Appendix

A. Proof of Result in Section 2

A.1. Proof of Lemma 2.1

We first provide the following technical result to help derive Lemma 2.1.

LEMMA A.1. *Let $(x^*, \gamma^*) \in \mathbb{R}^n \times \mathbb{R}^m_+$ be a saddle point to the following convex optimization problem*

$$\min_{x \in \mathcal{X}} \left\{ f(x) \mid g(x) \leq 0 \right\}, \tag{A.1}$$

where f and g_i, $i = 1, \ldots, m$ are closed proper convex functions and $\mathcal{X} \subseteq \mathbb{R}^n$ is a nonempty closed convex set. Then, it holds for all $x \in \mathcal{X}$ that

$$f(x) - f(x^*) \geq -\|\gamma^*\| \left\| g(x)_+ \right\|.$$

Proof: The arguments here essentially follow from the proof to (Lan and Monteiro 2013, Corollary 2). Consider the following Lagrangian function associated with (A.1):

$$\mathcal{L}(x; \gamma) = f(x) + \gamma^\top g(x), \quad \forall (x, \gamma) \in \mathcal{X} \times \mathbb{R}^m_+.$$

Since (x^*, γ^*) is a saddle point, we have that

$$\mathcal{L}(x^*; \gamma) \leq \mathcal{L}(x^*; \gamma^*) = f(x^*) \leq \mathcal{L}(x; \gamma^*), \quad \forall (x, \gamma) \in \mathcal{X} \times \mathbb{R}^m_+.$$

For any given $x \in \mathcal{X}$, the above inequality implies that

$$f(x) = \mathcal{L}(x; \gamma^*) - g(x)^\top \gamma^* \geq \mathcal{L}(x^*; \gamma^*) - g(x)^\top \gamma^* = f(x^*) - g(x)^\top \gamma^*.$$

Since $(g(x)_+ - g(x))^\top \gamma^* \geq 0$, we further have that

$$f(x) \geq f(x^*) - (g(x)_+)^\top \gamma^* \geq f(x^*) - \|\gamma^*\| \|g(x)_+\|.$$

This completes the proof. ∎

Now we are ready to prove Lemma 2.1.

Proof of Lemma 2.1: Recall that $(x^*, y^*, \gamma^*, \lambda^*)$ is a saddle point to problem (2.1). Then, by using (2.2), we observe that (x^*, γ^*) is saddle point to the following convex optimization problem

$$\min_{x \in \mathcal{X}} \left\{ F(x, y^*) \mid H(x) \leq 0 \right\}.$$

Hence, by using Lemma A.1, we have

$$F(x, y^*) - F(x^*, y^*) \geq -\|\gamma^*\| \left\| H(x)_+ \right\|, \quad \forall x \in \mathcal{X}.$$

Similarly, it holds that

$$-F(x^*, y) + F(x^*, y^*) \geq -\|\lambda^*\| \left\| G(y)_+ \right\|, \quad \forall y \in \mathcal{Y}.$$

The desired result is obtained by combining the two inequalities above. ∎

B. Proof of Results in Section 3

We first present a three-point lemma, which is frequently used throughout our analysis.

LEMMA B.1 **(Lemma 3.8 of Lan (2020))**. *Let Y be a given closed convex set and V be some Bregman distance, and assume function ϕ is μ-strongly convex such that $\phi(y)-\phi(\bar{y})-\phi'(\bar{y})^\top(y-\bar{y}) \geq \mu V(\bar{y},y)$ for all $y,y' \in Y$. For given π, if $\hat{y} \in \arg\min_{y\in Y}\{\pi^\top y+\phi(y)+\tau V(\bar{y},y)\}$, then*

$$(\hat{y}-y)^\top \pi + \phi(\hat{y}) - \phi(y) \leq \tau V(\bar{y},y) - (\tau+\mu)V(\hat{y},y) - \tau V(\bar{y},\hat{y}), \quad \forall y \in Y.$$

B.1. Lemma B.2 and Its Proof

We provide the following result to resolve the complex structure and facilitate our analysis.

LEMMA B.2. *Let $\{\delta_t\}_{t=0}^{N-1}$ be a sequence of random variables. Suppose δ_t is conditionally mean-zero such that $\mathbb{E}[\delta_t \mid \delta_0,\delta_1,\cdots,\delta_{t-1}]=0$ for all $t\geq 1$. For any π such that $\mathbb{E}[\|\pi\|^2]<+\infty$, it holds that:*

(a) Suppose $\mathbb{E}\Big[\|\delta_t\|^2 \mid \delta_0,\delta_1,\cdots,\delta_{t-1}\Big] \leq \sigma^2$, then

$$\mathbb{E}[\sum_{t=0}^{N-1}\pi^\top\delta_t] \leq \sqrt{N}\mathbb{E}[\|\pi\|]\sigma.$$

