A Proofs

In this section, we show the detail proofs of Theorem 1, 2 and 3.

A.1 Proof of Theorem 1

Proof.

(1) $\Longrightarrow$ (2) The proof is followed directly from the definition of $\mathcal{S}_k$.

(2) $\Longrightarrow$ (3) As $\mathcal{S}_k = \emptyset$, we obtain $\mathcal{F}_k = \{i : \mathbf{x}_i^k > 0 \vee \partial f(\mathbf{x}^k)/\partial \mathbf{x}_i = 0\}$. We obtain the desired result.

(3) $\Longrightarrow$ (4) The definition of $\mathbf{d}^k$ implies $\mathbf{d}^k = 0$.

(4) $\Longrightarrow$ (5) As $\mathbf{d}^k = 0$, we certainly have $\mathbf{x}^k(\alpha) = \mathbf{x}^k$ for all $\alpha > 0$.

(5) $\Longrightarrow$ (1) Assume $\mathbf{x}^k(\alpha) = \mathbf{x}^k$ for all $\alpha > 0$. For $i \in \mathcal{W}_k$, we must have

$$\mathbf{d}_i^k = 0, \qquad \text{if } \mathbf{x}_i^k > 0, \tag{28}$$

$$\mathbf{d}_i^k \leq 0, \qquad \text{if } \mathbf{x}_i^k = 0. \tag{29}$$

By the definition of $\mathbf{W}_k$, if $i \in \mathcal{W}_k$ and $\mathbf{x}_i^k > 0$, then we have $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i \cdot \mathbf{d}_i^k = 0$; if $i \in \mathcal{W}_k$ and $\mathbf{x}_i^k = 0$, then we have $\partial f(x^k)/\partial \mathbf{x}_i \leq 0$ and so $\partial f(x^k)/\partial \mathbf{x}_i \cdot \mathbf{d}_i^k \geq 0$. Therefore, we obtain

$$\sum_{i \in \mathcal{W}_k} \frac{\partial f(x^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k \geq 0. \tag{30}$$

By the definition of $\mathbf{d}^k$, however, we have

$$\nabla f(\mathbf{x}^k)^T \mathbf{d}^k = (\mathbf{g}^k)^T \mathbf{d}_{\mathcal{W}_k}^k = -(\mathbf{g}^k)^T \mathbf{M}_k \mathbf{g}^k < 0, \tag{31}$$

for all $\mathbf{g}^k \neq 0$ due to the positive definite matrix $\mathbf{M}_k$. Combining inequalities (30) and (31), we have $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i = 0$ for all $i \in \mathcal{W}_k$. Since the subspace $\mathcal{W}_k$ is valid, we have $\mathcal{W}_k \cap \mathcal{S}_k \neq \emptyset$. However, for all $i \in \mathcal{S}_k$, we have $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i \neq 0$. Therefore, we have $\mathcal{S}_k = \emptyset$ which implies $\mathbf{x}^k$ is a stationary point. □

A.2 Proof of Theorem 2

Proof. Denote $\mathcal{S}_k^+ = \mathcal{W}_k \cap \mathcal{S}_k$. Consider the index sets

$$\mathcal{W}_k^1 = \{i \in \mathcal{W}_k : \left(\mathbf{x}_i^k > 0 \text{ and } \mathbf{d}_i^k \neq 0\right), \text{ or } \left(\mathbf{x}_i^k = 0 \text{ and } \mathbf{d}_i^k > 0\right)\}, \tag{32}$$

$$\mathcal{W}_k^2 = \{i \in \mathcal{W}_k : \left(\mathbf{x}_i^k > 0 \text{ and } \mathbf{d}_i^k = 0\right), \text{ or } \left(\mathbf{x}_i^k = 0 \text{ and } \mathbf{d}_i^k \leq 0\right)\}. \tag{33}$$

Note that we have $\mathcal{W}_k = \mathcal{W}_k^1 \cup \mathcal{W}_k^2$ and, for all $i \in \mathcal{W}_k^2$, we have

$$\mathbf{x}_i^k(\alpha) = \mathbf{x}_i^k, \quad \forall \alpha > 0. \tag{34}$$

To make progress, the index set $\mathcal{W}_k^1$ cannot be empty. We assume the contrary, that is, the index set $\mathcal{W}_k^1 = \emptyset$. For $j \in \mathcal{W}_k \setminus \mathcal{S}_k^+$ and $\mathbf{x}_j^k > 0$, we have $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i = 0$ based on (8), and hence

$$\mathbf{d}_i^k = -\sum_{j \in \mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j}, \quad \forall i \in \mathcal{S}_k^+, \tag{35}$$

where $\mathbf{M}_{ij}^k$ is the i-th entry of the j-th column of the matrix $\mathbf{M}_k$. By the assumption of $\mathcal{W}_k^1 = \emptyset$, for all $i \in \mathcal{S}_k^+$, we have $\mathbf{d}_i^k = 0$ if $\mathbf{x}_i^k > 0$; and $\mathbf{d}_i^k \leq 0$ if $\mathbf{x}_i^k = 0$. Thus, we obtain for all $i \in \mathcal{S}_k^+$

$$\mathbf{d}_i^k = -\sum_{j \in \mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j} = 0, \quad \text{if } \mathbf{x}_i^k > 0, \tag{36}$$

$$\mathbf{d}_i^k = -\sum_{j \in \mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j} \leq 0, \quad \text{if } \mathbf{x}_i^k = 0. \tag{37}$$

