Supplementary Materials

A. Proof of Lemma 3

Note that the evolution $V_2^{k+1} - V_2^k$ can be structurally decomposed into two constituent terms:

$$V_2^{k+1}-V_2^k=C_3+\frac{1}{m}(g(\mathbf{1}\bar{x}^k,z^{k+1})-g(\mathbf{1}\bar{x}^k,z^k)), \quad \text{(A.1)}$$

where $C_3 = \frac{1}{m}g(\mathbf{1}\bar{x}^{k+1}, z^{k+1}) - \frac{1}{m}g(\mathbf{1}\bar{x}^{k+1}, \theta^*(\mathbf{1}\bar{x}^{k+1})) - \frac{1}{m}(g(\mathbf{1}\bar{x}^k, z^{k+1}) - g(\mathbf{1}\bar{x}^k, \theta^*(\mathbf{1}\bar{x}^k)))$. Thus, we begin by bounding the last term in (A.1) using the smoothness of g_i and the recursion (2) as follows:

$$\begin{aligned}
&\frac{1}{m}g\left(\mathbf{1}\bar{x}^k, z^{k+1}\right) - \frac{1}{m}g\left(\mathbf{1}\bar{x}^k, z^k\right) \quad \text{(A.2)}\\
\leqslant& -\frac{\gamma}{m}\left\langle \nabla_\theta g\left(\mathbf{1}\bar{x}^k, z^k\right), \nabla_\theta g\left(x^k, z^k\right)\right\rangle + \frac{L_{g,1}}{2m}\|z^{k+1}-z^k\|^2\\
\leqslant& \frac{L_{g,1}^2\gamma}{2}\frac{1}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2 - \frac{\gamma}{2}\frac{1}{m}\|\nabla_\theta g(\mathbf{1}\bar{x}^k, z^k)\|^2,
\end{aligned}$$

where the last inequality is derived through application of the polarization identity $-a^{\mathrm{T}}b = \frac{1}{2}\|a-b\|^2 - \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$, combined with the gradient-Lipschitz continuity of g_i and the condition that $\gamma \leqslant \frac{1}{L_{g,1}}$. For the term C_3 in (A.1), it follows from the smoothness of the functions g_i and $g_i^*(\hat{x})$ that

$$\begin{aligned}
&C_3 \quad \text{(A.3)}\\
\leqslant& \frac{1}{m}\left\langle \nabla_x g(\mathbf{1}\bar{x}^k, z^{k+1}) - \nabla_x g(\mathbf{1}\bar{x}^k, \theta^*(\mathbf{1}\bar{x}^k)), \mathbf{1}\bar{x}^{k+1} - \mathbf{1}\bar{x}^k\right\rangle\\
&+ \frac{L_{g,1} + L_{g^*}}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2\\
\leqslant& 16\alpha L_{g,1}^2\frac{1}{m}\|z^{k+1} - \theta^*(\mathbf{1}\bar{x}^k)\|^2 + \frac{\alpha}{16}\|\bar{x}^{k+1} - \bar{x}^k\|^2\\
\leqslant& \frac{16\alpha L_{g,1}^2}{m}\|z^{k+1} - \theta^*(\mathbf{1}\bar{x}^k)\|^2 + \frac{\alpha\lambda^2 L_{g,1}^2}{4m}\|z^{k+1} - \theta^*(\mathbf{1}x^k)\|^2\\
&+ \frac{\alpha L_{p\theta,\lambda}^2}{4m}\|\mathbf{1}\bar{x}^k - x^k\|^2 + \frac{\alpha}{4}\|\nabla_x p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k)\|^2,
\end{aligned}$$

where the second step follows form Cauchy-Schwart inequality and the condition that $\alpha \leqslant \frac{1}{16(L_{g,1} + L_{g^*})}$. In what follows, our analysis focuses on bounding the inner-level optimality gap $\|z^{k+1} - \theta^*(\mathbf{1}\bar{x}^k)\|^2$. To this end, we let $g^*(x^k) \triangleq \sum_{i=1}^m g_i^*(x_i^k)$. Then, considering the condition $\gamma < \min\{\frac{1}{L_{g,1}}, \frac{1}{\mu_g}\}$, we can derive from the smoothness of g_i, the PL-condition in θ and the recursion (2) that the following the exponential convergence property holds [25]:

$$g(x^k, z^{k+1}) - g^*(x^k) \leqslant (1-\mu\gamma)^N(g(x^k, z^k) - g^*(x^k)). \quad \text{(A.4)}$$

Then, we have that

$$\begin{aligned}
&\|z^{k+1} - \theta^*(\bar{x}^k)\|^2 \quad \text{(A.5)}\\
\leqslant& \frac{1}{2\mu_g}(1-\gamma\mu_g)^N(g(\mathbf{1}\bar{x}^k, z^k) - g(\mathbf{1}\bar{x}^k, \theta^*(\mathbf{1}\bar{x}^k)))\\
\leqslant& \frac{(1-\gamma\mu_g)^N}{4\mu_g^2}\|\nabla_\theta g(\mathbf{1}\bar{x}^k, z^k)\|^2,
\end{aligned}$$

where the first inequality harnesses the the quadratic growth property induced by the PL condition of g_i [25] and the inequality (A.4) with the condition that $\gamma \leqslant \frac{1}{\mu_g}$. Then, substituting the boundedness (A.2) and (A.3) into (A.1) and synthesizing the inequality (A.5) yields the derived result. This completes the proof. ■