(b) Let $\tau>0$ be a positive number, then

$$\mathbb{E}\Big[\sum_{t=0}^{N-1}\pi^\top\delta_t\Big] \leq \frac{\tau}{2}\mathbb{E}[\|\pi\|^2] + \sum_{t=0}^{N-1}\frac{1}{2\tau}\mathbb{E}[\|\delta_t\|^2].$$

(c) Let $\{\tau_t\},\{\rho_t\}$ be two positive sequences such that $\tau_t+\rho_t \geq \tau_{t+1}$, then

$$\mathbb{E}\Big[\sum_{t=0}^{N-1}\pi^\top\delta_t\Big] \leq \Big(\frac{\tau_0}{2}+\sum_{t=0}^{N-1}\frac{\rho_t}{2}\Big)\mathbb{E}[\|\pi\|^2] + \sum_{t=0}^{N-1}\frac{1}{2\tau_t}\mathbb{E}[\|\delta_t\|^2].$$

Proof: Part (a) comes from Lemma 2 of Zhang and Lan (2020), part (b) can be derived by similar ideas with $\tau>0$ been kept as a generic value, and part (c) is a generalization of part (b). Here we present the proof of part (c), and part b) can be derived from part c) by setting $\tau_t=\tau$ and $\rho_t=0$. Part (c): We define an auxiliary sequence $\{\pi_t\}$ by

$$\pi_t = \begin{cases} 0 & \text{if } t=0, \\ \arg\min_{\tilde{\pi}} -\delta_{t-1}^\top\tilde{\pi} + \frac{\tau_{t-1}}{2}\|\pi_{t-1}-\tilde{\pi}\|^2 + \frac{\rho_{t-1}}{2}\|\pi_0-\tilde{\pi}\|^2 & \text{if } t\geq 1. \end{cases}$$

We can see that π_t is conditionally independent of δ_t. By using the three-point lemma B.1, we have for all $t\geq 0$ that for any π,

$$-(\pi_{t+1}-\pi)^\top\delta_t \leq \frac{\tau_t}{2}\|\pi_t-\pi\|^2 - \frac{\tau_t}{2}\|\pi_{t+1}-\pi_t\|^2 - \frac{\tau_t+\rho_t}{2}\|\pi_{t+1}-\pi\|^2 - \frac{\rho_t}{2}\|\pi_{t+1}-\pi_0\|^2 + \frac{\rho_t}{2}\|\pi_0-\pi\|^2,$$

which further implies that

$$\begin{aligned}
-(\pi_t-\pi)^\top\delta_t &\le \frac{\tau_t}{2}\|\pi_t-\pi\|^2-\frac{\tau_t}{2}\|\pi_{t+1}-\pi_t\|^2-\frac{\tau_t+\rho_t}{2}\|\pi_{t+1}-\pi\|^2\\
&\quad -\frac{\rho_t}{2}\|\pi_{t+1}-\pi_0\|^2+\frac{\rho_t}{2}\|\pi_0-\pi\|^2+(\pi_{t+1}-\pi_t)^\top\delta_t\\
&\le \frac{\tau_t}{2}\|\pi_t-\pi\|^2-\frac{\tau_t+\rho_t}{2}\|\pi_{t+1}-\pi\|^2-\frac{\rho_t}{2}\|\pi_{t+1}-\pi_0\|^2+\frac{\rho_t}{2}\|\pi_0-\pi\|^2+\frac{1}{2\tau_t}\|\delta_t\|^2,
\end{aligned}$$

where the last inequality holds by the fact that $(\pi_{t+1}-\pi_t)^\top\delta_t \le \frac{1}{2\tau_t}\|\delta_t\|^2+\frac{\tau_t}{2}\|\pi_{t+1}-\pi_t\|^2$. By summing the above inequality over $t=0,1,\cdots,N-1$ and noting that $\pi_0=0$, $\tau_t+\rho_t\ge\tau_{t+1}$, we arrive at

$$\begin{aligned}
\sum_{t=0}^{N-1}\delta_t^\top\pi &\le \sum_{t=0}^{N}\delta_t^\top\pi_t+\Big(\frac{\tau_0}{2}+\sum_{t=0}^{N-1}\frac{\rho_t}{2}\Big)\|\pi\|^2+\sum_{t=0}^{N-1}\frac{1}{2\tau_t}\|\delta_t\|^2+\sum_{t=0}^{N-1}\frac{\tau_{t+1}-\tau_t-\rho_t}{2}\|\pi_{t+1}-\pi\|^2\\
&\le \sum_{t=0}^{N}\delta_t^\top\pi_t+\Big(\frac{\tau_0}{2}+\sum_{t=0}^{N-1}\frac{\rho_t}{2}\Big)\|\pi\|^2+\sum_{t=0}^{N-1}\frac{1}{2\tau_t}\|\delta_t\|^2
\end{aligned}$$