Multiplying $\mathbf{d}_i^k$ by $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i$ yields

$$\frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k = -\frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \sum_{j \in \mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j} = 0, \quad \text{if } \mathbf{x}_i > 0, \tag{38}$$

$$\frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k = -\frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \sum_{j \in \mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j} \geq 0, \quad \text{if } \mathbf{x}_i = 0, \tag{39}$$

where the sign is changed in (39) due to $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i < 0$ for $\mathbf{x}_i^k = 0$. Taking the summation over all $i \in \mathcal{S}_k^+$ gives

$$\sum_{i\in\mathcal{S}_k^+} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k = -\sum_{i\in\mathcal{S}_k^+} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \sum_{j\in\mathcal{S}_k^+} \mathbf{M}_{ij}^k \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_j} \geq 0. \tag{40}$$

Notice that the left-hand side of (40) is actually

$$\sum_{i\in\mathcal{S}_k^+} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k = \sum_{i\in\mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k = -(\mathbf{g}^k)^T\mathbf{M}_k\mathbf{g}^k < 0, \tag{41}$$

since $\mathbf{M}_k$ is positive definite. The inequality (41) contradicts (40). Therefore, the index set $\mathcal{W}_k^1 \neq \emptyset$.

Next, we shall show we can induce a feasible descent direction according to $\mathcal{W}_k^1$ so that an objective reduction is possible. As $\mathcal{W}_k^1 \neq \emptyset$, define a stepsize α_1 such that

$$\alpha_1 = \sup\{\alpha : \mathbf{x}_i^k + \alpha\mathbf{d}_i^k \geq 0, \forall i \in \mathcal{W}_k^1\} \tag{42}$$

Here, the stepsize α_1 is either a positive number or $+\infty$. Define a vector $\overline{\mathbf{d}}^k$ with coordinates

$$\mathbf{d}_i^k = \begin{cases} \mathbf{d}_i^k, & \text{if } i \in \mathcal{W}_k^1 \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

and it follows that for all $\alpha \in (0, \alpha_1]$, we have

$$\mathbf{x}_i^k(\alpha) = \mathbf{x}_i^k + \alpha\mathbf{d}_i^k, \quad \forall i \in [n], \tag{44}$$

which implies $\mathbf{d}^k$ is a feasible direction. For all $i \in \mathcal{W}_k^2$, if $\mathbf{x}_i^k = 0$, then $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i^k < 0$, and so $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i \cdot \mathbf{d}_i \geq 0$; and if $\mathbf{x}_i^k > 0$, then $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i^k \cdot \mathbf{d}_i = 0$. In summary, we obtain

$$\sum_{i\in\mathcal{W}_k^2} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k \geq 0. \tag{45}$$

It follows that

$$\sum_{i\in\mathcal{W}_k^1} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k \leq \sum_{i\in\mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}_i^k = -(\mathbf{g}^k)^T\mathbf{M}_k\mathbf{g}^k < 0, \tag{46}$$

where the last inequality is due to the definition of the vector $\mathbf{d}^k$ and the p.d. matrix $\mathbf{M}_k$. Combining the relations (46) and (44), it implies that the vector $\mathbf{d}^k$ is a feasible descent direction. Therefore, there must exist a stepsize $\overline{\alpha} \leq \alpha_1$ such that

$$f[\mathbf{x}^k(\alpha)] < f(\mathbf{x}^k), \quad \forall \alpha \in (0, \overline{\alpha}]. \tag{47}$$

□

A.3 Proof of Theorem 3

Proof. The sequence $\{\mathbf{x}^k\}$ has at least one limit point since the level set $\mathcal{X}$ is compact. We assume the contrary, that is, there exists a subsequence $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ converging to a vector $\overline{x}$ which is not a first-order stationary point. Since $\{f(\mathbf{x}^k)\}$ is monotonically decreasing and X is compact, it follows that $\{f(\mathbf{x}^k)\}_{k\in\mathcal{K}}$ converges to $f(\overline{x})$ and

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \to 0.$$

Then the right-hand side of (12) converges to 0, and we must have

$$-\sigma\alpha^k\nabla f(\mathbf{x}^k)^T\mathbf{d}^k \to 0. \tag{48}$$

As the eigenvalues of $\{\mathbf{M}_k\}$ is bounded above and away from zero by (14), we have

$$\nabla f(\mathbf{x}^k)^T\mathbf{d}^k = (\mathbf{g}^k)^T\mathbf{d}_{\mathcal{W}_k}^k = -(\mathbf{g}^k)^T\mathbf{M}_k\mathbf{g}^k \leq -c_1\left\|\mathbf{g}^k\right\|^2. \tag{49}$$

Since $\overline{x}$ is not first-order stationary point, we have $\mathbf{g}^k \neq 0$, and from (48) and (49) we must have

$$\liminf_{k\to\infty, k\in\mathcal{K}} \alpha^k \to 0. \tag{50}$$

We shall complete the proof by showing that $\{\alpha^k\}$ is bounded away from zero if $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ converges to a point that is not first-order stationary point, whereby shows the contradiction.