B. Proof of Lemma 4

This proof proceeds through tri-variate descent analysis of the function $p_\lambda(\hat{x}, \theta, \hat{s})$. First, leveraging the smoothness of the function $p_\lambda(\hat{x}, \theta, \hat{s})$ in $\hat{x}$ from Lemma (1) and incorporating the fact that $\bar{y}^k = \bar{q}^k + c(\bar{x}^k - \bar{s}^k) = J_n\nabla_x P_\lambda(\mathbf{1}\bar{x}^k, \theta^{k+1}, z^{k+1}; \mathbf{1}\bar{s}^k)$ derived by the gradient tracking step (5), we obtain that:

$$\begin{aligned}
&p_\lambda(\bar{x}^{k+1}, \theta^{k+1}; \bar{s}^k) - p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k) \quad \text{(A.6)}\\
\leqslant& -\alpha\left\langle \nabla_x p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k), \bar{y}^k\right\rangle + \frac{L_{ps,\lambda}}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2\\
\leqslant& 2\alpha L_{p\theta,\lambda}^2\frac{1}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2 + \frac{\alpha\lambda^2 L_{g,1}^2}{4\mu_g^2}\frac{1}{m}\|\nabla_\theta g(\mathbf{1}\bar{x}^k, z^k)\|^2\\
&- \frac{1}{2}\alpha\|\nabla_x p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k)\|^2,
\end{aligned}$$

where the last step employs the polarization identity and the condition that $\alpha \leqslant \frac{1}{L_{ps,\lambda}}$ as well as the following boundedness:

$$\begin{aligned}
&\|\nabla_x p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k) - \bar{y}^k\|^2 \quad \text{(A.7)}\\
\leqslant& 4L_{p\theta,\lambda}^2\frac{1}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2 + 2\lambda^2 L_{g,1}^2\frac{1}{m}\|z^{k+1} - \theta^*(\mathbf{1}\bar{x}^k)\|^2,\\
\leqslant& \frac{4L_{p\theta,\lambda}^2}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2 + \frac{2\lambda^2 L_{g,1}^2(1-\mu\gamma)^N}{4\mu_g^2 m}\|\nabla_\theta g(\mathbf{1}\bar{x}^k, z^k)\|^2.
\end{aligned}$$

where the first step employs the gradient-Lipschitz continuity of f_i and g_i, and the second step follows from the result (A.5). Similarly, using the smoothness of the function $p_\lambda(\hat{x}, \theta, \hat{s})$ in $\hat{x}$ from Lemma 1 and the fact that $\theta^{k+1} - \theta^k = -\beta\nabla_\theta P_\lambda\left(x^k, \theta^k; \mathbf{1}\bar{s}^k\right)$ from the recursion (3), we obtain that:

$$\begin{aligned}
&p_\lambda(\bar{x}^k, \theta^{k+1}; \bar{s}^k) - p_\lambda(\bar{x}^k, \theta^k; \bar{s}^k) \quad \text{(A.8)}\\
\leqslant& -\beta\left\langle \nabla_\theta p_\lambda(\bar{x}^k, \theta^k; \bar{s}^k), \nabla_\theta P_\lambda(x^k, \theta^k; \mathbf{1}\bar{s}^k)\right\rangle\\
&+ \frac{L_{p\theta,\lambda}}{2m}\|\theta^{k+1} - \theta^k\|^2\\
\leqslant& \frac{1}{2}\beta L_{p\theta,\lambda}^2\frac{1}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2 - \frac{1}{2}\beta m\|\nabla_\theta p_\lambda(\bar{x}^k, \theta^k; \bar{s}^k)\|^2,
\end{aligned}$$

where the last step utilizes the polarization identity and applies the condition $\alpha \leqslant \frac{1}{L_{p\theta,\lambda}}$ along with the expression $\nabla_\theta p_\lambda(\bar{x}^k, \theta^k; \bar{s}^k) = \frac{1}{m}\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k, \theta^k; \mathbf{1}\bar{s}^k)$. Additionally, it follows from the definition of the function $p_\lambda(\hat{x}, \theta, \hat{s})$ that:

$$p_\lambda(\bar{x}^{k+1}, \theta^{k+1}; \bar{s}^{k+1}) - p_\lambda(\bar{x}^{k+1}, \theta^{k+1}; \bar{s}^k)$$

$$=\frac{c}{2}\frac{1}{m}\sum_{i=1}^{m}(\|\bar{x}^{k+1}-\bar{s}^{k+1}\|^2-\|\bar{x}^{k+1}-\bar{s}^{k}\|^2)$$
$$=\frac{c}{2}(\frac{1-\eta}{\eta^2}\|\bar{s}^{k+1}-\bar{s}^{k}\|^2-\frac{1}{\eta^2}\|\bar{s}^{k+1}-\bar{s}^{k}\|^2)$$
$$=-\frac{c}{2\eta}\|\bar{s}^{k+1}-\bar{s}^{k}\|^2, \tag{A.9}$$

where the first step applies the recursion (7). Subsequently, by aggregating the descent inequalities (A.6), (A.8), and (A.9), we obtain the desired result. This completes the proof. ■

