A PROOFS

In this section, we provide detailed proofs of our theorems.

A.1 Proof of Theorem 1

Let $\Phi_k(\mathbf{x}) := \Phi(\mathbf{x}) + \rho\|\mathbf{x}-\mathbf{x}^k\|^2$ and $\Phi_k^* = \min_{\mathbf{x}} \Phi_k(\mathbf{x})$ for each $k \geq 0$. Note we have $\mathrm{dist}(\mathbf{0}, \partial\Phi_k(\mathbf{x}^{k+1})) \leq \delta = \frac{\varepsilon}{4}$, and also Φ_k is ρ-strongly convex. Hence $\Phi_k(\mathbf{x}^{k+1}) - \Phi_k^* \leq \frac{\delta^2}{2\rho}$, and $\Phi(\mathbf{x}^{k+1}) + \rho\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 - \Phi(\mathbf{x}^k) \leq \frac{\delta^2}{2\rho}$. Thus,

$$\begin{aligned}
&\Phi(\mathbf{x}^T) - \Phi(\mathbf{x}^0) + \rho\sum_{k=0}^{T-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq \frac{T\delta^2}{2\rho}\\
&T\min_{0\leq k\leq T-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq \frac{1}{\rho}\left(\frac{T\delta^2}{2\rho} + [\Phi(\mathbf{x}^0) - \Phi(\mathbf{x}^T)]\right)\\
&2\rho\min_{0\leq k\leq T-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\| \leq 2\sqrt{\frac{\delta^2}{2} + \frac{\rho[\Phi(\mathbf{x}^0)-\Phi^*]}{T}}.
\end{aligned} \tag{25}$$

Since $T \geq \frac{32\rho}{\varepsilon^2}[\Phi(\mathbf{x}^0) - \Phi^*]$ and $\delta = \frac{\varepsilon}{4}$, we have

$$\frac{\rho}{T}[\Phi(\mathbf{x}^0) - \Phi^*] \leq \frac{\varepsilon^2}{32}, \tag{26}$$

and thus (25) implies

$$2\rho\min_{0\leq k\leq T-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\| \leq \frac{\varepsilon}{2}. \tag{27}$$

Therefore, Algorithm 2 must stop within T iterations, from its stopping condition, and when it stops, the output $\mathbf{x}^S$ satisfies $2\rho\|\mathbf{x}^S - \mathbf{x}^{S-1}\| \leq \frac{\varepsilon}{2}$.

Now recall $\mathrm{dist}(0, \partial\Phi_k(\mathbf{x}^{k+1})) \leq \delta = \frac{\varepsilon}{4}$, i.e.,

$$\mathrm{dist}(0, \partial\Phi(\mathbf{x}^{k+1}) + 2\rho(\mathbf{x}^{k+1}-\mathbf{x}^k)) \leq \frac{\varepsilon}{2}, \forall k \geq 0. \tag{28}$$

The above inequality together with $2\rho\|\mathbf{x}^S - \mathbf{x}^{S-1}\| \leq \frac{\varepsilon}{2}$ gives

$$\mathrm{dist}(\mathbf{0}, \partial\Phi(\mathbf{x}^S)) \leq \varepsilon,$$

which implies that $\mathbf{x}^S$ is an ε-stationary point to (8).

Finally, we apply Lemma 1 to obtain the overal complexity and complete the proof.

A.2 Proof of Claim 1

Let X_* be the optimal solution set of

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2. \tag{29}$$

Then for any $\bar{\mathbf{x}} \in X_*$, $\mathbf{A}\bar{\mathbf{x}} - \mathbf{b} = \mathbf{0}$ by our assumption. From (Wang and Lin, 2014, Theorem 18), it follows that there is a constant $\kappa > 0$ such that

$$\|\mathbf{x} - \mathrm{Proj}_{X_*}(\mathbf{x})\| \leq \kappa\left\|\mathbf{x} - \mathrm{Proj}_X\big(\mathbf{x} - \nabla f(\mathbf{x})\big)\right\|, \forall\, \mathbf{x} \in X, \tag{30}$$

where Proj_X denotes the Euclidean projection onto X.

For any fixed $\mathbf{x} \in X$, denote $\mathbf{u} = \nabla f(\mathbf{x})$ and $\mathbf{v} = \mathrm{Proj}_X(\mathbf{x}-\mathbf{u})$. Then from the definition of the Euclidean projection, it follows that $\langle \mathbf{v}-\mathbf{x}+\mathbf{u}, \mathbf{v}-\mathbf{x}'\rangle \leq 0$, $\forall\, \mathbf{x}' \in X$. Letting $\mathbf{x}' = \mathbf{x}$, we have $\|\mathbf{v}-\mathbf{x}\|^2 \leq \langle \mathbf{u}, \mathbf{x}-\mathbf{v}\rangle$. On the other hand, for any $\mathbf{z} \in \mathcal{N}_X(\mathbf{x})$, we have from the definition of the normal cone that $\langle \mathbf{z}, \mathbf{x}-\mathbf{x}'\rangle \geq 0$, $\forall\, \mathbf{x}' \in X$. Hence, letting $\mathbf{x}' = \mathbf{v}$ gives $\langle \mathbf{z}, \mathbf{x}-\mathbf{v}\rangle \geq 0$. Therefore, we have