By Lipschitz continuity in (13) and boundedness of $\{\mathbf{x}^k\}$, for all $\alpha \geq 0$, we have

$$\begin{aligned}
f[\mathbf{x}^k(\alpha)] - f(\mathbf{x}^k) =& \int_0^1 \nabla f(\mathbf{x}^k + t(\mathbf{x}^k(\alpha) - \mathbf{x}^k))^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k)dt \\
=& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \int_0^1 (\nabla f(\mathbf{x}^k + t(\mathbf{x}^k(\alpha) - \mathbf{x}^k)) - \nabla f(\mathbf{x}^k))^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k)dt \\
\leq& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \left| \int_0^1 (\nabla f(\mathbf{x}^k + t(\mathbf{x}^k(\alpha) - \mathbf{x}^k)) - \nabla f(\mathbf{x}^k))^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k)dt \right| \\
\leq& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \int_0^1 \left\| \nabla f(\mathbf{x}^k + t(\mathbf{x}^k(\alpha) - \mathbf{x}^k)) - \nabla f(\mathbf{x}^k) \right\| \cdot \left\| \mathbf{x}^k(\alpha) - \mathbf{x}^k \right\| dt \\
\leq& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \left\| \mathbf{x}^k(\alpha) - \mathbf{x}^k \right\| \int_0^1 Lt \left\| \mathbf{x}^k(\alpha) - \mathbf{x}^k \right\| dt \\
=& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \frac{L}{2} \left\| \mathbf{x}^k(\alpha) - \mathbf{x}^k \right\|^2
\end{aligned}$$

By the definition of $\mathbf{d}^k$ in (3) and the nonexpansive property of projection, we obtain

$$\begin{aligned}
f[\mathbf{x}^k(\alpha)] - f(\mathbf{x}^k) \leq& \nabla f(\mathbf{x}^k)^T(\mathbf{x}^k(\alpha) - \mathbf{x}^k) + \frac{L}{2} \left\| \mathbf{x}^k(\alpha) - \mathbf{x}^k \right\|^2 \\
=& (\mathbf{g}^k)^T(\mathbf{x}^k_{\mathcal{W}_k}(\alpha) - \mathbf{x}^k_{\mathcal{W}_k}) + \frac{L}{2} \left\| \mathbf{x}^k_{\mathcal{W}_k}(\alpha) - \mathbf{x}^k_{\mathcal{W}_k} \right\|^2 \\
=& (\mathbf{g}^k)^T(\mathbf{x}^k_{\mathcal{W}_k}(\alpha) - \mathbf{x}^k_{\mathcal{W}_k}) + \frac{L}{2} \left\| \pi(\mathbf{x}^k_{\mathcal{W}_k} + \alpha\mathbf{d}^k_{\mathcal{W}_k}) - \mathbf{x}^k_{\mathcal{W}_k} \right\|^2 \\
\leq& (\mathbf{g}^k)^T(\mathbf{x}^k_{\mathcal{W}_k}(\alpha) - \mathbf{x}^k_{\mathcal{W}_k}) + \frac{L}{2} \left\| \mathbf{x}^k_{\mathcal{W}_k} + \alpha\mathbf{d}^k_{\mathcal{W}_k} - \mathbf{x}^k_{\mathcal{W}_k} \right\|^2 \\
=& (\mathbf{g}^k)^T(\mathbf{x}^k_{\mathcal{W}_k}(\alpha) - \mathbf{x}^k_{\mathcal{W}_k}) + \frac{\alpha^2 L}{2} \left\| \mathbf{d}^k_{\mathcal{W}_k} \right\|^2
\end{aligned}$$

By the definition of $\mathbf{d}^k_{\mathcal{W}_k} = -\mathbf{M}_k\mathbf{g}^k$ and inequalities in (14), we have

$$\left\| \mathbf{d}^k_{\mathcal{W}_k} \right\|^2 = (\mathbf{g}^k)^T M_k^2 \mathbf{g}^k \leq c_2(\mathbf{g}^k)^T\mathbf{M}_k\mathbf{g}^k = -c_2(\mathbf{g}^k)^T\mathbf{d}_{\mathcal{W}_k} \tag{51}$$

and it follows that

$$f[\mathbf{x}^k(\alpha)] - f(\mathbf{x}^k) \leq \sum_{i\in\mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}(\mathbf{x}^k_i(\alpha) - \mathbf{x}^k_i) - \frac{\alpha^2 L c_2}{2} \sum_{i\in\mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}^k_i \tag{52}$$

Consider the index sets

$$\mathcal{W}^1_k = \{i \in \mathcal{W}_k : \mathbf{x}^k_i > 0\}, \tag{53}$$

$$\mathcal{W}^2_k = \{i \in \mathcal{W}_k : \mathbf{x}^k_i = 0\}. \tag{54}$$

For all $i \in \mathcal{W}^1_k$, there exists a scalar $\varepsilon^k_i > 0$ such that $\mathbf{x}^k_i \geq \varepsilon^k_i$. Let $\varepsilon^k = \min\{\varepsilon^k_i : \forall i \in \mathcal{W}_k\}$. Since $\overline{x}$ is not a first-order stationary point, we have $\overline{x} \neq \overline{x}(\alpha)$ for some $\alpha > 0$, and we must have $\liminf_{k\to\infty, k\in\mathcal{K}} \varepsilon^k > 0$. Let $\overline{\varepsilon} > 0$ be a fixed scalar such that $\overline{\varepsilon} \leq \varepsilon^k$ for all k. The equation (14) implies that

$$\left\| \mathbf{d}^k \right\|^2 = \left\| \mathbf{d}^k_{\mathcal{W}_k} \right\|^2 = (\mathbf{g}^k)^T M_k^2 \mathbf{g}^k \leq c_2^2 \left\| \mathbf{g}^k \right\|^2. \tag{55}$$

Hence $\{\mathbf{d}^k\}$ is bounded as f is twice continuously differentiable and $\{\mathbf{x}^k\}$ is bounded. Then there must exist a scalar $\xi > 0$ such that $\left|\mathbf{d}^k_i\right| \leq \xi$ for all i and k. Therefore, for $\alpha \in [0, \overline{\varepsilon}/\xi]$, we have

$$\mathbf{x}^k_i + \alpha\mathbf{d}^k_i \geq \overline{\varepsilon} + \alpha\mathbf{d}^k_i \geq \overline{\varepsilon} - \alpha\xi \geq 0, \quad \forall i \in \mathcal{W}^1_k \tag{56}$$

which implies $\mathbf{x}^k_i(\alpha) = \mathbf{x}^k_i + \alpha\mathbf{d}^k_i$, and it follows

$$\sum_{i\in\mathcal{W}^1_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}(\mathbf{x}^k_i(\alpha) - \mathbf{x}^k_i) = \alpha \sum_{i\in\mathcal{W}^1_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i}\mathbf{d}^k_i, \quad \forall \alpha \in [0, \overline{\varepsilon}/\xi]. \tag{57}$$