C. Proof of Lemma 5

By the fact that $\psi_\lambda(\theta^{k+1};\bar{s}^{k+1}) = p_\lambda(x^*(\theta^{k+1},\bar{s}^{k+1}),\theta^{k+1};\bar{s}^{k+1})$ as well as the expression of the function $p_\lambda(\hat{x},\theta,\hat{s})$, we can derive that

$$\psi_\lambda(\theta^{k+1};\bar{s}^{k+1})-\psi_\lambda(\theta^{k+1};\bar{s}^{k}) \tag{A.10}$$
$$\leqslant p_\lambda(x^*(\theta^{k+1},\bar{s}^{k}),\theta^{k+1};\bar{s}^{k+1})-p_\lambda(x^*(\theta^{k+1},\bar{s}^{k}),\theta^{k+1};\bar{s}^{k})$$
$$=\frac{c}{2}\left\langle \bar{s}^{k+1}-\bar{s}^{k},\bar{s}^{k+1}-\bar{s}^{k}-2x^*(\theta^{k+1},\bar{s}^{k})\right\rangle.$$

Furthermore, by synthesizing the smoothness property of the value function $\psi_\lambda(\theta;\hat{s})$ in θ, as established in Lemma 1, with the recursion $\theta^{k+1}-\theta^k=\beta\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)$ in (3), we derive the following critical descent relationship:

$$\psi_\lambda(\theta^{k+1};\bar{s}^{k})-\psi_\lambda(\theta^{k};\bar{s}^{k}) \tag{A.11}$$
$$\leqslant\left\langle\nabla_\theta\psi_\lambda(\theta^k;\bar{s}^k),\theta^{k+1}-\theta^k\right\rangle+\frac{L_{\psi_\lambda}}{2m}\|\theta^{k+1}-\theta^k\|^2\triangleq C_1+C_2,$$

where $C_1\triangleq-\beta\left\langle\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k),\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\right\rangle$ and $C_2\triangleq\frac{L_{\psi_\lambda}}{2m}\beta^2\|\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2$. Noting that

$$-\beta\left\langle\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k),\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\right\rangle \tag{A.12}$$
$$\leqslant\frac{L_{p\theta,\lambda}^2\beta}{2m}\|\mathbf{1}x^k-x^k\|^2-\frac{\beta}{2m}\|\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2,$$

then it follows from the term C_1 that:

$$C_1\leqslant-\beta\left\langle\mathcal{A}_1,\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\right\rangle+\frac{\beta}{2}L_{p\theta,\lambda}^2\frac{1}{m}\|\mathbf{1}x^k-x^k\|^2$$
$$-\frac{\beta}{2}\frac{1}{m}\|\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2, \tag{A.13}$$

where $\mathcal{A}_1=\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)-\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)$. For the first term on the right-hand side of (A.13), we obtain the following bound

$$-\beta\left\langle\mathcal{A}_1,\nabla_\theta P_\lambda(x^k,\theta^k;1_m\bar{s}^k)\right\rangle \tag{A.14}$$
$$\leqslant\frac{L_{p\theta,\lambda}}{w}\|x^*(\theta^k,\bar{s}^k)-\bar{x}^k\|^2+\frac{\beta^2L_{p\theta,\lambda}wL_{p\theta,\lambda}^2}{2m}\left\|\mathbf{1}\bar{x}^k-x^k\right\|^2$$
$$+\frac{\beta^2}{2}L_{p\theta,\lambda}wm\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\|^2$$
$$\leqslant\frac{\beta^2}{2}L_{p\theta,\lambda}(w+\frac{3\sigma_2^2}{w})m\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\|^2$$
$$+\frac{3L_{p\theta,\lambda}\sigma_3^2}{w}\|\nabla_x p_\lambda(\bar{x}^k,\theta^{k+1};\bar{s}^k)\|^2$$
$$+\frac{\beta^2}{2}L_{p\theta,\lambda}^3(w+\frac{6\sigma_2^2}{w})\frac{1}{m}\|\mathbf{1}\bar{x}^k-x^k\|^2,$$

where the first step applies the Cauchy-Schwarz inequality and exploits the gradient-Lipschitz continuity of $p_\lambda(\hat{x},\theta,\hat{s})$ in the inner-level variables, followed by Young's inequality with the parameter $w=\frac{\beta L_{p\theta,\lambda}}{8}$; subsequently, we introduce the term $\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k,\theta^k;\mathbf{1}\bar{s}^k)$ and leverage its Lipschitz continuity in the outer-level variables in the first step; the last step leverages the boundedness of the term $\|x^*(\theta^k,\bar{s}^k)-\bar{x}^k\|$ given by:

