2
\end{aligned}$$

Hence, we have

$$\rho\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\right\rangle - \|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2 - \rho\|\tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{x}}^{1,k+1}\|_2^2 \geq \left\langle \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{y}}^k\right\rangle. \tag{52}$$

From (49), we have the following two inequalities:

$$\tilde{\mathbf{y}}^{k+1} - \tilde{\mathbf{y}}^{k} = \alpha(\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1})$$
$$\tilde{\mathbf{y}}^{k+1} + \tilde{\mathbf{y}}^{k} = \alpha(\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}) + 2\tilde{\mathbf{y}}^{k}$$

Taking the inner product of the left terms together and the right terms together, we obtain

$$\|\tilde{\mathbf{y}}^{k+1}\|_2^2 - \|\tilde{\mathbf{y}}^{k}\|_2^2 = \alpha^2\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\| + 2\alpha\left\langle \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{y}}^{k}\right\rangle \tag{53}$$
$$\leq \alpha(\alpha - 2\rho)\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 - 2\alpha\|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2 + 2\alpha\rho\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\right\rangle \tag{54}$$

where the inequality uses (52). Next, we will upper bound $\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\rangle$. Setting $\mathbf{x}^2 = \mathbf{x}^{2,k}$ in (46), we have

$$\left\langle M^\top(M\mathbf{x}^{2,k+1} - b) + \rho(\mathbf{x}^{2,k+1} - \mathbf{x}^{1,k+1}) - \mathbf{y}^{k}, \mathbf{x}^{2,k} - \mathbf{x}^{2,k+1}\right\rangle \geq 0. \tag{55}$$

Setting $k+1$ in (46) to k, and $\mathbf{x}^2 = \mathbf{x}^{2,k+1}$, we have

$$\left\langle M^\top(M\mathbf{x}^{2,k} - b) + \rho(\mathbf{x}^{2,k} - \mathbf{x}^{1,k}) - \mathbf{y}^{k-1}, \mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\right\rangle \geq 0. \tag{56}$$

Adding (55) and (56), we have

$$\begin{aligned}&\left\langle \mathbf{y}^{k} - \mathbf{y}^{k-1}, \mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\right\rangle - \rho\|\mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\|_2^2 + \rho\left\langle \mathbf{x}^{1,k+1} - \mathbf{x}^{1,k}, \mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\right\rangle \\ &\quad \geq \|M(\mathbf{x}^{2,k+1} - \mathbf{x}^{2,k})\|_2^2 \geq 0.\end{aligned} \tag{57}$$

From (49), we have $\mathbf{y}^k - \mathbf{y}^{k-1} = \alpha(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})$. Using it in (57) and rearranging terms, we obtain

$$\rho\left\langle \mathbf{x}^{1,k+1} - \mathbf{x}^{1,k}, \mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\right\rangle \geq \rho\|\mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\|_2^2 - \alpha\left\langle \mathbf{x}^{1,k} - \mathbf{x}^{2,k}, \mathbf{x}^{2,k+1} - \mathbf{x}^{2,k}\right\rangle.$$

Adding and subtracting $\mathbf{x}^{1,*}$ and $\mathbf{x}^{2,*}$ into each argument in (B) as needed, we have

$$\rho\left\langle \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{1,k}, \tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{x}}^{2,k}\right\rangle \geq \rho\|\tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{x}}^{2,k}\|_2^2 - \alpha\left\langle \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}, \tilde{\mathbf{x}}^{2,k+1} - \tilde{\mathbf{x}}^{2,k}\right\rangle. \tag{58}$$

The term $\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\rangle$ can be transformed as following:

$$\begin{aligned}&\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1}\right\rangle \\ &= \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{1,k} + \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k} + \tilde{\mathbf{x}}^{2,k}\right\rangle \\ &= \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{1,k}\right\rangle + \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\right\rangle + \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{2,k}\right\rangle \\ &= \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{1,k}\right\rangle + \left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\right\rangle + \frac{1}{2}(\|\tilde{\mathbf{x}}^{2,k}\|_2^2 - \|\tilde{\mathbf{x}}^{2,k+1}\|_2^2 + \|\tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2) \\ &\leq \frac{1}{2}(\|\tilde{\mathbf{x}}^{2,k}\|_2^2 - \|\tilde{\mathbf{x}}^{2,k+1}\|_2^2 - \|\tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2) + (1 - \frac{\alpha}{\rho})\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\right\rangle,\end{aligned} \tag{59}$$

where the last inequality follows from (58). Combining (53) and (59) and rearranging the terms, we obtain

$$\begin{aligned}&\|\tilde{\mathbf{y}}^{k+1}\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k+1}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 - (\|\tilde{\mathbf{y}}^{k}\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k}\|_2^2) \\ &\leq -\alpha\rho\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 - 2\alpha\|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2 - \alpha\rho\|\tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 + 2\alpha(\rho-\alpha)\left\langle \tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}, \tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\right\rangle\end{aligned} \tag{60}$$

By upper bounding the last term in (60) by the identity $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2$, we get

$$\begin{aligned}&\|\tilde{\mathbf{y}}^{k+1}\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k+1}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 - (\|\tilde{\mathbf{y}}^{k}\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\|_2^2) \\ &\leq -\alpha\rho\|\tilde{\mathbf{x}}^{1,k+1} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 - 2\alpha\|M\tilde{\mathbf{x}}^{2,k+1}\|_2^2 - \alpha^2\|\tilde{\mathbf{x}}^{2,k} - \tilde{\mathbf{x}}^{2,k+1}\|_2^2 \leq 0.\end{aligned} \tag{61}$$

We have shown that the sequence $\{\|\tilde{\mathbf{y}}^k\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\|_2^2\}_{k=1}^\infty$ is non-increasing. Once the initial point and saddle point are fixed, which are not related to α, then $\{\|\tilde{\mathbf{y}}^k\|_2^2 + \alpha\rho\|\tilde{\mathbf{x}}^{2,k}\|_2^2 + \alpha(\rho-\alpha)\|\tilde{\mathbf{x}}^{1,k} - \tilde{\mathbf{x}}^{2,k}\|_2^2\}_{k=1}^\infty$ is bounded by $\|\tilde{\mathbf{y}}^0\|_2^2 + \rho^2\|\tilde{\mathbf{x}}^{2,0}\|_2^2 + \frac{\rho^2}{4}\|\tilde{\mathbf{x}}^{1,0} - \tilde{\mathbf{x}}^{2,0}\|_2^2$. We concluded that the sequence $\{\mathbf{x}^{1,k}\}$, $\{\mathbf{x}^{2,k}\}$ and $\{\mathbf{y}^k\}$ generated by (15) is uniformly bounded for any $0 < \alpha < \rho$.