For all $i \in \mathcal{W}_k^2$, if $\mathbf{d}_i^k \leq 0$, then $\mathbf{x}_i(\alpha) = \mathbf{x}_i$; if $\mathbf{d}_i^k > 0$, then there exists a stepsize $\overline{\alpha}^k > 0$ such that $\mathbf{x}_i^k(\alpha) = \mathbf{x}_i^k + \alpha \mathbf{d}_i^k$ for all $\alpha \in [0, \overline{\alpha}^k]$. In summary, we have $\mathbf{x}_i^k(\alpha) - \mathbf{x}_i^k \geq \alpha \mathbf{d}_i^k$ for all $i \in \mathcal{W}_k^2$. Since $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i < 0$ for all $i \in \mathcal{W}_k^2$, we have

$$\sum_{i \in \mathcal{W}_k^2} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} (\mathbf{x}_i^k(\alpha) - x_i^k) \leq \alpha \sum_{i \in \mathcal{W}_k^2} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k, \quad \forall \alpha \in [0, \overline{\alpha}], \tag{58}$$

where $\overline{\alpha} = \min\{\overline{\alpha}^k : \forall k\}$.

By (52), (57) and (58), it follows that, for all $0 \leq \alpha \leq \min\{\overline{\varepsilon}/\xi, \overline{\alpha}\}$, we obtain

$$f[\mathbf{x}^k(\alpha)] - f(\mathbf{x}^k) \leq \alpha \sum_{i \in \mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k - \frac{\alpha^2 L c_2}{2} \sum_{i \in \mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k = \alpha \left(1 - \frac{\alpha L c_2}{2}\right) \sum_{i \in \mathcal{W}} \frac{\partial f(\mathbf{x}^k)}{\partial x_i} \mathbf{d}_i^k \tag{59}$$

Suppose α is chosen to satisfy

$$0 < \alpha \leq \frac{\overline{\varepsilon}}{\xi}, \quad 0 < \alpha \leq \frac{2(1-\sigma)}{L c_2}, \quad 0 < \alpha \leq \overline{\alpha}, \quad 0 < \alpha \leq 1, \tag{60}$$

or equivalently

$$0 < \alpha \leq \min\left\{\frac{\overline{\varepsilon}}{\xi}, \frac{2(1-\sigma)}{L c_2}, \overline{\alpha}, 1\right\}, \tag{61}$$

For all $k \in \mathcal{K}$, we must have

$$f[\mathbf{x}^k(\alpha)] - f(\mathbf{x}^k) \leq \sigma \alpha \sum_{i \in \mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k \tag{62}$$

That means that if (61) is satisfied with $\beta^m = \alpha$, then the condition (12) of the Armijo-like rule will be satisfied. Let α^k be such stepsize, then

$$\alpha^k \geq \beta \min\left\{\frac{\overline{\varepsilon}}{\xi}, \frac{2(1-\sigma)}{L c_2}, \overline{\alpha}, 1\right\} > 0 \quad \forall k \in \mathcal{K}, \tag{63}$$

which contradicts (50) and proves the convergence.

Next, we show the convergence rate of SPM is at least sublinear. Recall the definition of the optimality gap as the following:

$$\nabla \mathcal{G}(\mathbf{x}^k) = \mathbf{x}^k - \pi(\mathbf{x}^k - \nabla f(\mathbf{x}^k)). \tag{64}$$

If $\nabla \mathcal{G} = 0$, then $\mathbf{x}^k$ is a stationary point [Bertsekas, 2014]. We first that show $\mathbf{g}^k = 0$ implies $\nabla \mathcal{G} = 0$. By the definition of valid working set, $\mathbf{g}^k = 0$ implies $\mathcal{S}_k = \emptyset$, which further implies $\partial f(\mathbf{x}^k)/\mathbf{x}_i = 0$ for all $i \in \mathcal{F}_k$. In other words, we have if $\mathbf{x}_i^k > 0$, then $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i = 0$; if $\mathbf{x}_i^k = 0$, then $\partial f(\mathbf{x}^k)/\partial \mathbf{x}_i \geq 0$. Clearly, this condition implies $\nabla \mathcal{G}(\mathbf{x}^k) = 0$.

From (61) and (62), we have

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq \sigma \alpha^k \sum_{i \in \mathcal{W}_k} \frac{\partial f(\mathbf{x}^k)}{\partial \mathbf{x}_i} \mathbf{d}_i^k = \sigma \alpha^k (\mathbf{g}^k)^T \mathbf{d}^k = -\sigma \alpha^k (\mathbf{g}^k)^T \mathbf{M}_k \mathbf{g}^k \leq -\sigma c_1 \alpha^k \left\|\mathbf{g}^k\right\|^2 \tag{65}$$