$$\|x^*\left(\theta^k,\bar{s}^k\right)-\bar{x}^k\| \tag{A.15}$$
$$\leqslant\|x^*(\theta^{k+1},\bar{s}^k)-\bar{x}^k\|+\|x^*(\theta^k,\bar{s}^k)-x^*(\theta^{k+1},\bar{s}^k)\|$$
$$\leqslant\sigma_3\|\nabla_x p_\lambda(\bar{x}^k,\theta^{k+1};\bar{s}^k)\|+\sigma_2\beta\sqrt{m}\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\|$$
$$+\sigma_2\beta L_{p\theta,\lambda}\frac{1}{\sqrt{m}}\|\mathbf{1}x^k-x^k\|,$$

where the last inequality follows from the strong convexity of $p_\lambda(\hat{x},\theta,\hat{s})$ in $\hat{x}$ with the parameter $\sigma_3=\frac{1}{c-L_{px,\lambda}}$ and the Lipschitz continuity of $x^*(\theta,\hat{s})$ in θ. In what follows, we aim to bound the term C_2:

$$C_2\leqslant L_{\psi_\lambda}\beta^2\frac{1}{m}\|\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2 \tag{A.16}$$
$$+L_{\psi_\lambda}L_{p\theta,\lambda}^2\beta^2\frac{1}{m}\|\mathbf{1}\bar{x}^k-\bar{x}^k\|^2.$$

Then, by substituting the inequalities (A.13), (A.14), (A.16) and the expression $\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)=\frac{1}{m}\nabla_\theta P_\lambda(\mathbf{1}\bar{x}^k,\theta^k;\mathbf{1}\bar{s}^k)$ into (A.11) and combining the results (A.10) and (A.11), the desired result can be derived. This completes the proof. ■

D. Proof of Lemma 6

We begin by employing the definition of the value function $\phi_\lambda(\hat{s})$ in (8):

$$\varphi_\lambda(\bar{s}^k)-\varphi_\lambda(\bar{s}^{k+1})$$
$$\leqslant\psi_\lambda(\theta_\lambda^*(\bar{s}^{k+1});\bar{s}^k)-\psi_\lambda(\theta_\lambda^*(\bar{s}^{k+1});\bar{s}^{k+1})$$
$$\leqslant p_\lambda(x^*(\theta_\lambda^*(\bar{s}^{k+1})\bar{s}^{k+1}),\theta^*(\bar{s}^{k+1});\bar{s}^k) \tag{A.17}$$
$$-p_\lambda(x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1}),\theta_\lambda^*(\bar{s}^{k+1});\bar{s}^{k+1})$$
$$\leqslant-\frac{c}{2}\left\langle\bar{s}^{k+1}-\bar{s}^k,\bar{s}^{k+1}-\bar{s}^k-2x^*(\theta^{k+1},\bar{s}^k)\right\rangle+C_3,$$

where $C_3\triangleq c\left\langle\bar{s}^{k+1}-\bar{s}^k,x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1})-x^*(\theta^{k+1},\bar{s}^k)\right\rangle$; the first inequality follows from the optimality of $\theta_\lambda(\bar{s}^k)$, while the second inequality is derived from the definition of the value function $\psi_\lambda(\theta_\lambda^*(\bar{s}^{k+1});\bar{s}^k)$ in section III-A and the fact that $\psi_\lambda(\theta_\lambda^*(\bar{s}^{k+1});\bar{s}^k)\leqslant p_\lambda(x^*(\theta_\lambda^*(\bar{s}^{k+1})\bar{s}^{k+1}),\theta^*(\bar{s}^{k+1});\bar{s}^k)$. Next, we proceed with a further analysis of the term C_3 in (A.17):

$$C_3=c\left\langle\bar{s}^{k+1}-\bar{s}^k,x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1})-x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^k)\right\rangle$$
$$+c\left\langle\bar{s}^{k+1}-\bar{s}^k,x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^k)-x^*(\theta^{k+1},\bar{s}^k)\right\rangle$$
$$\leqslant(c\sigma_1+\frac{c}{12\eta})\|\bar{s}^{k+1}-\bar{s}^k\|^2 \tag{A.18}$$
$$+3c\eta\|x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^k)-x^*(\theta^{k+1},\bar{s}^k)\|^2,$$

where we employ the Lipschitz continuity of $x^*(\theta,\hat{s})$ in $\hat{s}$ along with Young's inequality. In what follows, we aim to bound the last term in (A.18).