$$\|\mathbf{v}-\mathbf{x}\|^2 \leq \langle \mathbf{u}, \mathbf{x}-\mathbf{v}\rangle + \langle \mathbf{z}, \mathbf{x}-\mathbf{v}\rangle \leq \|\mathbf{x}-\mathbf{v}\| \cdot \|\mathbf{u}+\mathbf{z}\|,$$

which implies $\|\mathbf{v}-\mathbf{x}\| \le \|\mathbf{u}+\mathbf{z}\|$. By the definition of $\mathbf{u}$ and $\mathbf{v}$ and noticing that $\mathbf{z}$ is an arbitrary vector in $\mathcal{N}_X(\mathbf{x})$, we obtain

$$\left\|\mathbf{x}-\mathrm{Proj}_X\big(\mathbf{x}-\nabla f(\mathbf{x})\big)\right\| \le \mathrm{dist}\left(\mathbf{0}, \nabla f(\mathbf{x})+\mathcal{N}_X(\mathbf{x})\right).$$

The above inequality together with (30) gives

$$\|\mathbf{x}-\mathrm{Proj}_{X_*}(\mathbf{x})\| \le \kappa \cdot \mathrm{dist}\left(\mathbf{0}, \nabla f(\mathbf{x})+\mathcal{N}_X(\mathbf{x})\right), \ \forall\, \mathbf{x}\in X. \tag{31}$$

Now by the fact $\mathbf{A}\mathrm{Proj}_{X_*}(\mathbf{x}) = \mathbf{b}$, we have $\|\mathbf{A}\mathbf{x}-\mathbf{b}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}-\mathrm{Proj}_{X_*}(\mathbf{x})\|$. Therefore, from (31) and also noting $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})$, we obtain (16) with $v = \frac{1}{\kappa\|\mathbf{A}\|}$.

A.3 Proof of Claim 2

Without loss of generality, we assume $r = 1$ and $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$, i.e., the row vectors of $\mathbf{A}$ are orthonormal. Notice that

$$\mathcal{N}_X(\mathbf{x}) = \begin{cases} \{\mathbf{0}\}, & \text{if } \|\mathbf{x}\| < 1, \\ \{\lambda\mathbf{x} : \lambda \ge 0\}, & \text{if } \|\mathbf{x}\| = 1. \end{cases} \tag{32}$$

Hence, if $\|\mathbf{x}\| < 1$, (16) holds with $v = 1$ because $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$. In the following, we focus on the case of $\|\mathbf{x}\| = 1$.

When $\|\mathbf{x}\| = 1$, we have from (32) that

$$\mathrm{dist}\left(\mathbf{0}, \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})+\mathcal{N}_X(\mathbf{x})\right) = \min_{\lambda\ge 0} \|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})+\lambda\mathbf{x}\|. \tag{33}$$

If the minimizer of the right hand side of (33) is achieved at $\lambda = 0$, then (16) holds with $v = 1$. Otherwise, the minimizer is $\lambda = -\mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \ge 0$. With this λ, we have

$$\begin{aligned} &\left[\mathrm{dist}\left(\mathbf{0}, \mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})+\mathcal{N}_X(\mathbf{x})\right)\right]^2 \\ &= \|\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) - \mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})\mathbf{x}\|^2 \\ &= (\mathbf{A}\mathbf{x}-\mathbf{b})^\top\mathbf{A}(\mathbf{I}-\mathbf{x}\mathbf{x}^\top)\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}). \end{aligned}$$

Let

$$\begin{aligned} v_* = \min_{\mathbf{x}} \Big\{ &\lambda_{\min}\left(\mathbf{A}(\mathbf{I}-\mathbf{x}\mathbf{x}^\top)\mathbf{A}^\top\right), \\ &\text{s.t. } \ \mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \le 0, \ \|\mathbf{x}\| = 1\Big\}, \end{aligned} \tag{34}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix. Then v_* must be a finite nonnegative number. We show $v_* > 0$. Otherwise suppose $v_* = 0$, i.e., there is a $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \le 0$ and $\|\mathbf{x}\| = 1$, and also $\mathbf{A}(\mathbf{I}-\mathbf{x}\mathbf{x}^\top)\mathbf{A}^\top$ is singular. Hence, there exists a $\mathbf{y} \ne \mathbf{0}$ such that

$$\mathbf{A}(\mathbf{I}-\mathbf{x}\mathbf{x}^\top)\mathbf{A}^\top\mathbf{y} = \mathbf{0}. \tag{35}$$