Since $\mathbb{E}[\pi_t^\top\delta_t\mid\delta_0,\delta_1,\cdots,\delta_{t-1}]=0$, we take expectations on both sides and conclude that

$$\mathbb{E}\Big[\sum_{t=0}^{N-1}\pi^\top\delta_t\Big]\le\Big(\frac{\tau_0}{2}+\sum_{t=0}^{N-1}\frac{\rho_t}{2}\Big)\mathbb{E}[\|\pi\|^2]+\sum_{t=0}^{N-1}\frac{1}{2\tau_t}\mathbb{E}[\|\delta_t\|^2].$$

This completes the proof. ∎

B.2. Proof of Lemma 3.1

Proof: For any $(x,y,\lambda,\gamma)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}_{m_1}\times\mathbb{R}_{m_2}$, we split the gap function $Q(z_{t+1},z)$ in the following manner:

$$\begin{aligned}
&\mathcal{L}(x_{t+1},y,\gamma,\lambda_{t+1})-\mathcal{L}(x,y_{t+1},\gamma_{t+1},\lambda)\\
&=F(x_{t+1},y)+\gamma^\top H(x_{t+1})-\lambda_{t+1}^\top G(y)-F(x,y_{t+1})-\gamma_{t+1}^\top H(x)+\lambda^\top G(y_{t+1})\\
&=F(x_{t+1},y)-F(x_t,y)+F(x_t,y)-F(x_t,y_t)+F(x_t,y_t)-F(x,y_t)+F(x,y_t)-F(x,y_{t+1})\\
&\quad+\gamma^\top H(x_{t+1})-\lambda_{t+1}^\top G(y)-\gamma_{t+1}^\top H(x)+\lambda^\top G(y^{t+1})\\
&\le C_f\|x_{t+1}-x_t\|+C_f\|y_{t+1}-y_t\|+Q_x^{t+1}+Q_y^{t+1}\\
&\le \frac{3C_f^2}{2\eta_t}+\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2+\frac{3C_f^2}{2\kappa_t}+\frac{\kappa_t}{6}\|y_{t+1}-y_t\|^2+Q_x^{t+1}+Q_y^{t+1},
\end{aligned}\tag{B.1}$$

where

$$\begin{aligned}
Q_x^{t+1}&:=F(x_t,y_t)-F(x,y_t)+\gamma^\top H(x_{t+1})-\gamma_{t+1}^\top H(x),\\
Q_y^{t+1}&:=F(x_t,y)-F(x_t,y_t)-\lambda_{t+1}^\top G(y)+\lambda^\top G(y_{t+1}),
\end{aligned}$$

and the first inequality holds since F is Lipschitz continuous under Assumption 2.1. Since $F(\bullet,y_t)+\gamma_{t+1}^\top H(\bullet)$ is a convex function, we have

$$\begin{aligned}
Q_x^{t+1}&=F(x_t,y_t)-F(x,y_t)+\gamma_{t+1}^\top\big(H(x_t)-H(x)\big)-\gamma_{t+1}^\top H(x_t)+\gamma^\top H(x_{t+1})\\
&\le\Big(\widetilde{\nabla}_xF(x_t,y_t)+\sum_{i=1}^{m_1}\gamma_{t+1,i}\widetilde{\nabla}H_i(x_t)\Big)^\top(x_t-x)+\gamma^\top H(x_{t+1})-\gamma_{t+1}^\top H(x_t).
\end{aligned}$$

Similarly, by using the concavity of $F(x_t,\bullet) - \lambda_{t+1}^\top G(\bullet)$, we have

$$\begin{aligned} Q_y^{t+1} &= F(x_t,y) - F(x_t,y_t) - \lambda_{t+1}^\top\big(G(y)-G(y_t)\big) + \lambda^\top G(y_{t+1}) - \lambda_{t+1}^\top G(y_t) \\ &\le \Big(\widetilde{\nabla}_y F(x_t,y_t) - \sum_{j=1}^{m_2}\lambda_{t+1,j}\widetilde{\nabla} G_j(y_t)\Big)^\top (y-y_t) + \lambda^\top G(y_{t+1}) - \lambda_{t+1}^\top G(y_t). \end{aligned}$$