To this end, we introduce an auxiliary variable $\theta_\lambda^+(\bar{s}^k) \triangleq \theta^k - \beta\nabla_\theta P_\lambda(\mathbf{1}x^*(\theta^k,\bar{s}^k),\theta^k;\mathbf{1}\bar{s}^k)$. Then, we obtain the following result:

$$\begin{aligned}
&\|x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^k) - x^*(\theta^{k+1},\bar{s}^k)\|^2 && \text{(A.19)}\\
\leqslant& C_4 + \underbrace{4\|x^*(\theta_\lambda^*(\bar{s}^k),\bar{s}^k) - x^*(\theta_\lambda^+\left(\bar{s}^k\right),\bar{s}^k)\|^2}_{C_5} && \text{(A.20)}\\
&+ \underbrace{4\|x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k) - x^*(\theta^{k+1},\bar{s}^k)\|^2}_{C_6},
\end{aligned}$$

where $C_4 = 4\|x^*(\theta_\lambda^*\left(\bar{s}^{k+1}\right),\bar{s}^k) - x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1})\|^2 + 4\|x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1}) - x^*(\theta_\lambda^*(\bar{s}^k),\bar{s}^k)\|^2$. For the term C_4 in (A.19), it follows from the Lipschitz continuity of $x^*(\theta,\hat{s})$ and $x^*(\hat{s})$ in $\hat{s}$ that

$$\begin{aligned}
&4\|x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^k) - x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1})\|^2\\
&+ 4\|x^*(\theta_\lambda^*(\bar{s}^{k+1}),\bar{s}^{k+1}) - x^*(\theta_\lambda^*(\bar{s}^k),\bar{s}^k)\|^2\\
\leqslant& 8\sigma_1^2\|\bar{s}^{k+1} - \bar{s}^k\|^2. && \text{(A.21)}
\end{aligned}$$

In addition, the term C_5 can be bounded by:

$$\begin{aligned}
&C_5\\
\leqslant& 8\sigma_2^2 L_{p\theta,\lambda}^2\beta^2\|x^*(\theta^k,\bar{s}^k)-\bar{x}^k\|^2+8\sigma_2^2L_{p\theta,\lambda}^2\beta^2\frac{1}{m}\|\mathbf{1}\bar{x}^k-x^k\|^2\\
\leqslant& 24\sigma_2^2\sigma_3^2L_{p\theta,\lambda}^2\beta^2\|\nabla_x p_\lambda(\bar{x}^k,\theta^{k+1};\bar{s}^k)\|^2 && \text{(A.22)}\\
&+24\sigma_2^4L_{p\theta,\lambda}^2\beta^4 m\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\|^2\\
&+ 24\sigma_2^4L_{p\theta,\lambda}^4\beta^4\frac{1}{m}\|\mathbf{1}\bar{x}^k-x^k\|^2+8\sigma_2^2L_{p\theta,\lambda}^2\beta^2\frac{1}{m}\|\mathbf{1}\bar{x}^k - x^k\|^2,
\end{aligned}$$

where the first step follows from the Lipschitz continuity of $x^*(\theta,\hat{s})$ with respect to θ, while leveraging the definition of $\theta_\lambda^+(\bar{s}^k)$ and the recursion (3); the second step incorporates the recursion (A.15). Next, for the term C_6, we can further derive that:

$$\begin{aligned}
&C_6\\
\leqslant& 8\sigma_3(\varphi_\lambda(x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k);\bar{s}^k) - \varphi_\lambda(x^*(\theta_\lambda^*(\bar{s}^k),\bar{s}^k);\bar{s}^k))\\
\leqslant& 8\sigma_3(p_\lambda(x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k),\theta_\lambda^+(\bar{s}^k);\bar{s}^k)-\varphi_\lambda(x^*(\theta_\lambda^*(\bar{s}^k),\bar{s}^k);\bar{s}^k))\\
\leqslant& \frac{8\sigma_3}{\mu_p}m\|\nabla_\theta p_\lambda(x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k),\theta_\lambda^+(\bar{s}^k);\bar{s}^k)\|^2\\
\leqslant& \frac{8\sigma_3}{\mu_p}\tau^2 m\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|^2,
\end{aligned}$$

where $\tau \triangleq 1 + \sigma_2 L_{p\theta,\lambda}\beta + L_{p\theta,\lambda}\beta$; the first step follows from the strong convexity of $p_\lambda(\hat{x},\theta,\hat{s})$ in $\hat{x}$; the third step is derived by the PL-condition of $p_\lambda(\hat{x},\theta,\hat{s})$ in θ; the last step is obtained by the following result:

$$\begin{aligned}
&\|\nabla_\theta p_\lambda(x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k),\theta_\lambda^+(\bar{s}^k);\bar{s}^k)\|\\
\leqslant& \|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\| + L_{p\theta,\lambda}\frac{1}{m}\|\theta_\lambda^+(\bar{s}^k) - \theta^k\|\\
&+ L_{p\theta,\lambda}\|x^*(\theta_\lambda^+(\bar{s}^k),\bar{s}^k) - x^*(\theta^k,\bar{s}^k)\|\\
\leqslant& \tau\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|,
\end{aligned}$$

where the first step introduces the term $\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)$ and leverages the gradient Lipschitz continuity of $p_\lambda(\hat{x},\theta,\hat{s})$ in θ, and the second step exploits the Lipschitz continuity of $x^*(\hat{x},\theta)$ in θ along with the definition of $\theta_\lambda^+\left(\bar{s}^k\right)$. Next, by substituting the boundedness of the terms C_4, C_5, and C_6, incorporating the inequality (A.19) into (A.18), and combining the resulting inequality with (A.17), we obtain the desired result. This completes the proof. ■