By scaling, we can assume $\|\mathbf{y}\| = 1$. Let $\mathbf{z} = \mathbf{A}^\top\mathbf{y}$. Then $\|\mathbf{z}\| = 1$, and from (35), we have $\mathbf{z}^\top(\mathbf{I}-\mathbf{x}\mathbf{x}^\top)\mathbf{z} = 1-(\mathbf{z}^\top\mathbf{x})^2 = 0$. This equation implies $\mathbf{z} = \mathbf{x}$ or $\mathbf{z} = -\mathbf{x}$, because both $\mathbf{x}$ and $\mathbf{z}$ are unit vectors. Without loss of generality, we can assume $\mathbf{z} = \mathbf{x}$. Now recall $\mathbf{b} = \mathbf{A}\hat{\mathbf{x}}$ with $\|\hat{\mathbf{x}}\| < 1$ and notice

$$\begin{aligned} \mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) &= \mathbf{z}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{z}-\mathbf{b}) \\ &= \mathbf{z}^\top\mathbf{A}^\top(\mathbf{y}-\mathbf{A}\hat{\mathbf{x}}) = 1-\mathbf{z}^\top\mathbf{A}^\top\mathbf{A}\hat{\mathbf{x}} > 0, \end{aligned}$$

where the inequality follows from $\|\mathbf{A}\| = 1$, $\|\mathbf{z}\| = 1$, and $\|\hat{\mathbf{x}}\| < 1$. Hence, we have a contradiction to $\mathbf{x}^\top\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) \le 0$. Therefore, $v_* > 0$.

Putting the above discussion together, we have that (16) holds with $v = \min\{1, v_*\}$, where v_* is defined in (34). This completes the proof.

A.4 Proof of Theorem 2

First, note that $\mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k)$ is $\hat{L}_k$-smooth and $\hat{\rho}_k$-weakly convex, with $\hat{L}_k$ and $\hat{\rho}_k$ defined in (11). Then by the $\mathbf{x}$ update in Algorithm 3, the stopping conditions of Algorithms 1 and 2, and following the same proof of ε stationarity as in Theorem 1, we have

$$\operatorname{dist}(\mathbf{0}, \partial_x \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k)) \le \varepsilon, \forall k \ge 0. \tag{36}$$

Next we give a uniform upper bound of the dual variable. By (12), (13), $\mathbf{y}^0 = \mathbf{0}$, and also the setting of γ_k, we have that $\forall k \ge 0$,

$$\begin{aligned} \|\mathbf{y}^k\| &\le \sum_{t=0}^{k-1} w_t \|\mathbf{c}(\mathbf{x}^{t+1})\| \le \sum_{t=0}^{\infty} w_t \|\mathbf{c}(\mathbf{x}^{t+1})\| \\ &\le \bar{c} w_0 \|\mathbf{c}(\mathbf{x}^1)\| (\log 2)^2 = y_{\max}, \end{aligned} \tag{37}$$

where we have defined $\bar{c} = \sum_{t=0}^{\infty} \frac{1}{(t+1)^2[\log(t+2)]^2}$ and $y_{\max} = \bar{c} w_0 \|\mathbf{c}(\mathbf{x}^1)\| (\log 2)^2$.

Combining the above bound with the regularity assumption (15), we have the following feasibility bound: for all $k \ge 1$,

$$\begin{aligned} &\|\mathbf{c}(\mathbf{x}^k)\| \le \frac{1}{v\beta_{k-1}} \operatorname{dist}\left(0, \partial h(\mathbf{x}^k) + \beta_{k-1} J_c(\mathbf{x}^k)^\top \mathbf{c}(\mathbf{x}^k)\right) \\ =& \frac{1}{v\beta_{k-1}} \operatorname{dist}\big(0, \partial_x \mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1}) - \nabla g(\mathbf{x}^k) \\ &- J_c(\mathbf{x}^k)^\top \mathbf{y}^{k-1}\big) \\ \le& \frac{1}{v\beta_{k-1}} \big(\operatorname{dist}\left(0, \partial_x \mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1})\right) + \|\nabla g(\mathbf{x}^k)\| \\ &+ \|J_c(\mathbf{x}^k)\| \|\mathbf{y}^{k-1}\|\big) \\ \le& \frac{1}{v\beta_{k-1}} (\varepsilon + B_0 + B_c y_{\max}), \end{aligned} \tag{38}$$

where the third inequality follows from (36), (10a), (10c), and (37).

Now we define

$$K = \lceil \log_\sigma C_\varepsilon \rceil + 1, \text{ with } C_\varepsilon = \frac{\varepsilon + B_0 + B_c y_{\max}}{v\beta_0 \varepsilon}. \tag{39}$$

Then by (38) and the setting of β_k in Algorithm 3, we have $\|\mathbf{c}(\mathbf{x}^K)\| \le \varepsilon$. Also recalling (36), we have

$$\operatorname{dist}(\mathbf{0}, \partial f_0(\mathbf{x}^{k+1}) + J_c(\mathbf{x}^{k+1}) \left(\mathbf{y}^k + \beta_k \mathbf{c}(\mathbf{x}^{k+1})\right)) \le \varepsilon.$$

Therefore, $\mathbf{x}^K$ is an ε-KKT point of (1) with the corresponding multiplier $\mathbf{y}^{K-1} + \beta_{K-1}\mathbf{c}(\mathbf{x}^K)$, according to Definition 1.