By substituting the above inequalities into (B.1), and conducting a similar analysis for Q_y^{t+1}, we conclude

$$\begin{aligned} &\mathcal{L}(x_{t+1},y,\gamma,\lambda_{t+1}) - \mathcal{L}(x,y_{t+1},\gamma_{t+1},\lambda) \\ &\le \Big(\widetilde{\nabla}_x F(x_t,y_t) + \sum_{i=1}^{m_1}\gamma_{t+1,i}\widetilde{\nabla} H_i(x_t)\Big)^\top (x_t - x) + \gamma^\top H(x_{t+1}) - \gamma_{t+1}^\top H(x_t) \\ &\quad + \Big(\widetilde{\nabla}_y F(x_t,y_t) - \sum_{j=1}^{m_2}\lambda_{t+1,j}\widetilde{\nabla} G_j(y_t)\Big)^\top (y-y_t) + \lambda^\top G(y_{t+1}) - \lambda_{t+1}^\top G(y_t) \\ &\quad + \frac{3C_f^2}{2\eta_t} + \frac{\eta_t}{6}\|x_{t+1}-x_t\|^2 + \frac{3C_f^2}{2\kappa_t} + \frac{\kappa_t}{6}\|y_{t+1}-y_t\|^2. \end{aligned}$$

This completes the proof. ■

B.3. Proof of Lemma 3.2

Proof: Recall the definitions of $\mathcal{H}$ and Δ_x^{t+1} in (3.4), (3.3), and (3.7). It holds that

$$\begin{aligned} \Delta_x^{t+1} &= \Big(\widetilde{\nabla} F(x_t,y_t) + \sum_{i=1}^{m_1}\gamma_{t+1,i}\widetilde{\nabla} H_i(x_t)\Big)^\top (x_t - x) = \widetilde{\nabla}_x \mathcal{L}(x_t,y_t,\gamma_{t+1})^\top (x_t-x) \\ &= \widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)^\top (x_{t+1}-x) + \Big(\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t,y_t,\gamma_{t+1})\Big)^\top (x-x_t) \\ &\quad + \widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)^\top (x_t - x_{t+1}). \end{aligned} \tag{B.2}$$

By recalling the update rule (3.2) for x_{t+1} and using the three-point Lemma B.1 in Appendix Section B, for all $x \in \mathcal{X}$, we have

$$\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)^\top (x_{t+1}-x) \le \frac{\eta_t}{2}\|x_t - x\|^2 - \frac{\eta_t}{2}\|x_t - x_{t+1}\|^2 - \frac{\eta_t}{2}\|x_{t+1}-x\|^2. \tag{B.3}$$

Meanwhile, it holds by simple calculations that

$$\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)^\top (x_t - x_{t+1}) \le \frac{3\|\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)\|^2}{2\eta_t} + \frac{\eta_t\|x_t - x_{t+1}\|^2}{6}$$

Therefore, we have from (B.2) that

$$\begin{aligned} \Delta_x^{t+1} + \frac{\eta_t}{3}\|x_t - x_{t+1}\|^2 &\le \frac{\eta_t}{2}\|x_t - x\|^2 - \frac{\eta_t}{2}\|x_{t+1}-x\|^2 + \Big(\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t,y_t,\gamma_{t+1})\Big)^\top (x-x_t) \\ &\quad + \frac{3\|\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)\|^2}{2\eta_t}. \end{aligned} \tag{B.4}$$

By summing (B.4) over $t=0,1,\cdots,K-1$, we see that for any $x\in\mathcal{X}$,

$$\begin{aligned}&\sum_{t=0}^{K-1}\big(\Delta_x^{t+1}+\frac{\eta_t}{3}\|x_t-x_{t+1}\|^2\big)+\frac{\eta_0}{2}\|x_K-x\|^2\\ \le&\frac{\eta_0}{2}\|x_0-x\|^2+\sum_{t=0}^{K-1}\frac{3\|\widetilde{\nabla}_xL(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)\|^2}{2\eta_t}\\ &+\sum_{t=0}^{K-1}\big(\widetilde{\nabla}_xL(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)-\widetilde{\nabla}_x\mathcal{L}(x_t,y_t,\gamma_{t+1})\big)^\top(x-x_t).\end{aligned}\tag{B.5}$$

Next, we focus on the term Δ_γ^{t+1} and split it as

$$\begin{aligned}\Delta_\gamma^{t+1}&=\gamma^\top H(x_{t+1})-\gamma_{t+1}^\top H(x_t)\\ &=(\gamma-\gamma_{t+1})^\top H(x_t)+\gamma^\top[H(x_{t+1})-H(x_t)]\\ &=\underbrace{-h(x_t,\xi_t^1)^\top(\gamma_{t+1}-\gamma)}_{\Delta_{\gamma,1}^{t+1}}+\underbrace{(\gamma-\gamma_{t+1})^\top[H(x_t)-h(x_t,\xi_t^1)]}_{\Delta_{\gamma,2}^{t+1}}+\underbrace{\gamma^\top\big(H(x_{t+1})-H(x_t)\big)}_{\Delta_{\gamma,3}^{t+1}}.\end{aligned}\tag{B.6}$$