Appendix C: Proof of Lemma 3

The proof extends the analysis of Tseng (2010) and Zhang et al. (2013). Since both works discuss primal methods, there are mainly two new ingredients in our proof: 1) dealing with the dual variable $\mathbf{y}$, and 2) splitting $\mathbf{x}$ into $\mathbf{x}^1$ and $\mathbf{x}^2$, where neither step is trivial.

Given $\mathbf{y}$, note that $\mathbf{X}(\mathbf{y})$ can be written as $(\mathbf{X}^1(\mathbf{y}), \mathbf{X}^2(\mathbf{y}))$. For a fixed $\mathbf{y}$, and for *any* sequence $\{(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y}) : \mathbf{x}^{2,k} \notin \mathbf{X}^2(\mathbf{y})\}_{k\geq 0}$, we define

$$\mathbf{r}^{1,k} \triangleq \tilde{\nabla}_{\mathbf{x}^1} L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y}), \tag{62}$$

$$\mathbf{r}^{2,k} \triangleq \tilde{\nabla}_{\mathbf{x}^2} L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y}) = M^T(M\mathbf{x}^{2,k} - \mathbf{b}) - \mathbf{y} + \rho(\mathbf{x}^{2,k} - \mathbf{x}^{1,k}), \tag{63}$$

$$\delta^k \triangleq \|\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}\|_2, \quad \text{where } \bar{\mathbf{x}}^{2,k} \triangleq \operatorname*{arg\,min}_{\mathbf{x}^2 \in \mathbf{X}^2(\mathbf{y})} \|\mathbf{x}^{2,k} - \mathbf{x}^2\|_2, \tag{64}$$

$$\bar{\mathbf{x}}^{1,k} \triangleq \operatorname*{arg\,min}_{\mathbf{x}^1} L_\rho(\mathbf{x}^1, \bar{\mathbf{x}}^{2,k}; \mathbf{y}), \tag{65}$$

$$\mathbf{u}^k \triangleq \frac{\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}}{\delta^k}. \tag{66}$$

Note that

$$\tilde{\nabla}_{\mathbf{x}^1} L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y}) = \mathbf{x}^1 - \operatorname{prox}_{\lambda \sum_{g\in\mathcal{G}} w_g \|\tilde{\mathbf{d}}_{j(g)}\|_2}(\mathbf{x}^1 - \mathbf{y} - \rho(\mathbf{x}^1 - \mathbf{x}^2)) \tag{67}$$

$$= \mathbf{x}^1 - \operatorname*{arg\,min}_{\tilde{\mathbf{d}}} \ \lambda \sum_{g\in\mathcal{G}} w_g \|\tilde{\mathbf{d}}_{j(g)}\|_2 + \frac{1}{2}\|\tilde{\mathbf{d}} - (\mathbf{x}^1 - \mathbf{y} + \rho(\mathbf{x}^1 - \mathbf{x}^2))\|_2^2 \tag{68}$$

$$= \operatorname*{arg\,min}_{\mathbf{d}} \lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{d}_{j(g)} - \mathbf{x}^1_{j(g)}\|_2 + \frac{1}{2}\|\mathbf{d} - \mathbf{y} - \rho(\mathbf{x}^1 - \mathbf{x}^2)\|_2^2, \tag{69}$$

where the second equality follows from the definition of the proximal operator and the third equality uses the transformation $\mathbf{d} \triangleq \mathbf{x}^1 - \tilde{\mathbf{d}}$. Furthermore, for any group $g \in \mathcal{G}$, we have

$$\left(\tilde{\nabla}_{\mathbf{x}^1} L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})\right)_{j(g)} = \operatorname*{arg\,min}_{\mathbf{d}_{j(g)}} \lambda w_g \|\mathbf{d}_{j(g)} - \mathbf{x}^1_{j(g)}\|_2 + \frac{1}{2}\|\mathbf{d}_{j(g)} - \mathbf{y}_{j(g)} - \rho(\mathbf{x}^1_{j(g)} - \mathbf{x}^2_{j(g)})\|_2^2 \tag{70}$$

$$= \begin{cases} \mathbf{x}^1_{j(g)}, & \text{if } \ \|\mathbf{x}^1_{j(g)} - \mathbf{y}_{j(g)} - \rho(\mathbf{x}^1_{j(g)} - \mathbf{x}^2_{j(g)})\|_2 \leq \lambda w_g, \\ \gamma_g \mathbf{x}^1_{j(g)} + (1-\gamma_g)(\mathbf{y}_{j(g)} + \rho(\mathbf{x}^1_{j(g)} - \mathbf{x}^2_{j(g)})), & \text{otherwise.} \end{cases} \tag{71}$$

where $\gamma_g = \lambda w_g / \|\mathbf{x}^1_j(g) - \mathbf{y}_{j(g)} - \rho(\mathbf{x}^1_{j(g)} - \mathbf{x}^2_{j(g)})\|_2$. Note that the two cases from the soft-thresholding operator in (71) yield $\mathbf{x}^1_{j(g)}$ at the boundary $\|\mathbf{x}^1_{j(g)} - \mathbf{y}_{j(g)} - \rho(\mathbf{x}^1_{j(g)} - \mathbf{x}^2_{j(g)})\|_2 = \lambda w_g$, i.e., $\tilde{\nabla}_{x^1} L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})_{j(g)}$ is continuous in $(\mathbf{x}^1, \mathbf{x}^2, \mathbf{y})$.

To prove this lemma, we will first prove that it suffices to show that there exists $0 < \tilde{\tau} < +\infty$ and $\delta > 0$ such that

$$\operatorname{dist}(\mathbf{x}^2, \mathbf{X}^2(\mathbf{y})) \leq \tilde{\tau} \|\tilde{\nabla}_{\mathbf{x}} L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})\|_2, \tag{72}$$

for all $(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})$ such that $\|\tilde{\nabla}_x L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})\|_2 \leq \delta$. Second, we will show (72).