E. Proof of Lemma 7

We begin the proof by deriving the relationship between the consensus errors of the outer-level variables and the proximal variables. First, by incorporating the recursion of the variable s^k in (7), we obatin $\mathbf{1}\bar{s}^{k+1} - s^{k+1} = (\mathcal{W} - \mathcal{J}_n)(\mathbf{1}\bar{s}^k - s^k) + \eta(\mathbf{1}\bar{x}^{k+1} - x^{k+1} - (\mathbf{1}\bar{s}^k - s^k))$ with $\mathcal{W} = W\otimes I_n$. Applying Young's inequality with parameter $\frac{1-\rho}{2\rho}$ and the fact $\|\mathcal{W} - \mathcal{J}_n\|^2 = \|W - \frac{1_m^T 1_m}{m}\|^2 = \rho \in [0,1)$, this further yields the following evolution of the consensus errors:

$$\begin{aligned}
&\|\mathbf{1}\bar{s}^{k+1} - s^{k+1}\|^2\\
\leqslant& \frac{1+\rho}{2}\|\mathbf{1}\bar{s}^k - s^k\|^2 + \frac{4\eta^2}{1-\rho}\|\mathbf{1}\bar{s}^k - s^k\|^2\\
&+ \frac{4\eta^2}{1-\rho}\|\mathbf{1}\bar{x}^{k+1} - x^{k+1}\|^2.
\end{aligned}$$

Then, through recursive derivation, we obtain

$$\begin{aligned}
&(1 - \frac{1+\rho}{2})\sum_{k=0}^K\|\mathbf{1}\bar{s}^k - s^k\|^2 && \text{(A.23)}\\
\leqslant& \|\mathbf{1}\bar{s}^0 - s^0\|^2 + \frac{4\eta^2}{1-\rho}\sum_{k=0}^K\|\mathbf{1}\bar{s}^k - s^k\|^2\\
&+ \frac{4\eta^2}{1-\rho}\sum_{k=0}^K\|\mathbf{1}\bar{x}^k - x^k\|^2.
\end{aligned}$$

With the initialization $s_i^0 = s_j^0, \forall i \in \mathcal{V}$, it follows that

$$\sum_{k=0}^K\|\mathbf{1}\bar{s}^k - s^k\|^2 \leqslant b_1\sum_{k=0}^K\|\mathbf{1}\bar{x}^k - x^k\|^2, \tag{A.24}$$

where $b_1 \triangleq \frac{1}{1-\frac{8\eta^2}{(1-\rho)^2}}\frac{8\eta^2}{(1-\rho)^2}$. Similarly, leveraging the recursion of the outer-level variable in (6) and the initialization $x_i^0 = x_j^0$ for all $i \in \mathcal{V}$, we obtain:

$$\begin{aligned}
&\sum_{k=0}^K\|\mathbf{1}\bar{x}^k - x^k\|^2 && \text{(A.25)}\\
\leqslant& \frac{8\alpha^2}{(1-\rho)^2}\sum_{k=0}^K\|\mathbf{1}\bar{q}^k - q^k\|^2 + \frac{16\alpha^2\eta^2}{(1-\rho)^2}\sum_{k=0}^K\|\mathbf{1}\bar{x}^k - x^k\|^2\\
&+ \frac{16\alpha^2\eta^2}{(1-\rho)^2}\sum_{k=0}^K\|\mathbf{1}\bar{s}^k - s^k\|^2.
\end{aligned}$$

Then, substituting (A.24) into (A.25) and utilizing the condition $1-\frac{16\alpha^2\eta^2}{(1-\rho)^2}(1+b_1) \geqslant \frac{1}{2}$ which follows from $\alpha\eta \leqslant \frac{1-\rho}{8}$ and $\eta \leqslant \frac{1-\rho}{4}$, we can derive an upper bound for the consensus error term associated with the outer-level variable:

$$\sum_{k=0}^{K}\left\|\mathbf{1}\bar{x}^k - x^k\right\|^2 \leqslant \frac{16\alpha^2}{(1-\rho)^2}\sum_{k=0}^{K}\left\|\mathbf{1}\bar{q}^k - q^k\right\|^2. \quad \text{(A.26)}$$

Similarly, leveraging the recursion of the variable q^k in (5) along with the initialization condition $q_i^0 = q_j^0, \forall i \in \mathcal{V}$, we obtain:

$$\big(1-\frac{1+\rho}{2}\big)\sum_{k=0}^{K}\|\mathbf{1}\bar{q}^k - q^k\|^2 \leqslant \frac{2}{1-\rho}\sum_{k=0}^{K}\|u^{k+1}-u^k\|^2. \quad \text{(A.27)}$$

where the last term can be bounded by

$$\begin{aligned}&\|u^{k+1}-u^k\|^2 \quad \text{(A.28)}\\ \leqslant & 6L_{px,\lambda}^2\,\|x^k - x^{k-1}\|^2 + 3L_{p\theta,\lambda}^2\beta^2\|\nabla_\theta P_\lambda(x^k,\theta^k;s^k)\|^2\\ &+3\lambda^2 L_{g,1}^2\gamma^2\|\nabla_\theta g(x^k,z^k)\|^2.\end{aligned}$$