First, consider $\Delta_{\gamma,1}^{t+1}$, recall the update rule that

$$\gamma_{t+1}=\operatorname*{arg\,min}_{\gamma\in\mathbb{R}_+^{m_1}}\Big\{-h(x_t,\xi_t^1)^\top\gamma+\frac{\beta_t}{2}\|\gamma_t-\gamma\|^2\Big\}.$$

By using the three-point Lemma B.1 in Appendix Section B, we have for any $\gamma\in\mathbb{R}_+^{m_1}$ that

$$\Delta_{\gamma,1}^{t+1}=-h(x_t,\xi_t)^\top(\gamma_{t+1}-\gamma)\le\frac{\beta_t}{2}\Big(\|\gamma_t-\gamma\|^2-\|\gamma_t-\gamma_{t+1}\|^2-\|\gamma_{t+1}-\gamma\|^2\Big).\tag{B.7}$$

Meanwhile, it holds that

$$\begin{aligned}\Delta_{\gamma,2}^{t+1}&=\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma-\gamma_t)+\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma_t-\gamma_{t+1})\\ &\le\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma-\gamma_t)+\frac{\|H(x_t)-h(x_t,\xi_t^1)\|^2}{2\beta_t}+\frac{\beta_t}{2}\|\gamma_t-\gamma_{t+1}\|^2.\end{aligned}\tag{B.8}$$

By the Lipschitz continuity of H and some simple computations, we have

$$\begin{aligned}\Delta_{\gamma,3}^{t+1}&\le\|\gamma\|\|H(x_{t+1})-H(x_t)\|\\ &\le C_h\|\gamma\|\|x_{t+1}-x_t\|\le\frac{3\|\gamma\|^2C_h^2}{2\eta_t}+\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2.\end{aligned}\tag{B.9}$$

Thus, it holds from (B.6), (B.7), (B.8) and (B.9) that for any $\gamma\in\Re_+^{m_1}$,

$$\begin{aligned}\sum_{t=0}^{K-1}\Delta_\gamma^{t+1}+\frac{\beta_0}{2}\|\gamma_K-\gamma\|^2\le\frac{\beta_0}{2}\|\gamma_0-\gamma\|^2+\sum_{t=0}^{K-1}\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma-\gamma_t)\\ +\sum_{t=0}^{K-1}\Big(\frac{\|H(x_t)-h(x_t,\xi_t^1)\|^2}{2\beta_t}+\frac{3\|\gamma\|^2C_h^2}{2\eta_t}+\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2\Big).\end{aligned}\tag{B.10}$$

Summing (B.5) and (B.8), we obtain the desired inequality and complete the proof. ∎

B.4. Proof of Lemma 3.4

Recall the definition of U_t in (3.8):

$$U_t(x,\gamma) = \big(\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})\big)^\top (x - x_t) + \frac{3\|\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2)\|^2}{2\eta_t} + \big(H(x_t) - h(x_t, \xi_t^1)\big)^\top (\gamma - \gamma_t) + \frac{\|H(x_t) - h(x_t, \xi_t^1)\|^2}{2\beta_t}.$$

We first note from the independence between $\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})$ and x_t that

$$\mathbb{E}\Big[\big(\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})\big)^\top x_t\Big] = 0 \tag{B.11}$$

Meanwhile, for any $x \in \mathcal{X}$ satisfying $\mathbb{E}[\|x\|^2] \le +\infty$, we know from Lemma B.2 (b) that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=0}^{K-1} \big(\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})\big)^\top x\Big] \\
&\le \frac{\eta_0}{2}\mathbb{E}[\|x\|^2] + \sum_{t=0}^{K-1} \frac{1}{2\eta_t}\mathbb{E}[\|\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})\|^2] \\
&\le \frac{\eta_0}{2}\mathbb{E}[\|x\|^2] + \sum_{t=0}^{K-1} \frac{1}{\eta_t}\mathbb{E}\Big[\|\widetilde{\nabla} f(x_t, y_t, \omega_t^1) - \widetilde{\nabla} F(x_t, y_t)\|^2 + \|(\widetilde{\nabla} h(x, \xi_t^2) - \widetilde{\nabla} H(x_t))\gamma_{t+1}\|^2\Big] \\
&\le \frac{\eta_0}{2}\mathbb{E}[\|x\|^2] + \sum_{t=0}^{K-1} \frac{1}{\eta_t}(C_f^2 + C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]),
\end{aligned} \tag{B.12}$$

where we use the facts that $\eta_t \equiv \eta_0$, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, the independence between γ_{t+1} and $\nabla h(x_t, \xi_t^2)$ in the update of Algorithm 1, Assumption 2.1 that $\mathbb{E}[\|\widetilde{\nabla} f(x_t, y_t, \omega_t^1) - \widetilde{\nabla} F(x_t, y_t)\|^2] \le C_f^2$, and Assumption 2.2 that $\mathbb{E}[\|\widetilde{\nabla} h(x, \xi_t^2) - \widetilde{\nabla} H(x_t)\|^2] \le C_h^2$ in the last inequality. Assumptions 2.1 and 2.2 also imply

$$\mathbb{E}\Big[\|\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2)\|^2\Big] \le 2C_f^2 + 2C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]. \tag{B.13}$$