In the rest of the proof, we bound the maximum number of iPPM iterations needed to stop Algorithm 2, and the number of APG iterations per iPPM iteration needed to stop Algorithm 1, for each iALM outer iteration.

Denote $\mathbf{x}_k^t$ as the t-th iPPM iterate within the k-th outer iteration of iALM. Then at $\mathbf{x}_k^t$, we use APG to minimize $F_k^t(\cdot) := \mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k) + \hat{\rho}_k \|\cdot - \mathbf{x}_k^t\|^2$, which is $\tilde{L}_k := (\hat{L}_k + 2\hat{\rho}_k)$-smooth and $\hat{\rho}_k$-strongly convex. Hence, by Lemma 1, at most T_k^{APG} (that is independent of t) APG iterations are required to find an $\frac{\varepsilon}{4}$ stationary point of $F_k^t(\cdot)$, where

$$T_k^{\mathrm{APG}} = \left\lceil \sqrt{\frac{\tilde{L}_k}{\hat{\rho}_k}} \log \frac{1024 \tilde{L}_k^2 (\tilde{L}_k + \hat{\rho}_k) D^2}{\varepsilon^2 \hat{\rho}_k} \right\rceil + 1, \forall k \ge 0. \tag{40}$$

In addition, recalling the definition of $\mathcal{L}_\beta$ in (2), observe that for all $k \geq 1$,

$$\begin{aligned}\mathcal{L}_{\beta_k}(\mathbf{x}^k, \mathbf{y}^k) \leq & B_0 + \frac{\varepsilon + B_0 + B_c y_{\max}}{v\beta_0} \\ & \left(y_{\max} + \frac{\sigma(\varepsilon + B_0 + B_c y_{\max})}{2v}\right)\sigma^{1-k} \\ \leq & B_0 + \tilde{c}, \forall k \geq 1,\end{aligned} \tag{41}$$

where B_0 is given in (10a) and
$\tilde{c} := \frac{\varepsilon+B_0+B_c y_{\max}}{v\beta_0}\left(y_{\max} + \frac{\sigma(\varepsilon+B_0+B_c y_{\max})}{2v}\right)$. Furthermore,

$$\mathcal{L}_{\beta_0}(\mathbf{x}^0, \mathbf{y}^0) \leq B_0 + \frac{\beta_0}{2}\|\mathbf{c}(\mathbf{x}^0)\|^2,$$

and $\forall k \geq 0, \forall \mathbf{x} \in \mathrm{dom}(h)$,

$$\mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k) \geq f_0(\mathbf{x}) + \langle \mathbf{y}^k, \mathbf{c}(\mathbf{x})\rangle \geq -B_0 - y_{\max}\bar{B}_c, \tag{42}$$

where $\bar{B}_c$ is given in (10c).

Combining all three inequalities above with Theorem 1 and $\hat{\rho}_k$-weak convexity of $\mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k)$, we conclude at most T_k^{PPM} iPPM iterations are needed to guarantee that $\mathbf{x}^{k+1}$ is an ε stationary point of $\mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k)$, with

$$T_k^{\mathrm{PPM}} = \left\lceil \frac{32(\rho_0 + y_{\max}\bar{L} + \beta_k\rho_c)(2B_0 + y_{\max}\bar{B}_c + \tilde{c})}{\varepsilon^2} \right\rceil, \forall k \geq 1 \tag{43}$$

$$T_0^{\mathrm{PPM}} = \left\lceil \frac{32\rho_0}{\varepsilon^2}(2B_0 + y_{\max}\bar{B}_c + \frac{\beta_0}{2}\|\mathbf{c}(\mathbf{x}^0)\|^2) \right\rceil. \tag{44}$$

Consequently, we have shown that at most T total APG iterations are needed to find an ε-KKT point of (1), where

$$T = \sum_{k=0}^{K-1} T_k^{\mathrm{PPM}} T_k^{\mathrm{APG}}, \tag{45}$$

with K given in (39), T_k^{APG} given in (40), and T_k^{PPM} given in (43).

The result in (45) immediately gives us the following complexity results.

By (39), we have $K = \tilde{O}(1)$ and $\beta_K = O(\varepsilon^{-1})$. Hence from (11), we have $\hat{\rho}_k = O(\beta_k), \hat{L}_k = O(\beta_k), \forall k \geq 0$. Then by (40), $T_k^{\mathrm{APG}} = \tilde{O}(1), \forall k \geq 0$, and by (43), we have $T_k^{\mathrm{PPM}} = O(\varepsilon^{-3}), \forall k \geq 0$. Therefore, in (45), $T = \tilde{O}(\varepsilon^{-3})$ for a general nonlinear $\mathbf{c}(\cdot)$.