Furthermore, for the first term on the right-hand side of (A.28), it can be rewritten as $x^k - x^{k-1} = (I_{mn} - \mathcal{W})(\mathbf{1}\bar{x}^{k-1} - x^{k-1}) + \alpha(\mathbf{1}\bar{q}^{k-1} - q^{k-1}) - \alpha\mathbf{1}\bar{y}^{k-1}$. Thus, we can derive that

$$\begin{aligned}&\|x^k - x^{k-1}\|^2 \quad \text{(A.29)}\\ \leqslant & 12\|\mathbf{1}\bar{x}^{k-1}-x^{k-1}\|^2+3\alpha^2\|\mathbf{1}\bar{q}^{k-1}-q^{k-1}\|^2+3m\alpha^2\|\bar{y}^{k-1}\|^2\\ \leqslant & 15\|\mathbf{1}\bar{x}^{k-1}-x^{k-1}\|^2+3\alpha^2\|\mathbf{1}\bar{q}^{k-1}-q^{k-1}\|^2\\ &+9\alpha^2 m\|\nabla_x p_\lambda(\bar{x}^{k-1},\theta^k;\bar{s}^{k-1})\|^2\\ &+\frac{9\lambda^2L_{g,1}^2\alpha^2(1-\mu\gamma)^N}{4\mu_g^2}\|\nabla_\theta g(\mathbf{1}\bar{x}^{k-1},z^{k-1})\|^2,\end{aligned}$$

where the last step exploits the relation $\bar{y}^{k-1} = J_n\nabla_x P_\lambda(x^{k-1},\theta^k,z^k;s^{k-1})$ and introduces the term $\nabla_x p_\lambda(\bar{x}^{k-1},\theta^k;\bar{s}^{k-1})$ to upper bound $\|\bar{y}^{k-1}\|^2$, leveraging the gradient-Lipschitz continuity of the involved functions and the result (A.5). Before addressing the second term on the right-hand side of (A.28), we first establish an upper bound for the term $\|x^*(\theta^k,\bar{s}^k)-\bar{x}^k\|$. By combining the inequality $\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\| \leqslant \frac{L_{p\theta,\lambda}}{\sqrt{m}}\|x^*(\theta^k,\bar{s}^k)-\bar{x}^k\|$ $+\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|$ with the result in (A.15) and applying the condition $1-2\sigma_2 L_{p\theta,\lambda} \geqslant \frac{1}{2}$ which follows from $\beta \leqslant \frac{1}{2\sigma_2 L_{p\theta,\lambda}}$, we can derive the following result:

$$\begin{aligned}&\|x^*(\theta^k,s^k)-x^k\| \quad \text{(A.30)}\\ \leqslant & 2\sigma_3\|\nabla_x p_\lambda(\bar{x}^k,\theta^{k+1};s^k)\| + \frac{2\sigma_2\beta L_{p\theta,\lambda}}{\sqrt{m}}\|\mathbf{1}\bar{x}^k - x^k\|\\ &+2\sigma_2\beta\sqrt{m}\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|.\end{aligned}$$

Next, by employing the upper bound of the term $\|x^*(\theta^k,\bar{s}^k)-x^k\|^2$ in (A.30) and the condition $\beta \leqslant \frac{1}{2\sigma_2 L_{p\theta,\lambda}}$, we can bound the second term on the right-hand side of (A.28) as:

$$\begin{aligned}&\left\|\nabla_\theta P_\lambda\left(x^k,\theta^k;\mathbf{1}\bar{s}^k\right)\right\|^2 \quad \text{(A.31)}\\ \leqslant & 2m^2\left\|\nabla_\theta p_\lambda\left(\bar{x}^k,\theta^k;\bar{s}^k\right)\right\|^2 + 2L_{p\theta,\lambda}^2\left\|\mathbf{1}\bar{x}^k - x^k\right\|^2\\ \leqslant & 8m^2\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|^2+48L_{p\theta,\lambda}^4\sigma_2^2\beta^2\|\mathbf{1}\bar{x}^k-x^k\|^2\\ &+48L_{px,\theta}^2\sigma_3^2 m\left\|\nabla_x p_\lambda\left(\bar{x}^k,\theta^{k+1};\bar{s}^k\right)\right\|^2.\end{aligned}$$

Then, by resorting to the fact that $\|\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2 = m^2\|\nabla_\theta p_\lambda(\bar{x}^k,\theta^k;\bar{s}^k)\|^2$, we can obtain an upper bound for the term $\|\nabla_\theta P_\lambda(x^k,\theta^k;\mathbf{1}\bar{s}^k)\|^2$ in (A.28). Then, we deal with the last term in (A.28) as follows:

$$\|\nabla_\theta g(x^k,z^k)\| \leqslant \|\nabla_\theta g(\mathbf{1}\bar{x}^k,z^k)\| + L_{g,1}\|\mathbf{1}\bar{x}^k - x^k\|. \quad \text{(A.32)}$$