Assume (72) holds. Given $(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})$, pick $(\mathbf{x}^{1,*}, \mathbf{x}^{2,*}) \in \mathbf{X}(\mathbf{y})$, such that $\operatorname{dist}(\mathbf{x}^2, \mathbf{x}^{2,*}) = \operatorname{dist}(\mathbf{x}^2, \mathbf{X}^2(\mathbf{y}))$, and $\mathbf{x}^{1,*}$ such that it satisfies the optimality condition (74). Recall that

$$\tilde{\nabla}_{\mathbf{x}^2} L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y}) = M^T(M\mathbf{x}^2 - \mathbf{b}) - \mathbf{y} + \rho(\mathbf{x}^2 - \mathbf{x}^1). \tag{73}$$

Hence, from the optimality condition, we have

$$\tilde{\nabla}_{\mathbf{x}^2} L_\rho(\mathbf{x}^{1,*}, \mathbf{x}^{2,*}; \mathbf{y}) = M^T(M\mathbf{x}^{2,*} - \mathbf{b}) - \mathbf{y} + \rho(\mathbf{x}^{2,*} - \mathbf{x}^{1,*}) = 0 \tag{74}$$

Subtracting (74) from (73) and rearranging the terms, we obtain

$$\mathbf{x}^1 - \mathbf{x}^{1,*} = (\frac{1}{\rho}M^TM + \mathbf{I})(\mathbf{x}^2 - \mathbf{x}^{2,*}) - \frac{1}{\rho}\tilde{\nabla}_{x^2}L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y}). \tag{75}$$

Thus,

$$\begin{aligned} \mathrm{dist}(\mathbf{x}, \mathbf{X}(\mathbf{y}))^2 &\le \|\mathbf{x}^1 - \mathbf{x}^{1,*}\|_2^2 + \|\mathbf{x}^2 - \mathbf{x}^{2,*}\|_2^2 &(76)\\ &\le \|(\frac{1}{\rho}M^TM + \mathbf{I})(\mathbf{x}^2 - \mathbf{x}^{2,*})\|_2^2 + \|\frac{1}{\rho}\tilde{\nabla}_{x^2}L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})\|_2^2 + \|\mathbf{x}^2 - \mathbf{x}^{2,*}\|_2^2. &(77)\end{aligned}$$

Upper bounding $\|\mathbf{x}^2 - \mathbf{x}^{2,*}\|_2^2$ in (77) with (72), we have (31).

Next, we will show (72) by contradiction. Suppose (72) does not hold, then there exists a sequence $\{(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y}) : \mathbf{x}^{2,k} \notin \mathbf{X}^2(\mathbf{y})\}_{k\ge 0}$ satisfying

$$\|\tilde{\nabla}_{\mathbf{x}}L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}, \mathbf{y})\|_2/\delta^k \to 0, \quad \text{and} \quad \|\tilde{\nabla}_{\mathbf{x}}L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}, \mathbf{y})\|_2 \to 0. \tag{78}$$

Note that

$$\frac{\|\mathbf{r}^{1,k}\| + \|\mathbf{r}^{2,k}\|}{\sqrt{2}} \le \|\tilde{\nabla}_x L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y})\| \le \|\mathbf{r}^{1,k}\| + \|\mathbf{r}^{2,k}\| \tag{79}$$

where $\mathbf{r}^{1,k}$ and $\mathbf{r}^{2,k}$ are defined in (62) and (63), respectively. Hence, using the left inequality in (79), (78) implies

$$\{\mathbf{r}^{1,k}\} \to \mathbf{0}, \quad \{\mathbf{r}^{2,k}\} \to \mathbf{0}, \quad \{\frac{\|\mathbf{r}^{1,k}\| + \|\mathbf{r}^{2,k}\|}{\delta^k}\} \to 0. \tag{80}$$

We will show that (80) does not hold. Since $(\mathbf{x}^{1,k}, \mathbf{x}^{2,k})$ is in a compact set, by passing to a subsequence if necessary, we can assume that $\{(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}) \to (\bar{\mathbf{x}}^1, \bar{\mathbf{x}}^2)\}$. Since $\{\mathbf{r}^{1,k}\} \to \mathbf{0}$, and $\{\mathbf{r}^{2,k}\} \to \mathbf{0}$, then by the right inequality in (79), $\tilde{\nabla}_x L_\rho(\mathbf{x}^{1,k}, \mathbf{x}^{2,k}; \mathbf{y}) \to \mathbf{0}$. Furthermore, since $\tilde{\nabla}_x L_\rho(\mathbf{x}^1, \mathbf{x}^2; \mathbf{y})$ is continuous, this implies $\tilde{\nabla}_x L_\rho(\bar{x}^1, \bar{x}^2; \mathbf{y}) = \mathbf{0}$. It further implies that $(\bar{\mathbf{x}}^1, \bar{\mathbf{x}}^2) \in \mathbf{X}(\mathbf{y})$. Hence $\delta^k \le \|\mathbf{x}^{2,k} - \bar{\mathbf{x}}^2\| \to 0$, as $k \to \infty$, so that $\{\bar{\mathbf{x}}^{2,k}\} \to \bar{\mathbf{x}}^2$. And based on (75), we have

$$\{\bar{\mathbf{x}}^{1,k}\} \to \bar{\mathbf{x}}^1. \tag{81}$$

Next, we claim there exists $\kappa > 0$ such that,

$$\|\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}\| \le \kappa \|M\mathbf{x}^{2,k} - M\bar{\mathbf{x}}^{2,k}\|, \quad \forall k \tag{82}$$

Again, we argue (82) by contraction. Suppose (82) does not hold, then by passing to a subsequence if necessary, we can assume

$$\{\frac{\|M\mathbf{x}^{2,k} - M\bar{\mathbf{x}}^{2,k}\|}{\|\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}\|}\} \to 0. \tag{83}$$

This implies that $\{M\mathbf{u}^k\} \to 0$, where $\mathbf{u}^k$ is defined in (66). Note that $\|\mathbf{u}^k\| = 1$, we can assume $\mathbf{u}^k \to \bar{\mathbf{u}} \ne \mathbf{0}$ (by further passing to a subsequence if necessary); hence, we have $M\bar{\mathbf{u}} = \mathbf{0}$ by continuity. Combining (73) and (80), we have

$$M^T(M\mathbf{x}^{2,k} - \mathbf{b}) - \mathbf{y} + \rho(\mathbf{x}^{2,k} - \mathbf{x}^{1,k}) = o(\delta_k).$$

Furthermore,

$$M^T(M\bar{x}^{2,k} - \mathbf{b}) - \mathbf{y} + \rho(\bar{x}^{2,k} - \bar{x}^{1,k}) = 0.$$

Subtracting the above two equalities and using (83), we get

$$\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k} = \mathbf{x}^{1,k} - \bar{\mathbf{x}}^{1,k} + o(\delta_k). \tag{84}$$

Thus,

$$\bar{\mathbf{u}} = \lim_{k\to\infty} \frac{\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}}{\delta^k} = \lim_{k\to\infty} \frac{\mathbf{x}^{1,k} - \bar{\mathbf{x}}^{1,k}}{\delta^k}.$$