For the special case when $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $\|\mathbf{c}(\mathbf{x})\|^2$ is convex, so we have $\rho_c = 0$. Thus by (11), $\hat{\rho}_k = O(1), \forall k \geq 0$. Hence in (43), $T_k^{\mathrm{PPM}} = O(\varepsilon^{-2}), \forall k \geq 0$, and in (40), $T_k^{\mathrm{APG}} = \tilde{O}(\varepsilon^{-\frac{1}{2}}), \forall k \geq 0$. Therefore, by (45), $T = \tilde{O}(\varepsilon^{-\frac{5}{2}})$ for an affine $\mathbf{c}(\cdot)$. This completes the proof.

A.5 Proof of Theorem 3

First, by (12), (17) and $\mathbf{y}^0 = \mathbf{0}$, we have

$$\begin{aligned}\|\mathbf{y}^k\| \leq & \sum_{t=0}^{k-1} w_t \|\mathbf{c}(\mathbf{x}^{t+1})\| = \sum_{t=0}^{k-1} M(t+1)^q := y_k \\ = & O(k^{q+1}), \forall k \geq 0.\end{aligned} \tag{46}$$

Following the first part of the proof of Theorem 2, we can easily show that at most $K = O(\log \varepsilon^{-1})$ outer iALM iterations are needed to guarantee $\mathbf{x}^K$ to be an ε-KKT point of (1). Hence, $\beta_k = O(\varepsilon^{-1}), \forall\, 0 \leq k \leq K$.

Combining the above bound on K with (46), we have

$$\begin{aligned}\|\mathbf{y}^k\| \le y_K &:= \sum_{t=0}^{K-1} M(K+1)^q = O(K^{q+1}) \\ &= O\big((\log\varepsilon^{-1})^{q+1}\big),\ \forall 1 \le k \le K.\end{aligned}$$

Hence from (11), we have $\hat{\rho}_k = O(\beta_k) = O(\varepsilon^{-1}), \hat{L}_k = O(\beta_k) = O(\varepsilon^{-1}),\ \forall\, 0 \le k \le K$.

Notice that (41) and (42) still hold with $y_{\max}$ replaced by y_k. Hence, $\forall k \le K, \forall \mathbf{x} \in \mathrm{dom}(h)$,

$$\mathcal{L}_{\beta_k}(\mathbf{x}^k, \mathbf{y}^k) - \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k) = O\left(y_k\left(1 + \frac{y_k}{\beta_k}\right)\right).$$

The above equation together with Theorem 1 gives that for any $k \le K$, at most T_k^{PPM} iPPM iterations are needed to terminate Algorithm 2 at the k-th outer iALM iteration, where

$$\begin{aligned}T_k^{\mathrm{PPM}} &= \left\lceil \frac{32\hat{\rho}_k}{\varepsilon^2}\big(\mathcal{L}_{\beta_k}(\mathbf{x}^k, \mathbf{y}^k) - \min_{\mathbf{x}} \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k)\big)\right\rceil \\ &= O\left(\frac{\hat{\rho}_k y_k\left(1 + \frac{y_k}{\beta_k}\right)}{\varepsilon^2}\right).\end{aligned}$$

Also, by Lemma 1, at most T_k^{APG} APG iterations are needed to terminate Algorithm 1, where

$$T_k^{\mathrm{APG}} = O\left(\sqrt{\frac{\hat{L}_k}{\hat{\rho}_k}}\log\varepsilon^{-1}\right), \forall k \ge 0.$$

Therefore, for all $k \le K$,

$$\begin{aligned}T_k^{\mathrm{PPM}} T_k^{\mathrm{APG}} &= O\left(\frac{\sqrt{\hat{L}_k\hat{\rho}_k}\log\varepsilon^{-1}}{\varepsilon^2} y_k\left(1 + \frac{y_k}{\beta_k}\right)\right) \\ &= O\left(\frac{y_k\log\varepsilon^{-1}}{\varepsilon^2}(\beta_k + y_k)\right) \\ &= O\left(\frac{k^{q+1}\log\varepsilon^{-1}}{\varepsilon^2}(\sigma^k + k^{q+1})\right) \\ &= O\left(\frac{K^{q+1}\log\varepsilon^{-1}}{\varepsilon^2}(\sigma^K + K^{q+1})\right) \\ &= O\left(\frac{(\log\varepsilon^{-1})^{q+2}}{\varepsilon^2}\left(\frac{1}{\varepsilon} + (\log\varepsilon^{-1})^{q+1}\right)\right) \\ &= O\left(\frac{(\log\varepsilon^{-1})^{q+2}}{\varepsilon^3}\right),\end{aligned}$$

where the second equation is from $\hat{L}_k = O(\beta_k)$ and $\hat{\rho}_k = O(\beta_k)$ for a general nonlinear $\mathbf{c}(\cdot)$, and the fifth one is obtained by $K = O(\log\varepsilon^{-1})$.