By applying the telescope sum, we have

$$\sum_{i=0}^{k-1} \sigma c_1 \alpha^k \left\|\mathbf{g}^i\right\|^2 \leq \sum_{i=0}^{k-1} f(\mathbf{x}^{i+1}) - f(\mathbf{x}^i) = f(\mathbf{x}^0) - f(\mathbf{x}^k) \leq f(\mathbf{x}^0) - f^\star, \tag{66}$$

where $f^\star = \inf_{\mathbf{x} \geq 0} f(\mathbf{x})$. It implies

$$\left(\min_i \left\|\mathbf{g}^i\right\|^2\right) \sum_{i=0}^{k-1} \sigma c_1 \alpha^k \leq \sum_{i=0}^{k-1} \sigma c_1 \alpha^k \left\|\mathbf{g}^i\right\|^2 \leq f(\mathbf{x}^0) - f^*. \tag{67}$$

By (63), there must exists a scale $\delta > 0$ such that $\alpha^k \geq \delta$ for k. Then we obtain

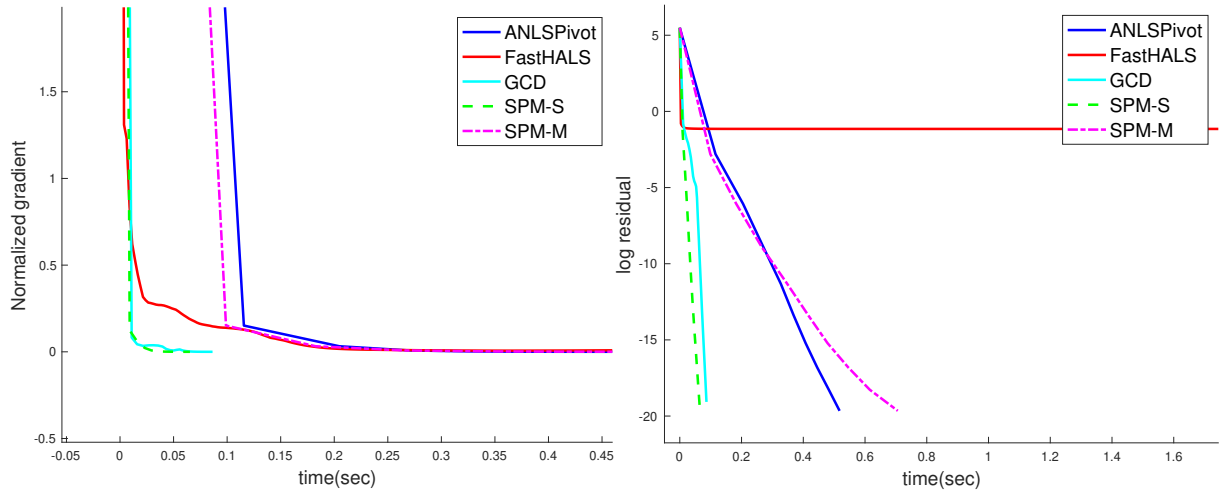
$$\min_i \left\|\mathbf{g}^i\right\|^2 \leq \frac{f(\mathbf{x}^0) - f^*}{\sigma c_1 \delta k}. \tag{68}$$

To obtain $\left\|\mathbf{g}^k\right\| \leq \epsilon$, therefore, we need $\mathcal{O}\left(1/\epsilon^2\right)$ number of iterations. $\square$

B Experimental results

In this supplemental material section, we include converges behavior plots for all datasets.

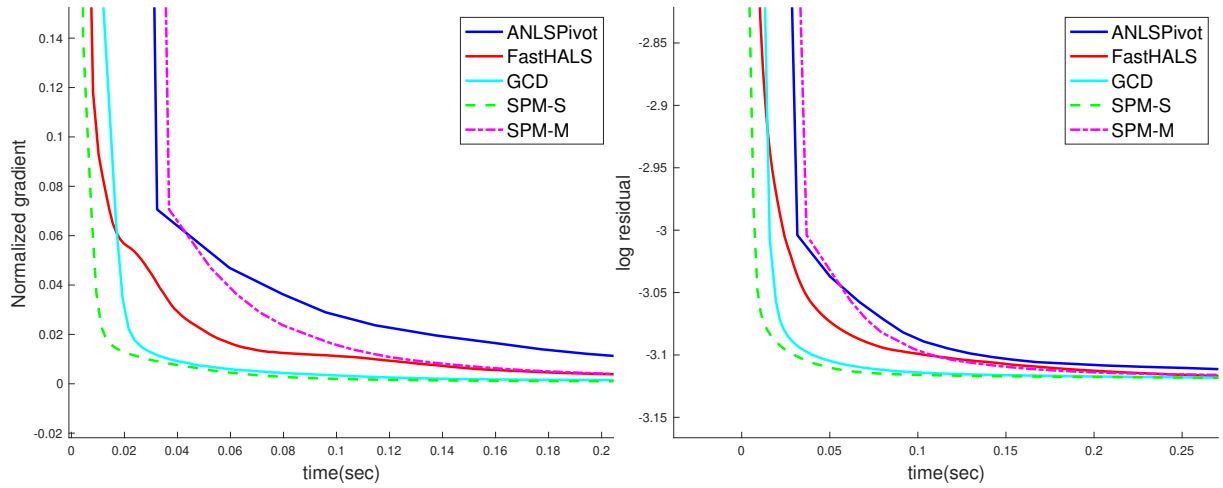
B.1 Dense datasets



(a) Optimality gap.

(b) Relative residual in log scale.