Since $\mathbf{u}^k \to \bar{\mathbf{u}} \neq \mathbf{0}$, we have $\langle \mathbf{u}^k, \bar{\mathbf{u}} \rangle > 0$ for k sufficiently large. Select k such that $\langle \mathbf{u}^k, \bar{\mathbf{u}} \rangle > 0$ and let

$$\hat{\mathbf{x}}^{2,k} \triangleq \bar{\mathbf{x}}^{2,k} + \epsilon \bar{\mathbf{u}} \tag{85}$$

for some $\epsilon > 0$. We can show that for $\epsilon > 0$ sufficiently small

$$\hat{\mathbf{x}}^{2,k} \in \mathbf{X}^2(\mathbf{y}), \tag{86}$$

whose proof is relegated to Appendix D. Now, assume $\hat{\mathbf{x}}^{2,k} \in \mathbf{X}^2(\mathbf{y}^k)$ for $\epsilon > 0$ sufficiently small. This leads to the following contradiction:

$$\|\mathbf{x}^{2,k} - \hat{\mathbf{x}}^{2,k}\|_2 = \|\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k} - \epsilon\bar{\mathbf{u}}\|_2 = \delta^k + \epsilon^2 - 2\epsilon \left\langle \mathbf{u}^k, \bar{\mathbf{u}} \right\rangle < \delta^k \tag{87}$$

for ϵ sufficiently small, which contradicts the definition of $\bar{\mathbf{x}}^{2,k}$ in (64). So (82) holds.

By (69), we have

$$\mathbf{0} \in \lambda \partial \sum_{g\in\mathcal{G}} w_g \|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2 + (\mathbf{r}^{1,k} - \mathbf{y} - \rho(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})), \tag{88}$$

which is the optimal condition to

$$\mathbf{r}^{1,k} \in \arg\min_{\mathbf{d}} \lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{d}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2 + \left\langle \mathbf{r}^{1,k} - \mathbf{y} - \rho(\mathbf{x}^{1,k} - \mathbf{x}^{2,k}), \mathbf{d} \right\rangle. \tag{89}$$

From (89), we have

$$\begin{aligned} &\lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2 + \left\langle \mathbf{r}^{1,k} - \mathbf{y} - \rho(\mathbf{x}^{1,k} - \mathbf{x}^{2,k}), \mathbf{r}^{1,k} \right\rangle \\ &\qquad \leq \lambda \sum_{g\in\mathcal{G}} w_g \|\bar{\mathbf{x}}^{1,k}\|_2 + \left\langle \mathbf{r}^{1,k} - \mathbf{y} - \rho(\mathbf{x}^{1,k} - \mathbf{x}^{2,k}), \mathbf{x}^{1,k} - \bar{\mathbf{x}}^{1,k} \right\rangle. \end{aligned} \tag{90}$$

From $\tilde{\nabla}_{\mathbf{x}^1} L_\rho(\bar{\mathbf{x}}^{1,k}, \bar{\mathbf{x}}^{2,k}; \mathbf{y}) = 0$, we have

$$\mathbf{0} = \arg\min_{\mathbf{d}} \lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{d}_{j(g)} - \bar{\mathbf{x}}^1_{j(g)}\|_2 + \frac{1}{2}\|\mathbf{d} - \mathbf{y} - \rho(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)\|_2^2. \tag{91}$$

Similar to (88), we have

$$\mathbf{0} \in \lambda \partial \sum_{g\in\mathcal{G}} w_g \|\bar{\mathbf{x}}^{1,k}_{j(g)}\|_2 + (-\mathbf{y} - \rho(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k})), \tag{92}$$

which is the optimal condition to

$$\mathbf{0} \in \arg\min_{\mathbf{d}} \lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{d}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2 + \left\langle -\mathbf{y} - \rho(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}), \mathbf{d} \right\rangle. \tag{93}$$

From (93), we have

$$\begin{aligned} &\lambda \sum_{g\in\mathcal{G}} w_g \|\bar{\mathbf{x}}^{1,k}_{j(g)}\|_2 \\ &\qquad \leq \lambda \sum_{g\in\mathcal{G}} w_g \|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2 + \left\langle -\mathbf{y} - \rho(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}), \bar{\mathbf{x}}^{1,k} + \mathbf{r}^{1,k} - \mathbf{x}^{1,k} \right\rangle. \end{aligned} \tag{94}$$

Adding (90) and (94), and using (63), we obtain

$$\begin{aligned}&\left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, \mathbf{r}^{1,k}\right\rangle + \left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), \mathbf{x}^{1,k}-\bar{\mathbf{x}}^{1,k}\right\rangle \\ &\quad\le \left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, \mathbf{x}^{1,k}-\bar{\mathbf{x}}^{1,k}\right\rangle + \left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), \mathbf{r}^{1,k}\right\rangle.\end{aligned} \tag{95}$$

From (63), we have

$$\mathbf{x}^{1,k}-\bar{\mathbf{x}}^{1,k} = (\frac{1}{\rho}M^TM+\mathbf{I})(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}) - \frac{1}{\rho}\mathbf{r}^{2,k}. \tag{96}$$

Defining $A \triangleq \frac{1}{\rho}M^TM + I$ and using (96) in (95) and rearranging the terms, we obtain

$$\begin{aligned}&\left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, \mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\right\rangle + \left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), A(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\right\rangle \\ &\quad\le \left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, A(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\right\rangle + \left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), \mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\right\rangle.\end{aligned} \tag{97}$$

Let us consider term by term. We have

$$\left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, \mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\right\rangle \ge \|\mathbf{r}^{1,k}\|_2^2 + \frac{1}{\rho}\|\mathbf{r}^{2,k}\|_2^2 - (\frac{1}{\rho}+1)\|\mathbf{r}^{1,k}\|_2\|\mathbf{r}^{2,k}\|_2. \tag{98}$$

Next, using (82), we have

$$\left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), A(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\right\rangle = \frac{1}{\rho}\|M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\|_2^2 + \|M(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\|_2^2 \tag{99}$$

$$\ge \kappa^2\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2^2. \tag{100}$$

Denote the largest eigenvalue of matrix A by L_1, we have

$$\left\langle \mathbf{r}^{1,k}+\mathbf{r}^{2,k}, A(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k})\right\rangle \le L_1\|\mathbf{r}^{1,k}+\mathbf{r}^{2,k}\|_2\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2. \tag{101}$$

Denote $L_2 \triangleq \max_{\|\mathbf{d}\|=1}\|M\mathbf{d}\|$,

$$\left\langle M^TM(\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}), \mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\right\rangle \le L_2^2\|\mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\|_2\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2. \tag{102}$$