Consequently, for a general nonlinear $\mathbf{c}(\cdot)$, at most T APG iterations in total are needed to find the ε-KKT point $\mathbf{x}^K$, where

$$T = \sum_{k=0}^{K-1} T_k^{\mathrm{PPM}} T_k^{\mathrm{APG}} = O\left(K\varepsilon^{-3}(\log\varepsilon^{-1})^{q+2}\right) = \tilde{O}\left(\varepsilon^{-3}\right).$$

In the special case when $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, the term $\|\mathbf{c}(\mathbf{x})\|^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is convex, so we have $\rho_c = 0$. Hence, by

(11), $\hat{\rho}_k = O(1), \forall k \geq 0$. Then following the same arguments as above, we obtain that for any $k \leq K$,

$$T_k^{\mathrm{PPM}} T_k^{\mathrm{APG}} = O\left(\frac{\sqrt{\hat{L}_k \hat{\rho}_k}}{\varepsilon^2}(\log \varepsilon^{-1})^{q+2}\right) = O\left(\varepsilon^{-\frac{5}{2}}(\log \varepsilon^{-1})^{q+2}\right).$$

Therefore, at most T total APG iterations are needed to find the ε-KKT point $\mathbf{x}^K$, where

$$T = \sum_{k=0}^{K-1} T_k^{\mathrm{PPM}} T_k^{\mathrm{APG}} = \tilde{O}\left(\varepsilon^{-\frac{5}{2}}\right),$$

which completes the proof.

B ADDITIONAL TABLES

We provide more detailed experimental results on the LCQP and EV problems to demonstrate the empirical performance of the proposed iALM from another perspective. We compare our method with the iALM in (Sahin et al., 2019) on LCQP and EV, and the HiAPeM in (Li and Xu, 2020) on LCQP.

For each method, we report the primal residual, dual residual, running time (in seconds), and the number of gradient evaluation, shortened as `pres`, `dres`, `time`, and `#Grad`, respectively. The results for all trials are shown in Tables 4 and 5 for the LCQP problem, and in Tables 6 and 7 for the EV problem. From the results, we conclude that for both of the LCQP and EV problems, to reach the same-accurate KKT point of each tested instance, the proposed improved iALM needs significantly fewer gradient evaluations and takes far less time than all other compared methods.

Table 4: Results by the proposed improved iALM, the iALM by Sahin et al. (2019), and the HiAPeM by Li and Xu (2020) on solving a 1-weakly convex LCQP (22) of size $m = 10$ and $n = 200$.

trial	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad
	proposed improved iALM				iALM by Sahin et al. (2019)				HiAPeM with $N_0 = 10, N_1 = 2$				HiAPeM with $N_0 = 1, N_1 = 10^6$			
1	2.29e-4	8.31e-4	2.09	47468	7.06e-4	1.00e-3	15.56	1569788	3.77e-5	9.64e-4	2.61	150653	2.28e-4	7.25e-4	3.93	323020
2	1.94e-4	9.24e-4	1.00	26107	1.94e-4	1.00e-3	6.68	713807	4.02e-4	6.45e-4	2.51	154519	3.72e-4	4.83e-4	6.23	531680
3	2.23e-4	3.29e-4	1.35	33392	1.40e-4	1.00e-3	5.37	636043	7.16e-5	6.37e-4	2.06	135379	3.41e-4	9.35e-4	5.54	458308
4	6.58e-4	7.18e-4	2.21	41325	6.58e-4	1.00e-3	9.39	1048446	1.33e-4	8.29e-4	1.53	82087	3.49e-4	7.10e-4	4.67	389567
5	2.22e-4	5.43e-4	1.04	29252	1.80e-4	1.00e-3	9.56	1100625	1.46e-4	4.60e-4	3.11	216479	2.95e-4	9.21e-4	8.97	735546
6	1.75e-4	5.04e-4	1.25	34488	8.96e-4	1.00e-3	11.03	1339160	9.82e-5	7.36e-4	0.64	31099	3.35e-4	7.94e-4	3.32	272395
7	4.03e-4	5.04e-4	1.10	28636	1.98e-4	1.00e-3	7.97	927075	3.00e-4	7.38e-4	3.00	199126	3.89e-4	8.39e-4	6.69	544974
8	5.83e-4	4.58e-4	1.70	39719	8.62e-4	1.00e-3	8.77	982164	3.93e-4	7.13e-4	2.85	189818	4.62e-4	9.09e-4	4.18	338027
9	5.98e-4	3.70e-4	1.66	37379	5.98e-4	1.00e-3	5.23	560382	1.45e-4	9.63e-4	4.34	286666	2.80e-4	9.45e-4	9.78	751636
10	8.11e-4	3.07e-4	1.05	25170	8.23e-4	1.00e-3	30.75	3474626	2.45e-4	8.45e-4	4.49	278127	4.65e-4	9.30e-4	7.47	594326
avg.	4.10e-4	5.49e-4	1.44	34294	5.26e-4	1.00e-3	11.03	1235210	1.97e-4	7.53e-4	2.71	172395	3.52e-4	8.20e-4	6.08	493948

In Table 8 below, we also compare our proposed iALM with the iPPP method in (Lin et al., 2019) on one representative instance of the LCQP problem in Section 4.1. For iPPP, we tune $\beta_k = \beta_0 \cdot k$ with $\beta_0 = 10$.