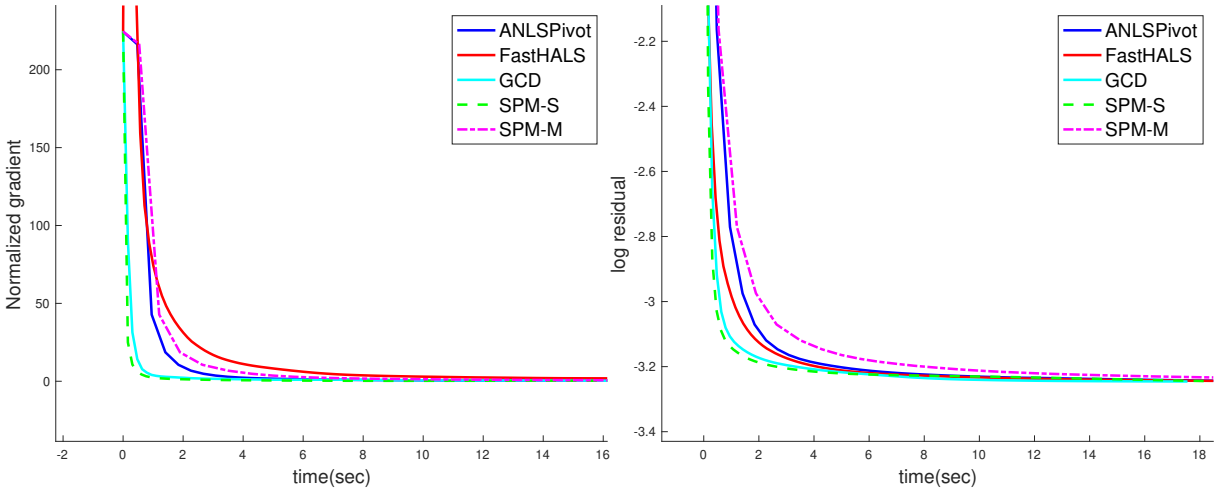
Figure 4: Convergence behaviors for synthetic dense datasets where the exact factorization exists.



(a) Optimality gap.

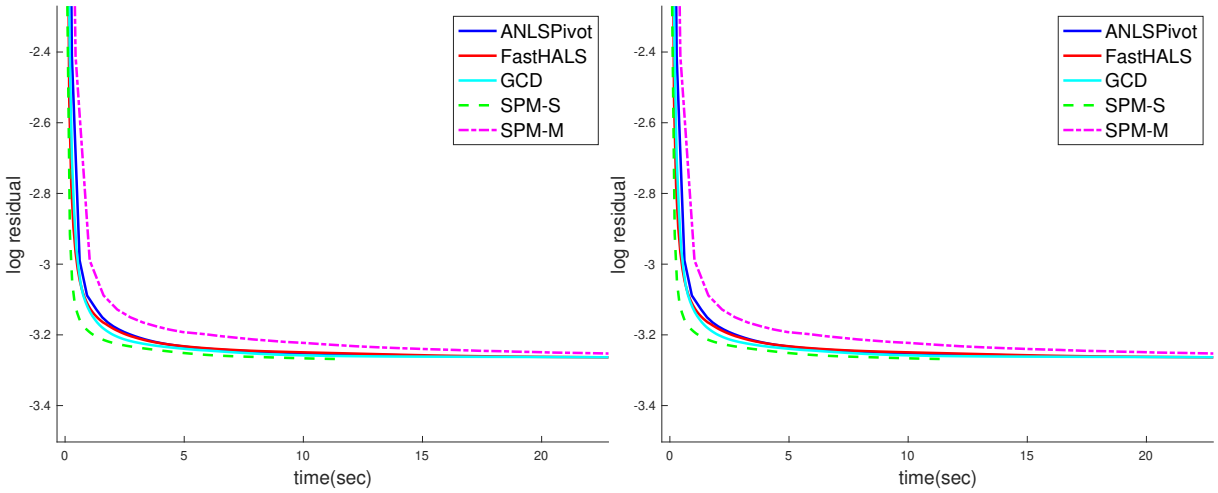
(b) Relative residual in log scale.

Figure 5: Convergence behaviors for synthetic dense datasets where the matrix cannot be factorized exactly.



(a) Optimality gap. (b) Relative residual in log scale.

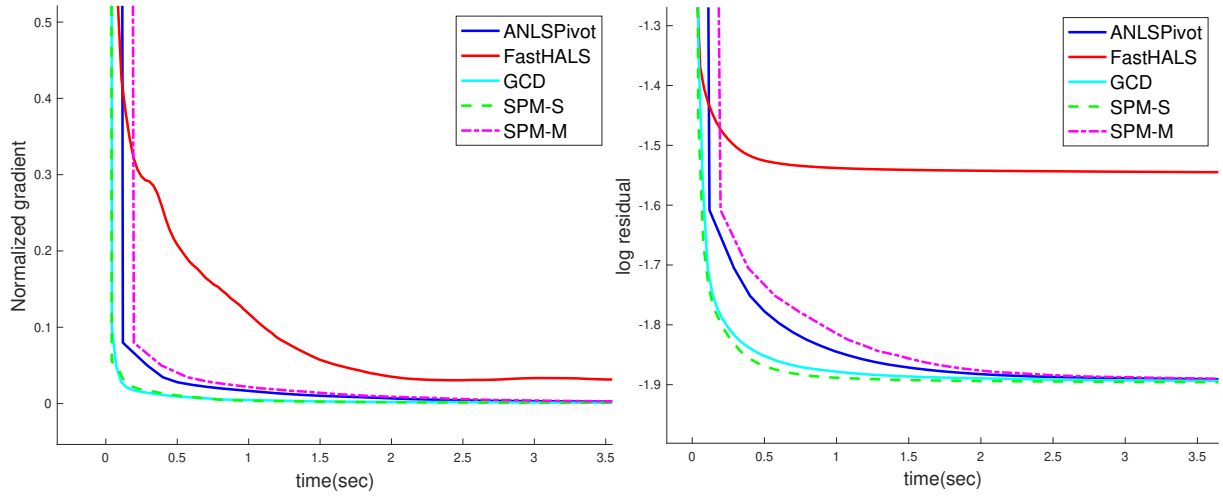
Figure 6: Convergence behaviors for real dense dataset of UMist.