Combining the four inequalities above, we have

$$\begin{aligned}&\kappa^2\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2^2 + \|\mathbf{r}^{1,k}\|_2^2 + \frac{1}{\rho}\|\mathbf{r}^{2,k}\|_2^2 - (\frac{1}{\rho}+1)\|\mathbf{r}^{1,k}\|_2\|\mathbf{r}^{2,k}\|_2 \\ &\quad\le (L_1\|\mathbf{r}^{1,k}+\mathbf{r}^{2,k}\|_2 + L_2^2\|\mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\|_2)\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2.\end{aligned} \tag{103}$$

Denote $b \triangleq L_1\|\mathbf{r}^{1,k}+\mathbf{r}^{2,k}\|_2 + L_2^2\|\mathbf{r}^{1,k}+\frac{1}{\rho}\mathbf{r}^{2,k}\|_2$, $c \triangleq \|\mathbf{r}^{1,k}\|_2^2 + \frac{1}{\rho}\|\mathbf{r}^{2,k}\|_2^2 - (\frac{1}{\rho}+1)\|\mathbf{r}^{1,k}\|_2\|\mathbf{r}^{2,k}\|_2$. Using quadratic formula, (103) implies

$$\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|_2 \le \frac{b+\sqrt{b^2-4\kappa^2 c}}{2\kappa^2}. \tag{104}$$

Note that the right-hand-side of (104) is $\mathcal{O}(\|\mathbf{r}^{1,k}\|+\|\mathbf{r}^{2,k}\|)$, so (104) contradicts (80), which says

$$\|\mathbf{r}^{1,k}\|+\|\mathbf{r}^{2,k}\| = o(\|\mathbf{x}^{2,k}-\bar{\mathbf{x}}^{2,k}\|).$$

So far, we have shown that for a fixed $\mathbf{y}$, there exist τ and δ satisfying (31) and the inequality below it, accordingly. From (70) and (63), we know $\tilde{\nabla}_{\mathbf{x}}L_\rho(\mathbf{x}^1,\mathbf{x}^2;\mathbf{y})$ is continuous in $\mathbf{y}$. Since $\mathbf{X}(\mathbf{y})$ is characterized by $\tilde{\nabla}_{\mathbf{x}}L_\rho(\mathbf{x}^1,\mathbf{x}^2;\mathbf{y})=0$, so $\mathrm{dist}(\mathbf{x},\mathbf{X}(\mathbf{y}))$ is also continuous in $\mathbf{y}$, which implies that we can define a continuous mapping from $\mathbf{y}$ to τ and δ. Note that from the above proof, we know that for any $\mathbf{y}$, τ is finite, i.e., $\tau<\infty$, and $\delta>0$. Hence, since $\mathbf{y}$ is in a compact set, we can find $\bar{\tau} \triangleq \sup\{\tau\} < \infty$ and $\bar{\delta} \triangleq \inf\{\delta\} > 0$. This finishes the proof of the lemma.

Appendix D: Proof of the inclusion (86) in Lemma 3

The following proof is inspired by Tseng (2010) and Zhang et al. (2013). Denote $\mathbf{t}^k \triangleq \mathbf{y} + \rho(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})$. Note that if we write (10) as a function of $\mathbf{x}^2$ and $\mathbf{z} \triangleq \mathbf{x}^1 - \mathbf{x}^2$, then the last term is strongly convex in $\mathbf{z}$. This implies that the value of $\bar{\mathbf{t}} \triangleq \mathbf{y} + \rho(\mathbf{x}^1 - \mathbf{x}^2)$ is unique for $\forall (\mathbf{x}^1, \mathbf{x}^2) \in \mathbf{X}(\mathbf{y})$. Recall the definition in (65) and (85), we will show (86) is equivalent to

$$\mathbf{0} \in \lambda \partial w_g \|(\bar{\mathbf{x}}^{1,k} + \epsilon \bar{\mathbf{u}})_{j(g)}\| + \bar{\mathbf{t}}_{j(g)}, \quad \forall g. \tag{105}$$

From the optimality condition of (10), we know $\hat{\mathbf{x}}^{2,k} \in \mathbf{X}^2(\mathbf{y})$ is equivalent to

$$\begin{cases} \mathbf{0} \in \lambda \partial \sum_{g \in \mathcal{G}} w_g \|\mathbf{x}^1_{j(g)}\|_2 + (\mathbf{y} + \rho(\mathbf{x}^1 - \hat{\mathbf{x}}^2))_{j(g)}, \ \forall g \\ \mathbf{0} = M^T(M\hat{\mathbf{x}}^{2,k} - \mathbf{b}) - (\mathbf{y} + \rho(\mathbf{x}^1 - \hat{\mathbf{x}}^2)), \end{cases} \tag{106}$$

is satisfied for some $\mathbf{x}^1$. From (85), we have

$$M\hat{\mathbf{x}}^2 = M\bar{\mathbf{x}}^2 \tag{107}$$

since $M\bar{\mathbf{u}} = \mathbf{0}$. So the second equality of (106) holds if and only if $\mathbf{x}^1 = \bar{\mathbf{x}}^1 + \epsilon\bar{\mathbf{u}}$.[2] Since $\mathbf{0} = M^T(M\bar{\mathbf{x}}^{2,k} - \mathbf{b}) - (\mathbf{y} + \rho(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2))$ holds by definitions (65) and (64), (106) is equivalent to

$$\mathbf{0} \in \lambda \partial \sum_{g \in \mathcal{G}} w_g \|\bar{\mathbf{x}}^1_{j(g)} + \epsilon\bar{\mathbf{u}}\|_2 + (\mathbf{y} + \rho(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2))_{j(g)} \ \forall g. \tag{108}$$

Using $\bar{\mathbf{t}}$ to replace $(\mathbf{y} + \rho(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2))$, we have (105).

Based on (80) and (83), we have

$$\mathbf{t}^k - \bar{\mathbf{t}} = M^T M(\mathbf{x}^{2,k} - \bar{\mathbf{x}}^{2,k}) - \mathbf{r}^{2,k} = o(\delta^k). \tag{109}$$

By further passing to a subsequence if necessary, we can assume that, for each $g \in \mathcal{G}$, either

1. $\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2 \le \lambda w_g, \quad \forall k$, or,
2. $\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2 > \lambda w_g$, and $\bar{\mathbf{x}}^{1,k}_{j(g)} \neq \mathbf{0}, \quad \forall k$, or,
3. $\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2 > \lambda w_g$, and $\bar{\mathbf{x}}^{1,k}_{j(g)} = \mathbf{0}, \quad \forall k$,

is true. We will show that in any of the three above cases, $\bar{\mathbf{u}}_{j(g)}$ is a certain multiple of $\bar{\mathbf{t}}_{j(g)}$ and then (105) is satisfied.