Table 5: Results by the proposed improved iALM, the iALM by Sahin et al. (2019), and the HiAPeM by Li and Xu (2020) on solving a 1-weakly convex LCQP (22) of size $m = 100$ and $n = 1000$.

trial	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad	pres	dres	time	#Grad
	proposed improved iALM				iALM by Sahin et al. (2019)				HiAPeM with $N_0 = 10, N_1 = 2$				HiAPeM with $N_0 = 1, N_1 = 10^6$			
1	4.36e-4	8.65e-4	109.90	220937	5.80e-4	8.1e-3	2281.8	13098032	1.05e-4	9.96e-4	550.18	2823733	5.35e-4	8.24e-4	897.68	5228014
2	4.07e-4	7.47e-4	144.23	280500	5.90e-4	1.1e-3	1682.5	10207308	1.67e-4	9.04e-4	597.60	2879969	5.51e-4	8.05e-4	740.28	4540532
3	5.99e-4	9.70e-4	99.37	228324	8.73e-4	1.00e-3	1281.3	8587300	8.22e-4	6.92e-4	474.76	2697241	5.67e-4	9.97e-4	1314.3	6986241
4	4.59e-4	8.53e-4	179.91	311724	4.05e-4	2.1e-3	1548.6	8474538	4.10e-5	8.20e-4	747.18	3804152	5.16e-4	8.62e-4	741.43	4281876
5	6.69e-4	9.57e-4	162.06	367321	3.96e-4	1.33e-2	1802.0	12464010	1.17e-4	9.82e-4	603.44	3008964	5.16e-4	9.11e-4	667.01	3830799
6	6.85e-4	8.84e-4	104.30	200256	1.49e-4	1.6e-3	2010.8	13071595	5.16e-4	9.11e-4	667.01	3830799	5.79e-4	9.82e-4	1396.0	8174370
7	6.10e-4	9.30e-4	124.50	244074	4.56e-4	1.4e-3	1843.8	11843900	4.78e-4	7.73e-4	712.36	3658514	5.53e-4	9.25e-4	615.96	3609496
8	8.47e-4	7.40e-4	122.57	261206	4.81e-4	2.3e-3	1520.6	10298480	7.69e-4	6.36e-4	402.49	2036351	5.47e-4	9.78e-4	520.07	2681970
9	5.16e-4	8.91e-4	165.14	316827	2.08e-4	1.3e-3	2334.9	14446205	5.08e-4	4.83e-4	561.30	3268825	5.43e-4	8.26e-4	1059.6	6958198
10	3.46e-4	9.72e-4	142.67	352781	3.13e-4	1.5e-3	1519.9	9370342	8.36e-5	9.60e-4	542.09	2807758	5.54e-4	8.98e-4	1963.1	11091867
avg.	5.57e-4	8.81e-4	135.47	278395	4.45e-4	3.37e-3	1782.6	11186171	3.61e-4	8.16e-4	585.84	3081631	5.46e-4	9.01e-4	991.54	5738336

Table 6: Results by the proposed improved iALM and the iALM by Sahin et al. (2019) on solving a generalized eigenvalue problem (23) of size $n = 200$.

trial	pres	dres	time	#Obj	#Grad	pres	dres	time	#Grad
	proposed improved iALM					iALM by Sahin et al. (2019)			
1	1.39e-4	9.98e-4	1.09	46140	38245	1.39e-4	1.00e-3	2.84	233367
2	5.69e-4	9.87e-4	0.48	31456	25592	5.69e-4	1.00e-3	1.32	144750
3	2.57e-4	9.92e-4	0.60	32933	26112	2.57e-4	1.00e-3	2.21	150136
4	1.45e-4	9.98e-4	0.59	29408	25203	1.45e-4	1.00e-3	2.24	153485
5	1.52e-4	1.00e-3	0.93	37477	27434	1.51e-4	1.00e-3	1.63	153596
6	2.34e-4	9.71e-4	0.29	17765	14353	2.34e-4	1.00e-3	0.59	60643
7	9.06e-4	9.98e-4	0.42	26032	20886	9.06e-4	1.00e-3	1.05	109958
8	6.57e-4	9.97e-4	0.42	24184	19974	6.57e-4	1.00e-3	1.53	104508
9	2.44e-4	9.95e-4	0.45	27125	22390	2.44e-4	1.00e-3	1.20	126874
10	2.16e-4	9.98e-4	0.49	31238	26527	2.16e-4	1.00e-3	1.55	160941
avg.	3.52e-4	9.03e-4	0.58	30376	24672	3.52e-4	1.00e-3	1.62	139823

Table 7: Results by the proposed improved iALM and the iALM by Sahin et al. (2019) on solving a generalized eigenvalue problem (23) of size $n = 1000$.