Then, substituting the result (A.29), (A.31) and (A.32) into (A.28), we can derive that:

$$\begin{aligned}&(1-\frac{1+\rho}{2}-\frac{36L_{p\theta,\lambda}^2\alpha^2}{1-\rho})\frac{4}{1-\rho}\sum_{k=0}^{K}\|\mathbf{1}\bar{q}^k - q^k\|^2\\ \leqslant & \frac{8}{(1-\rho)^2}\sum_{k=0}^{K}d_1 m\|\nabla_x p_\lambda(\bar{x}^k,\theta^{k+1};\bar{s}^k)\|^2 \quad \text{(A.33)}\\ &+\frac{8}{(1-\rho)^2}\sum_{k=0}^{K}d_2 m^2\|\nabla_\theta p_\lambda(x^*(\theta^k,\bar{s}^k),\theta^k;\bar{s}^k)\|^2\\ &+\frac{8}{(1-\rho)^2}\sum_{k=0}^{K}d_3\|\nabla_\theta g(\mathbf{1}\bar{x}^k,z^k)\|^2\\ &+\frac{8}{(1-\rho)^2}\sum_{k=0}^{K}d_4\|\mathbf{1}\bar{x}^k - x^k\|^2,\end{aligned}$$

where $d_4 = 90L_{p\theta,\lambda}^2 + 144L_{p\theta,\lambda}^6\sigma_2^2\beta^4 + 6\lambda^2L_{g,1}^4\gamma^2 \leqslant 108L_{p\theta,\lambda}^2 \triangleq \tilde{d}_4$ by the condition that $\gamma \leqslant \frac{1}{L_{g,1}}$ and $\beta \leqslant \min\{\frac{1}{4\sigma_2 L_{p\theta,\lambda}},\frac{1}{L_{p\theta,\lambda}}\}$. Noting that $(1-\frac{1+\rho}{2}-\frac{36L_{p\theta,\lambda}^2\alpha^2}{1-\rho})\frac{4}{1-\rho} \geqslant 1$ holds by the condition $\alpha \leqslant \frac{1-\rho}{12L_{p\theta,\lambda}}$ and $1-\frac{8}{(1-\rho)^2}\frac{16\alpha^2}{(1-\rho)^2}\tilde{d}_4 \geqslant \frac{1}{2}$ holds by the condition $\alpha \leqslant \frac{(1-\rho)^2}{120L_{p\theta,\lambda}}$, the derived result can be obtained by substituting (A.33) into (A.26). This completes the proof. ■

F. Proof of Corollary 2

First, building upon the result in [15], when Assumptions 1–3 hold, along with an additional assumption on the smoothness of ∇f_i and ∇g_i, we establish the relationship between the gradient of the penalty function and the hypergradient as follows: $\|\nabla\phi(\hat{x})-\nabla p_\lambda(\hat{x})\| \leqslant \mathcal{O}(\frac{\kappa^3}{\lambda})$. Furthermore, leveraging Corollary 1 and the discussion in Remark 6, we derive $\frac{1}{K+1}\sum_{k=0}^{K}\nabla p_\lambda(\bar{x}^k) \leqslant \mathcal{O}(\frac{\kappa}{\alpha K}+\frac{\kappa}{\alpha K}(1-\frac{1}{\kappa})^{(\kappa K)}) \approx \mathcal{O}(\frac{\kappa}{\alpha K})$. Thus, integrating the above result, we obtain an upper bound for the hypergradient measure:

$$\frac{1}{K+1}\sum_{k=0}^{K}\nabla\phi(\bar{x}^k) \leqslant \mathcal{O}(\frac{\kappa}{\alpha K}+\frac{\kappa^6}{\lambda^2}). \quad \text{(A.34)}$$

Next, we determine the dependence of the relevant parameters on the condition number. Notably, when $N \geqslant \frac{-3\ln\kappa}{\ln(1-\mu\gamma)}$ is satisfied, it follows that $\frac{\mu_g^2}{(1-\mu\gamma)^N L_{g,1}^2} \geqslant \kappa$. Under this setting, one of the step size conditions on α in Theorem 1 requires that $\alpha \leqslant \mathcal{O}(\frac{(1-\rho)\kappa\gamma}{\lambda^2})$, while the condition $\frac{\lambda^2 L_{g,1}^2(1-\mu\gamma)^N}{\mu_g^2} \geqslant 1$ in Corollary 1 imposes the requirement $\lambda^2 \geqslant \mathcal{O}(\kappa^{-1})$. Accordingly, we set the step sizes as $\alpha = \mathcal{O}((1-\rho)^2\kappa^{-2.5}K^{-\frac{1}{3}})$, $\beta = \mathcal{O}((1-\rho)^2\kappa^{-2.5}K^{-\frac{1}{3}})$, $\gamma = \mathcal{O}(1-\rho)$, $\lambda = \mathcal{O}(\kappa^{1.5}K^{\frac{1}{6}})$, which ensures that the step size condition in Theorem 1 is satisfied. Substituting these parameter choices into (A.34) yields the desired result. This completes the proof. ■