(a) optimality gap. (b) Relative residual in log scale.

Figure 7: Convergence behaviors for real dense dataset of YaleB.

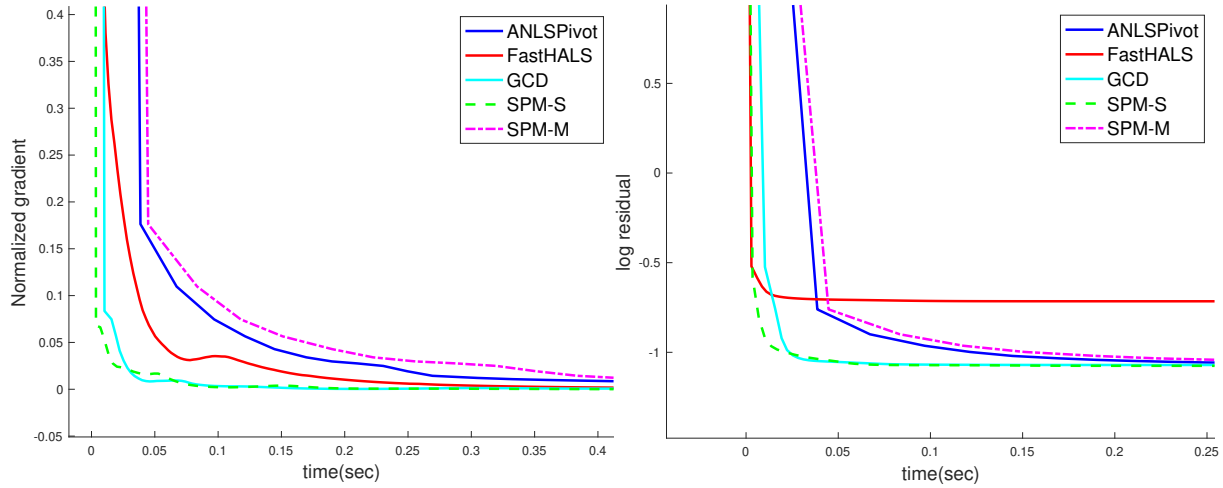
B.2 Sparse datasets



(a) optimality gap.

(b) Relative residual in log scale.

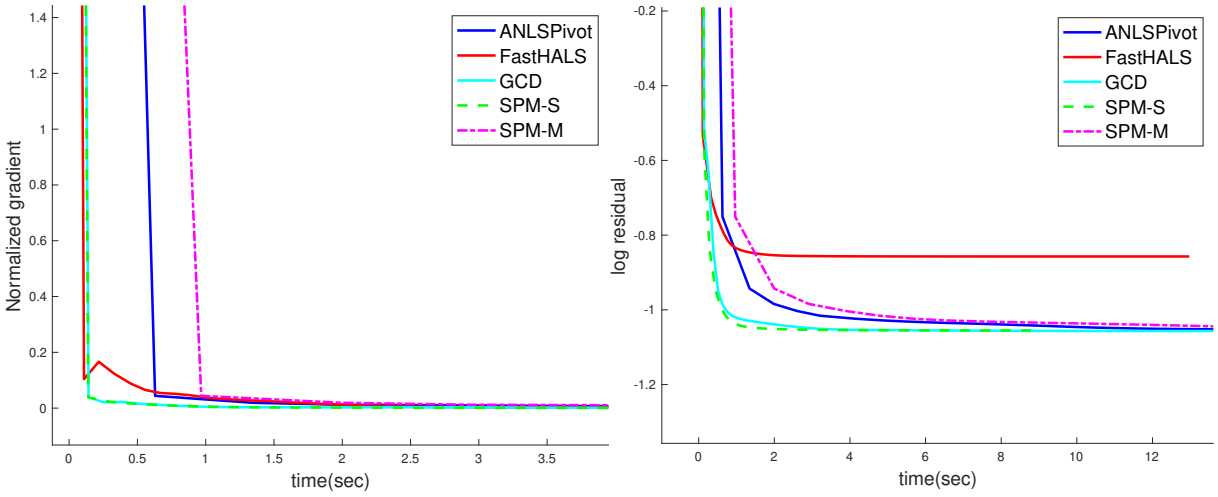
Figure 8: Convergence behaviors for synthetic sparse datasets where the exact factorization exists.



(a) Optimality gap.

(b) Relative residual in log scale.

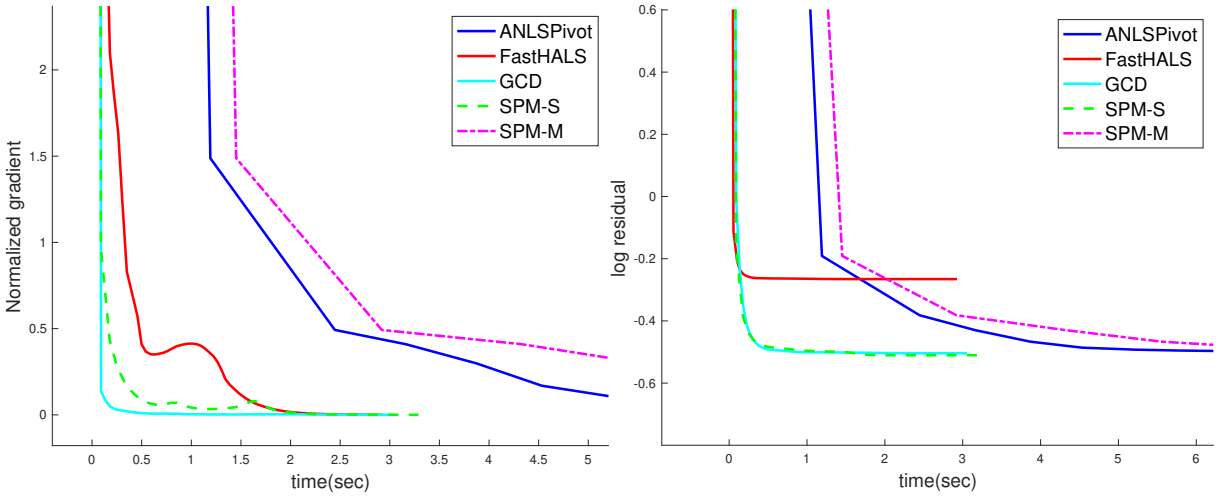
Figure 9: Convergence behaviors for synthetic sparse datasets where the matrix cannot be factorized exactly.