trial	pres	dres	time	#Obj	#Grad	pres	dres	time	#Grad
	proposed improved iALM					iALM by Sahin et al. (2019)			
1	6.87e-4	9.78e-4	60.77	56805	42626	6.86e-4	2.5e-3	5671.9	9329514
2	1.39e-4	9.85e-4	63.29	80454	60765	1.38e-4	4.3e-3	8128.5	13295555
3	5.94e-4	9.92e-4	60.87	70884	49616	5.94e-4	1.00e-3	5070.0	8585272
4	4.20e-4	9.97e-4	51.08	73494	51707	4.20e-4	1.00e-3	6045.3	10008459
5	6.27e-4	9.99e-4	65.20	72763	52095	6.27e-4	1.6e-3	6733.4	10820619
6	2.92e-4	9.82e-4	36.16	41402	32164	2.90e-4	3.1e-3	3936.9	6588034
7	3.35e-4	9.95e-4	87.89	104069	74808	3.35e-4	2.1e-3	9183.8	15689148
8	4.47e-4	9.91e-4	51.12	60555	45578	4.46e-4	2.6e-3	5300.0	9039022
9	4.02e-4	9.91e-4	44.23	51399	39064	4.01e-4	2.6e-3	4771.7	8466906
10	9.32e-4	9.95e-4	79.42	98130	69322	9.32e-4	1.6e-3	8846.8	14688990
avg.	4.88e-4	9.91e-4	60.00	70996	51775	4.87e-4	2.24e-3	5975.1	10651152

Table 8: Results by the proposed improved iALM and the iPPP by Lin et al. (2019) on solving an LCQP problem (23) of size $m = 100$ and $n = 1000$.

method	pres	dres	time	#Grad
proposed iALM	4.08e-4	7.47e-4	293.2	280500
iPPP in (Lin et al., 2019)	9.98e-4	9.98e-4	930.7	1644496

C BETTER SUBROUTINE BY INEXACT PROXIMAL POINT METHOD

We mentioned at the end of Section 2 that our iPPM is more stable and more efficient on solving nonconvex subproblems in the form of (7) than the subroutine by Sahin et al. (2019). An intuitive explanation is as follows. The iPPM tackles the nonconvex problem by solving a sequence of perturbed strongly convex problems, which can be solved by Nesterov's accelerated first-order method. In contrast, the subroutine of the iALM by Sahin et al. (2019) applies Nesterov's acceleration technique directly while performing proximal gradient update to solve the nonconvex problem. We believe such a combination of acceleration with nonconvexity attributes to the instability or inefficiency of the iALM by Sahin et al. (2019).

In this section, we provide numerical results to support the claim above. In Figure 2 below, we plot representative trajectories of the violation of stationarity for the first subproblem in all of our experiments (namely, LCQP, EV and clustering problems) using our iPPM and the subsolver by Sahin et al. (2019) started from the same initial points, where the violation of stationarity is measured as $\mathrm{dist}\big(\mathbf{0}, \partial F(\mathbf{x})\big)$. From the figure, we can clearly observe that our iPPM method is more efficient than the subsolver by Sahin et al. (2019).

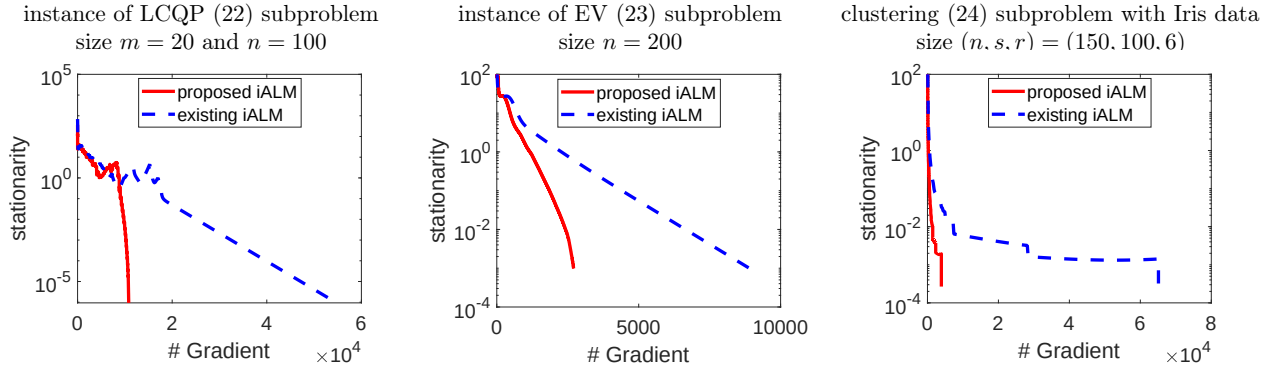

Figure 2: Comparison of iPPM and the subsolver of an existing iALM in (Sahin et al., 2019) on solving the first subproblem of LCQP, EV, and clustering problems. Each plot shows the violation of stationarity.