1. In this case, from (71), we know

$$\bar{\mathbf{u}}_{j(g)} = \lim_{k\to\infty} \frac{\mathbf{x}^{1,k}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}}{\delta^k} = \lim_{k\to\infty} \frac{\mathbf{r}^{1,k}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}}{\delta^k} = \lim_{k\to\infty} \frac{-\bar{\mathbf{x}}^{1,k}_{j(g)}}{\delta^k} \tag{110}$$

   where the last equation comes from (80). Suppose that $\bar{\mathbf{u}}_{j(g)} \neq \mathbf{0}$. (Otherwise, $\hat{\mathbf{x}}^{2,k} = \bar{\mathbf{x}}^{2,k}$.) Then $\bar{\mathbf{x}}^{1,k}_{j(g)} \neq \mathbf{0}$ for all k sufficiently large. From the optimality condition for (10), we have

$$\mathbf{0} = \lambda w_g \frac{\bar{\mathbf{x}}^{1,k}_{j(g)}}{\|\bar{\mathbf{x}}^{1,k}_{j(g)}\|_2} + \bar{\mathbf{t}}_{j(g)}, \tag{111}$$

   for k sufficiently large. By continuity, we have $\bar{\mathbf{u}}_{j(g)}$ is a positive multiple of $\bar{\mathbf{t}}_{j(g)}$. Furthermore, $\bar{\mathbf{x}}^{1,k}_{j(g)}$ is a negative multiple of $\bar{\mathbf{t}}_{j(g)}$. Therefore, for ϵ sufficiently small, (105) is satisfied.

[2] In fact, from our discussion on the uniqueness of $\mathbf{y} + \rho(\mathbf{x}^1 - \mathbf{x}^2)$ for $\forall (\mathbf{x}^1, \mathbf{x}^2) \in \mathbf{X}(\mathbf{y})$, we can also conclude that $\mathbf{x}^1$ must be $\bar{\mathbf{x}}^1 + \epsilon\bar{\mathbf{u}}$.

2. In this case, since we assumed $\bar{\mathbf{x}}^{1,k}_{j(g)} \neq \mathbf{0} \ \ \forall k$, (111) is always satisfied. It implies

$$\bar{\mathbf{t}}_{j(g)} = \lambda w_g \frac{\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2}. \tag{112}$$

From (71), we have

$$\begin{aligned}
\mathbf{r}^{1,k}_{j(g)} &= \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2}\mathbf{x}^{1,k}_{j(g)} + (\frac{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)})\|_2 - \lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2})\mathbf{t}^k_{j(g)} \\
&= \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2}(\bar{\mathbf{x}}^{1,k}_{j(g)} + \delta^k \mathbf{u}^k_{j(g)} + o(\delta^k)) + (\frac{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)})\|_2 - \lambda w_g}{\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2})(\bar{\mathbf{t}}_{j(g)} + o(\delta^k)) \\
&= \frac{\lambda w_g \delta^k}{\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2}\mathbf{u}^k_{j(g)} + \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2}(\bar{\mathbf{x}}^{1,k}_{j(g)} - \bar{\mathbf{t}}_{j(g)}) + \bar{\mathbf{t}}_{j(g)} + o(\delta^k) \\
&= \frac{\lambda w_g \delta^k}{\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2}\mathbf{u}^k_{j(g)} + (\frac{\lambda w_g}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2} - \frac{\lambda w_g}{\|\mathbf{t}^k_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2})(\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}) + o(\delta^k) \\
&= \frac{\lambda w_g \delta^k}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)} - \delta^k \mathbf{u}^k_{j(g)} + o(\delta^k)\|_2}\mathbf{u}^k_{j(g)} \\
&+ (\frac{\lambda w_g}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2} - \frac{\lambda w_g}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)} - \delta^k \mathbf{u}^k_{j(g)} + o(\delta^k)\|_2})(\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}) + o(\delta^k)
\end{aligned}$$

where the second equality comes from (84) and (109). The forth equality follows from (112). Finally, we use (84) and (109) in the last equality. From the Taylor expansion of $\|\cdot\|_2^{-1}$ and given that $\nabla_{\mathbf{x}}\|\mathbf{x}\|_2^{-1} = -\mathbf{x}/\|\mathbf{x}\|_2^3$, we have

$$\begin{aligned}
&\frac{1}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)} - \delta^k \mathbf{u}^k_{j(g)} + o(\delta^k)\|_2} = \frac{1}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2} - \frac{\left\langle \bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}, -\delta^k \mathbf{u}^k_{j(g)} + o(\delta^k) \right\rangle}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2^3} \\
&+ o(\| - \delta^k \mathbf{u}^k_{j(g)} + o(\delta^k)\|_2) \\
&= \frac{1}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2} + \frac{\left\langle \bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}, \delta^k \mathbf{u}^k_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2^3} + o(\delta^k).
\end{aligned}$$

Using this back in the last equation for $\mathbf{r}^{1,k}_{j(g)}$ and rearranging the terms, we have

$$\begin{aligned}
\mathbf{r}^{1,k}_{j(g)} &= \frac{\lambda w_g \delta^k}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2}\mathbf{u}^k_{j(g)} - \frac{\lambda w_g \left\langle \bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}, \delta^k \mathbf{u}^k_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2^3}(\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}) + o(\delta^k) \\
&= \frac{\lambda w_g \delta^k}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2}\mathbf{u}^k_{j(g)} - \frac{\left\langle \bar{\mathbf{t}}_{j(g)}, \delta^k \mathbf{u}^k_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2}(\frac{\bar{\mathbf{t}}_{j(g)}}{\lambda w_g}) + o(\delta^k),
\end{aligned}$$

where the second equality uses (112). Multiplying both sides by $\frac{\|\bar{\mathbf{t}}_{j(g)} - \bar{\mathbf{x}}^{1,k}_{j(g)}\|_2}{\lambda w_g \delta^k}$ and using (80),(81) and $\|\bar{\mathbf{t}}_{j(g)}\|_2 = \lambda w_g$ (from (112)) yields in the limit

$$\mathbf{0} = \bar{\mathbf{u}}_{j(g)} - \frac{\left\langle \bar{\mathbf{t}}_{j(g)}, \bar{\mathbf{u}}_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)}\|_2^2}\bar{\mathbf{t}}_{j(g)}. \tag{113}$$

Thus $\bar{\mathbf{u}}_{j(g)}$ is a nonzero multiple of $\bar{\mathbf{t}}_{j(g)}$. In this case, since we assume $\bar{\mathbf{x}}^{1,k}_{j(g)} \neq \mathbf{0}$, from (111), we know $\bar{\mathbf{x}}^{1,k}_{j(g)}$ is a negative multiple of $\bar{\mathbf{t}}_{j(g)}$. So (105) is satisfied for ϵ sufficiently small.