Next, we focus on terms in $U_t(x,\gamma)$ involving γ. By using the independency between γ_t and $H(x_t) - h(x_t, \xi_t^1)$ and Lemma B.2 (a), we have for all γ satisfying $\mathbb{E}[\|\gamma\|^2] < +\infty$ that

$$\mathbb{E}\Big[\sum_{t=0}^{K-1} \big(H(x_t) - h(x_t, \xi_t^1)\big)^\top (\gamma - \gamma_t)\Big] \le \sqrt{K}\mathbb{E}[\|\gamma\|]\sigma_h. \tag{B.14}$$

Moreover, Assumption 2.2 implies that

$$\mathbb{E}\big[\|H(x_t) - h(x_t, \xi_t^1)\|^2\big] \le \sigma_h^2. \tag{B.15}$$

Combining (B.11), (B.12), (B.13), (B.14) and (B.15) with the definition of U_t, we obtain the following inequality

$$\mathbb{E}\Big[\sum_{t=0}^{K-1} U_t(x,\gamma)\Big] \le \frac{\eta_0}{2}\mathbb{E}[\|x\|^2] + \sum_{t=0}^{K-1} \frac{4}{\eta_t}\big(C_f^2 + C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]\big) + \sqrt{K}\mathbb{E}[\|\gamma\|]\sigma_h + \frac{K\sigma_h^2}{2\beta_0}.$$

The desired inequality of V_t follows in the similar way. ∎

B.5. Proof of Proposition 3.1

We present a technical result from Lemma 2.8 of Boob et al. (2023).

LEMMA B.3. *Let $\{a_t\}_{t\geq 0}$ be a nonnegative sequence and $b_1, b_2 \geq 0$ be two constants such that $a_0 \leq b_1$. Suppose for all $K \geq 1$, it holds that*

$$a_K \leq b_1 + b_2 \sum_{t=0}^{K-1} a_t.$$

Then we have $a_K \leq b_1(1+b_2)^K$ for all $K \geq 1$.

Proof of Proposition 3.1: By setting $x = x^*$, $y = y^*$, $\gamma = \gamma^*$ and $\lambda = \lambda^*$, and using the minimax relationship (2.2), we have

$$0 \leq \mathbb{E}[\sum_{t=0}^{K-1} Q(z_{t+1}, z^*)] = \mathbb{E}\Big[\sum_{t=0}^{K-1}\big(\mathcal{L}(x_{t+1}, y^*, \gamma^*, \lambda_{t+1}) - \mathcal{L}(x^*, y_{t+1}, \gamma_{t+1}, \lambda^*)\big)\Big], \quad K = 1, \ldots, N.$$

Combining the above inequality with Theorem 3.1, we have

$$\begin{aligned}
&\frac{\beta_0}{2}\mathbb{E}[\|\gamma^* - \gamma_K\|^2] + \frac{\eta_0}{2}\mathbb{E}[\|\lambda_K - \lambda^*\|^2] + \frac{\kappa_0}{2}\mathbb{E}[\|x^* - x_K\|^2] + \frac{\eta_0}{2}\mathbb{E}[\|y^* - y_K\|^2]\\
&\leq \frac{3KC_f^2}{2\eta_0} + \frac{\beta_0}{2}\mathbb{E}[\|\gamma_0 - \gamma^*\|^2] + \sqrt{K}\|\gamma^*\|\sigma_h + \frac{K\sigma_h^2}{2\beta_0} + \frac{3K\|\gamma^*\|^2C_h^2}{2\eta_0} + \frac{\eta_0}{2}\|x_0 - x^*\|^2 + \frac{\eta_0}{2}\|x^*\|^2\\
&\quad + \frac{3KC_f^2}{2\kappa_0} + \frac{\alpha_0}{2}\|\lambda_0 - \lambda^*\|^2 + \sqrt{K}\|\lambda^*\|\sigma_g + \frac{K\sigma_g^2}{2\alpha_0} + \frac{3K\|\lambda^*\|^2C_g^2}{2\kappa_0} + \frac{\kappa_0}{2}\|y_0 - y^*\|^2 + \frac{\kappa_0}{2}\|y^*\|^2\\
&\quad + \sum_{t=0}^{K-1}\frac{4}{\eta_t}(C_f^2 + C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]) + \sum_{t=0}^{K-1}\frac{4}{\kappa_t}(C_f^2 + C_g^2\mathbb{E}[\|\lambda_{t+1}\|^2]).
\end{aligned}$$