(a) Optimality gap.

(b) Relative residual in log scale.

Figure 10: Convergence behaviors for real sparse dataset of MNIST.



(a) Optimality gap.

(b) Relative residual in log scale.

Figure 11: Convergence behaviors for real sparse dataset of 20News.

C NNLS Experiments

As NNLS is a commonly subproblem for NMF, we compare the performance of SPM-M and BPP to solve NNLS. In addition, the results also include other state-of-art techniques to solve NNLS such as interior-point method (Interior), and trust-region method (Trust). These two methods are executed by Matlab command `quadprog` and the original author [Kim and Park, 2011] provides a Matlab implementation of BPP. The proposed algorithm SPM-M is also implemented in Matlab. We do not include the results for other methods such as active-set method, and sequential quadratic programming, because the runtimes are too long on our experiments.

The synthetic datasets are generated with the variable dimension n in the range of $[10^3, 10^4]$. The random numbers in the synthetic datasets are created by the Matlab commands `rand` and `sprand` that are uniformly distributed in the interval $[0, 1]$. Here `rand` is to generate dense data, while `sprand` is to generate sparse data using 1% sparsity. Throughout the experiments, we set $s = 1$, $\sigma = 0.1$, and $\beta = 0.5$ for the Armijo-like rule. We set the initial value in NNLS to zero and run each synthetic dataset 100 times, and record the runtime in seconds using `tic` and `toc`, then compute the mean and standard deviation. Due to the space limit and the variances are very small, we only include the mean values in Table 2.

n	Interior	Trust	BPP	SPM-M
Dense Synthetic Datasets				
$1,000$	0.3617	1.2284	0.0624	**0.0421**
$3,000$	5.3592	14.3689	0.5590	**0.4403**
$5,000$	22.4852	49.2219	2.0701	**1.6211**
$6,000$	41.8525	61.6149	3.3092	**2.4207**
$8,000$	112.6763	177.5038	7.8735	**6.4591**
$10,000$	206.8235	259.5436	15.0423	**12.2824**
Sparse Synthetic Datasets				
$1,000$	0.8931	0.0840	0.0062	**0.0035**
$3,000$	16.5547	1.5904	0.0638	**0.0532**
$5,000$	74.1925	5.6859	0.4513	**0.3985**
$6,000$	134.5539	12.0374	0.8821	**0.7490**
$8,000$	299.6275	30.5833	3.1698	**2.2887**
$10,000$	1601.5326	66.1288	8.7001	**6.0603**

Table 2: Performance comparison on NNLS in seconds.

From the results in Table 2, we can observe SPM-M is consistently faster than the other three algorithms in both dense and sparse datasets. As BPP and SPM both maintain a *working set*, we plot the results for dense dataset with $n = 5000$ to further analyze the performance. As illustrated in Figure 12, the Cholesky factorizations are the most expensive part regarding computation. We record and plot the size of working set at each iteration for SPM-M and BPP in Figure 13. BPP and SPM-M identify the optimal free set at the end, while SPM maintains a relatively *small* working set. In addition, an interesting observation can be made in Figure 14. Even maintaining a relatively larger working set, BPP does not have bigger objective reduce than SPM. Instead, SPM reduces even more objective values using relatively small working sets. However, we can also see using smaller working sets, SPM usually needs to use more iterations at the end to converges.

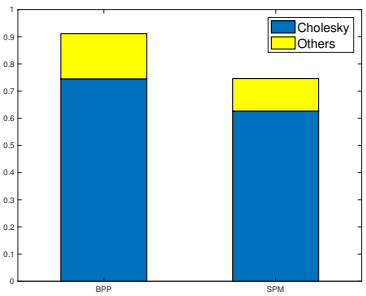


Figure 12: Runtime histogram

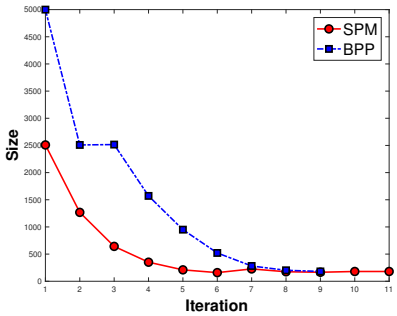


Figure 13: Size of working set

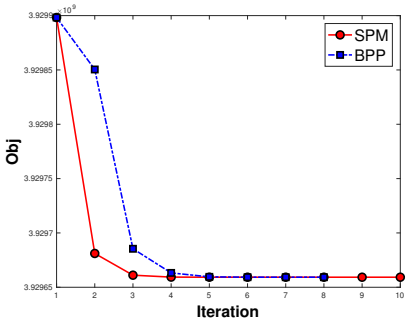


Figure 14: Objective reduction