3. In this case, we assume $\bar{\mathbf{x}}^{1,k}_{j(g)} = \mathbf{0}, \forall k$, from (81), we have $\bar{\mathbf{x}}^{1}_{j(g)} = \mathbf{0}$. We also assume that $\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^{k}_{j(g)}\|_2 > \lambda w_g$ for all k, this implies $\|\bar{\mathbf{t}}_{j(g)}\|_2 \geq \lambda w_g$. From the optimality condition for (10) for $\mathbf{x}^1$ we have

$$\mathbf{0} = \bar{\mathbf{t}}_{j(g)} + \lambda w_g \partial \|\mathbf{0}\|_2,$$

which implies $\|\bar{\mathbf{t}}_{j(g)}\|_2 \leq \lambda w_g$. Thus $\|\bar{\mathbf{t}}_{j(g)}\|_2 = \lambda w_g$. Then (71) implies

$$\begin{aligned}
\mathbf{r}^{1,k}_{j(g)} &= \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2}\mathbf{x}^{1,k}_{j(g)} + (\frac{\lambda w_g}{\|\bar{\mathbf{t}}_{j(g)}\|_2} - \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2})\mathbf{t}^k_{j(g)} \\
&= \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2}\mathbf{x}^{1,k}_{j(g)} + \frac{\lambda w_g \left\langle \bar{\mathbf{t}}_{j(g)}, \mathbf{t}^k_{j(g)} - \mathbf{x}^{1,k}_{j(g)} - \bar{\mathbf{t}}_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)}\|_2^3}\mathbf{t}_{j(g)} \\
&+ o(\|\mathbf{t}^k_{j(g)} - \mathbf{x}^{1,k}_j(g) - \bar{\mathbf{t}}_{j(g)}\|_2) \\
&= \frac{\lambda w_g}{\|\mathbf{x}^{1,k}_j(g) - \mathbf{t}^k_{j(g)}\|_2}\mathbf{x}^{1,k}_{j(g)} - \frac{\lambda w_g \left\langle \bar{\mathbf{t}}_{j(g)}, \mathbf{x}^{1,k}_{j(g)} \right\rangle}{\|\bar{\mathbf{t}}_{j(g)}\|_2^3}\mathbf{t}_{j(g)} + o(\delta^k)
\end{aligned}$$

where the second equality uses Taylor expansion similar to the case 2. The third equality follows from (109) and $\{\mathbf{x}^{1,k}_{j(g)}\} \to \mathbf{0}$. Dividing both sides by δ^k yield in the limit (113), where it uses

$$\{\frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k}\} = \{\mathbf{u}^k_{j(g)} + \frac{o(\delta^k)}{\delta^k}\} \to \bar{\mathbf{u}}_{j(g)}.$$

Since we assume $\|\mathbf{x}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}\|_2 > \lambda w_g$ for all k, we have the following equality from (70)

$$\mathbf{0} = \lambda w_g \frac{\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}}{\|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2} + \mathbf{r}^{1,k}_{j(g)} - \mathbf{t}^k_{j(g)}. \tag{114}$$

Suppose $\bar{\mathbf{u}}_{j(g)} \neq \mathbf{0}$. Then $\mathbf{u}^k_{j(g)} = \frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k} + \frac{o(\delta^k)}{\delta^k} \neq \mathbf{0}$, for k sufficiently large. It implies that $\mathbf{x}^{1,k}_{j(g)} \neq \mathbf{0}$, for k sufficiently large. Hence,

$$\begin{aligned}
\left\langle \bar{\mathbf{t}}_{j(g)}, \bar{\mathbf{u}}_{j(g)} \right\rangle &= \lim_{k\to+\infty} \left\langle \mathbf{t}^k_{j(g)}, \mathbf{u}^k_{j(g)} \right\rangle \\
&= \lim_{k\to+\infty} \left\langle \mathbf{r}^{1,k}_{j(g)}, \frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k} \right\rangle + \left\langle \lambda w_g \frac{\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}}{\|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2}, \frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k} \right\rangle \\
&= \lim_{k\to+\infty} \frac{\lambda w_g}{\|\mathbf{r}^{1,k}_{j(g)} - \mathbf{x}^{1,k}_{j(g)}\|_2}(\frac{\left\langle \mathbf{r}^{1,k}_{j(g)}, \mathbf{x}^{1,k}_{j(g)} \right\rangle}{\delta^k} - \frac{\|\mathbf{x}^{1,k}_{j(g)}\|_2^2}{\delta^k}) \\
&= \lim_{k\to+\infty} \frac{\lambda w_g}{\|\frac{\mathbf{r}^{1,k}_{j(g)}}{\|\mathbf{x}^{1,k}_{j(g)}\|_2} - \frac{\mathbf{x}^{1,k}_{j(g)}}{\|\mathbf{x}^{1,k}_{j(g)}\|_2}\|_2}(\left\langle \frac{\mathbf{r}^{1,k}_{j(g)}}{\delta^k}, \frac{\mathbf{x}^{1,k}_{j(g)}}{\|\mathbf{x}^{1,k}_{j(g)}\|_2} \right\rangle - \|\mathbf{u}^k_{j(g)}\|_2) \\
&= -\lambda w_g \|\mathbf{u}^k_{j(g)}\|_2 < 0,
\end{aligned}$$

where the second equality is based on (114) and $\lim_{k\to+\infty} \mathbf{u}^k_{j(g)} = \lim_{k\to+\infty} \frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k}$, the third equality is based on $\frac{\mathbf{r}^{1,k}_{j(g)}}{\delta^k} \to \mathbf{0}$ (by (80)), the forth equality is based on $\mathbf{x}^{1,k}_{j(g)} \neq \mathbf{0}$ and $\lim_{k\to+\infty} \mathbf{u}^k_{j(g)} = \lim_{k\to+\infty} \frac{\mathbf{x}^{1,k}_{j(g)}}{\delta^k}$, and, the fifth equality is based on $\mathbf{r}^{1,k}_{j(g)} \to \mathbf{0}$ and $\frac{\mathbf{r}^{1,k}_{j(g)}}{\delta^k} \to \mathbf{0}$ (by (80)). Finally, combined with (113), we obtain $\bar{\mathbf{u}}_{j(g)}$ is a negative multiplier of $\bar{\mathbf{t}}_{j(g)}$. Since $\bar{\mathbf{x}}^1_{j(g)} = \mathbf{0}$ in this case, (105) is satisfied.