By using the facts that $\|a\|^2 \leq 2\|a-b\|^2 + 2\|b\|^2$, we further obtain

$$\begin{aligned}
&\frac{\beta_0}{4}\mathbb{E}[\|\lambda_K\|^2] + \frac{\alpha_0}{4}\mathbb{E}[\|\gamma_K\|^2]\\
&\leq \frac{\beta_0}{4}\mathbb{E}[\|\lambda_K\|^2] + \frac{\alpha_0}{4}\mathbb{E}[\|\gamma_K\|^2] + \frac{\eta_0}{4}\mathbb{E}[\|x_K\|^2] + \frac{\kappa_0}{4}\mathbb{E}[\|y_K\|^2]\\
&\leq \frac{3KC_f^2}{2\eta_0} + \frac{\beta_0}{2}\|\gamma_0 - \gamma^*\|^2 + \sqrt{K}\|\gamma^*\|\sigma_h + \frac{K\sigma_h^2}{2\beta_0} + \frac{3K\|\gamma^*\|^2C_h^2}{2\eta_0} + \frac{\eta_0}{2}\|x_0 - x^*\|^2 + \eta_0\|x^*\|^2\\
&\quad + \frac{3KC_f^2}{2\kappa_0} + \frac{\alpha_0}{2}\|\lambda_0 - \lambda^*\|^2 + \sqrt{K}\|\lambda^*\|\sigma_g + \frac{K\sigma_g^2}{2\alpha_0} + \frac{3K\|\lambda^*\|^2C_g^2}{2\kappa_0} + \frac{\kappa_0}{2}\|y_0 - y^*\|^2 + \kappa_0\|y^*\|^2\\
&\quad + \sum_{t=0}^{K-1}\frac{4}{\eta_t}(C_f^2 + C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]) + \sum_{t=0}^{K-1}\frac{4}{\kappa_t}(C_f^2 + C_g^2\mathbb{E}[\|\lambda_{t+1}\|^2]) + \frac{\beta_0}{2}\|\lambda^*\|^2 + \frac{\alpha_0}{2}\|\gamma^*\|^2.
\end{aligned} \tag{B.16}$$

By setting $\beta_t = \alpha_t = 4\sqrt{N}$, $\eta_t = 4C_h^2\sqrt{N}$, and $\kappa_t = 4C_g^2\sqrt{N}$, and dividing $\sqrt{N}$ on both sides of the above inequality, we further have

$$\begin{aligned}
\mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2] &\leq \mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2] + C_h^2\mathbb{E}[\|x_K\|^2] + C_g^2\mathbb{E}[\|y_K\|^2]\\
&\leq R_K + \frac{1}{N}\sum_{t=0}^{K-1}(\mathbb{E}[\|\gamma_{t+1}\|^2] + \mathbb{E}[\|\lambda_{t+1}\|^2])
\end{aligned} \tag{B.17}$$

where

$$R_K := \frac{11KC_f^2}{8C_h^2N} + 2\|\gamma_0 - \gamma^*\|^2 + \frac{\sqrt{K}\|\gamma^*\|}{\sqrt{N}}\sigma_h + \frac{K\sigma_h^2}{8N} + (\frac{3K}{8N} + 2)\|\gamma^*\|^2 + 2C_h^2\|x_0 - x^*\|^2 + 4C_h^2\|x^*\|^2$$
$$+ \frac{11KC_f^2}{8C_g^2N} + 2\|\lambda_0 - \lambda^*\|^2 + \frac{\sqrt{K}\|\lambda^*\|}{\sqrt{N}}\sigma_g + \frac{K\sigma_g^2}{8N} + (\frac{3K}{8N} + 2)\|\lambda^*\|^2 + 2C_g^2\|y_0 - y^*\|^2 + 4C_g^2\|y^*\|^2.$$

Furthermore, let R be the constant defined in (3.13), we observe that $R_t \le R$ for $t = 1, 2, \cdots, N-1$. Therefore, by rearranging the terms within (B.17), we have

$$(1 - \frac{1}{N})\Big(\mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2]\Big) \le R + \frac{1}{N}\sum_{t=1}^{K-1}\Big(\mathbb{E}[\|\gamma_t\|^2] + \mathbb{E}[\|\lambda_t\|^2]\Big).$$

For $N \ge 2$, the above inequality further implies

$$\mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2] \le \Big(1 - \frac{1}{N}\Big)^{-1}\left(R + \frac{1}{N}\sum_{t=1}^{K-1}\Big(\mathbb{E}[\|\gamma_t\|^2] + \mathbb{E}[\|\lambda_t\|^2]\Big)\right)$$
$$\le 2R + \frac{2}{N}\sum_{t=1}^{K-1}\Big(\mathbb{E}[\|\gamma_t\|^2] + \mathbb{E}[\|\lambda_t\|^2]\Big).$$