Appendix E: Proof of Theorem 1

From Lemmas 2.4 and 2.5 in Hong and Luo (2017), we have the following two identities, respectively:

$$L(\mathbf{x}^k;\mathbf{y}^k) - L(\mathbf{x}^{k+1};\mathbf{y}^k) \geq \rho\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2, \tag{115}$$

$$\|\tilde{\nabla} L(\mathbf{x}^k;\mathbf{y}^k)\| \leq \sigma\|\mathbf{x}^k - \mathbf{x}^{k+1}\|, \tag{116}$$

where $\sigma = \sqrt{2}(\max\{1 + \|M^T\|\|M\|, \rho\} + 1)$.

Lemma 2 in our paper establishes the uniform boundedness of iterates $\{(\mathbf{x}^k, \mathbf{y}^k)\}$, based on which the compactness condition of Lemma 3 is satisfied. Lemma 3 quantifies the primal error bound with the proximal gradient of the Lagrangian function as

$$\|\mathbf{x}^k - \bar{\mathbf{x}}^k\| \leq \tau_p\|\tilde{\nabla} L(\mathbf{x}^k;\mathbf{y}^k)\| \leq \tau_p\sigma\|\mathbf{x}^k - \mathbf{x}^{k+1}\|, \tag{117}$$

where $\bar{\mathbf{x}}^k = \arg\min_{\bar{\mathbf{x}}\in\mathbf{X}(\mathbf{y}^k)} \|\bar{\mathbf{x}} - \mathbf{x}^k\|$ and the second inequality comes from (116).

In Lemma 1, we show

$$\mathrm{dist}(\mathbf{y}, Y^*) \leq \tau_d\|\nabla g_\rho(\mathbf{y})\|_2, \tag{118}$$

where $\tau_d = \max\{\|M^T\|^2, 2\rho\}$ as shown in the proof of Lemma 1. Based on the dual error bound (118) and following Lemma 3.1 in Hong and Luo (2017), we have

$$\Delta_d^k \leq \tau'\|\nabla g_\rho(\mathbf{y}^k)\|^2 = \tau'\|\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}\|^2, \tag{119}$$

$$\Delta_p^k \leq \zeta\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 + \zeta'\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq (\xi + \xi'\tau_p^2\sigma^2)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \tag{120}$$

where $\bar{\mathbf{x}}^{1,k}$ and $\bar{\mathbf{x}}^{2,k}$ represent the upper half and lower half of the vector $\bar{\mathbf{x}}^k$, correspondingly, $\tau' = \tau_d^2/\rho$, and

$$\zeta = 2A + \frac{3\sqrt{2}}{2}(\sigma - 1), \tag{121}$$

$$\zeta' = 2A + \frac{1}{2} + \frac{\sqrt{2}}{2}(\sigma - 1), \tag{122}$$

where $A = \|M^T\|\|M\| + \sqrt{2}\rho$.

Following Lemmas 3.2 and 3.3 in Hong and Luo (2017), we have

$$\Delta_d^k - \Delta_d^{k-1} \leq -\alpha(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})^T(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}), \tag{123}$$

$$\Delta_p^k - \Delta_p^{k-1} \leq \alpha\|\mathbf{x}^{1,k} - \mathbf{x}^{2,k}\|^2 - \gamma\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \alpha(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})^T(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}), \tag{124}$$

where Δ_p^k and Δ_d^k are primal and dual optimality gaps at iteration k (defined in the statement of Theorem 1), α is the stepsize, and in (124), we have used (115). Adding (123) and (124), we have

$$[\Delta_d^k + \Delta_p^k] - [\Delta_d^{k-1} + \Delta_p^{k-1}] \leq \alpha\|\mathbf{x}^{1,k} - \mathbf{x}^{2,k}\|^2 - \gamma\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - 2\alpha(\mathbf{x}^{1,k} - \mathbf{x}^{2,k})^T(\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}) \tag{125}$$

$$= \alpha\|\mathbf{x}^{1,k} - \mathbf{x}^{2,k} - \bar{\mathbf{x}}^{1,k} + \bar{\mathbf{x}}^{2,k}\|^2 - \alpha\|\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}\|^2 - \rho\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \tag{126}$$

$$\leq (2\alpha\tau_p^2\sigma^2 - \rho)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \alpha\|\bar{\mathbf{x}}^{1,k} - \bar{\mathbf{x}}^{2,k}\|^2, \tag{127}$$

where the last inequality comes from (117) and the Cauchy-Schwarz inequality.

Assuming that the stepsize α is chosen sufficient small such that $0 < \alpha < \frac{\rho}{2\tau_p^2\sigma^2}$, and substituting (119) and (120) into (127), we have

$$[\Delta_d^k + \Delta_p^k] - [\Delta_d^{k-1} + \Delta_p^{k-1}] \leq -\frac{\rho - 2\alpha\tau_p^2\sigma^2}{\xi + \xi'\tau_p\sigma^2}\Delta_p^k - \frac{\alpha}{\tau'}\Delta_d^k \tag{128}$$

$$\leq -\min\{\frac{\rho - 2\alpha\tau_p^2\sigma^2}{\xi + \xi'\tau_p\sigma^2}, \frac{\alpha}{\tau'}\}[\Delta_d^k + \Delta_p^k]. \tag{129}$$

Therefore, we have

$$0 \leq [\Delta_p^k + \Delta_d^k] \leq \frac{1}{\lambda + 1}[\Delta_p^{k-1} + \Delta_d^{k-1}], \tag{130}$$

where $\lambda = \min\{\frac{\rho - 2\alpha\tau_p^2\sigma^2}{\zeta + \zeta'\tau_p^2\sigma^2}, \frac{\alpha}{\tau'}\} > 0$. Therefore, $[\Delta_p^k + \Delta_d^k] \leq (\frac{1}{\lambda+1})^k[\Delta_p^0 + \Delta_d^0]$, implying that Δ_p^k and Δ_d^k converges to zero Q-linearly.

Appendix F: Reference list of some parameters used in the analysis

- λ: Check Equation (4)
- ρ: Check Equation (10)
- α: Check Equation (15)
- τ_d: Check Lemma 1, and Equation (40) in the online supplement
- τ_p: Check Lemma 3, and Equation (77) in the online supplement
- δ: Check Lemma 3, and Equation (72) in the online supplement
- σ: Check Theorem 1, and Equation (116) in the online supplement
- ζ: Check Theorem 1, and Equation (121) in the online supplement
- ζ': Check Theorem 1, and Equation (122) in the online supplement
- τ': Check Theorem 1, and Equation (119) in the online supplement