By using Lemma B.3, setting $a_t = \mathbb{E}[\|\lambda_t\|^2] + \mathbb{E}[\|\gamma_t\|^2]$, $b_1 = 2R$, and $b_2 = \frac{2}{N}$, we conclude that

$$\mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2] \le 2R\prod_{t=1}^{K-1}\big(1 + \frac{2}{N}\big) = 2R(1 + \frac{2}{N})^{K-1} \le 2Re^2,$$

where the last inequality uses the fact that $1 \le K \le N$. Then, (B.17) further implies that

$$\mathbb{E}[\|\lambda_K\|^2] + \mathbb{E}[\|\gamma_K\|^2] + C_h^2\mathbb{E}[\|x_K\|^2] + C_g^2\mathbb{E}[\|y_K\|^2] \le (2e^2 + 1)R, \quad \forall 1 \le K \le N.$$

This completes the proof. ∎

B.6. Proof of Theorem 3.3

Proof: Denote $\bar{\lambda}_N = \frac{1}{N}\sum_{t=1}^N \lambda_t$, and $\bar{\gamma}_N = \frac{1}{N}\sum_{t=1}^N \gamma_t$. We start by proving inequality (3.3). By using Theorem 3.2 and setting $\gamma = 0$ and $\lambda = 0$, we have for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathbb{E}[\|x\|^2] < +\infty$ and $\mathbb{E}[\|y\|^2] < +\infty$ that

$$\frac{1}{N}\mathbb{E}\Big[\sum_{t=1}^N\big(\mathcal{L}(x_t, y, 0, \lambda_t) - \mathcal{L}(x, y_t, \gamma_t, 0)\big)\Big] = \frac{1}{N}\mathbb{E}\Big[\sum_{t=1}^N Q(z_t, (x, y, 0, 0))\Big]$$
$$\le \frac{1}{\sqrt{N}}\Big(2Re^2 + \frac{11C_f^2}{8C_h^2} + 2\|\gamma_0\|^2 + \frac{\sigma_h^2}{8} + 2C_h^2\mathbb{E}[\|x_0 - x\|^2] + 2C_h^2\mathbb{E}[\|x\|^2]\Big)$$
$$+ \frac{1}{\sqrt{N}}\Big(\frac{11C_f^2}{8C_g^2} + 2\|\lambda_0\|^2 + \frac{\sigma_g^2}{8} + 2C_g^2\mathbb{E}[\|y_0 - y\|^2] + 2C_g^2\mathbb{E}[\|y\|^2]\Big).$$

Since $\mathcal{L}(x,y,\gamma,\lambda)$ is convex in x,λ and concave in y,γ, we have

$$\begin{aligned}&\frac{1}{N}\sum_{t=1}^{N}\Big(\mathcal{L}(x_t,y,0,\lambda_t)-\mathcal{L}(x,y_t,\gamma_t,0)\Big)\\&\geq \mathcal{L}(\bar{x}_N,y,0,\bar{\lambda}_N)-\mathcal{L}(x,\bar{y}_N,\bar{\gamma}_N,0)\\&=F(\bar{x}_N,y)-G(y)^\top\bar{\lambda}_N-F(x,\bar{y}_N)-H(x)^\top\bar{\gamma}_N\geq F(\bar{x}_N,y)-F(x,\bar{y}_N),\end{aligned}$$

where the last inequality holds since $G(y)^\top\bar{\lambda}_N\leq 0$ and $H(x)^\top\bar{\gamma}_N\leq 0$ for any feasible (x,y). Then we conclude that

$$\begin{aligned}&\mathbb{E}\Big[F(\bar{x}_N,y)-F(x,\bar{y}_N)\Big]\\&\leq\frac{1}{\sqrt{N}}\Big(2Re^2+\frac{11C_f^2}{8C_h^2}+2\|\gamma_0\|^2+\frac{\sigma_h^2}{8}+2C_h^2\mathbb{E}[\|x_0-x\|^2]+2C_h^2\mathbb{E}[\|x\|^2]\Big)\\&\quad+\frac{1}{\sqrt{N}}\Big(\frac{11C_f^2}{8C_g^2}+2\|\lambda_0\|^2+\frac{\sigma_g^2}{8}+2C_g^2\mathbb{E}[\|y_0-y\|^2]+2C_g^2\mathbb{E}[\|y\|^2]\Big).\end{aligned}\tag{B.18}$$

Next, we derive the upper bounds for feasibility residuals. Let $\tilde{\gamma}=(\|\gamma^*\|_2+1)\frac{H(\bar{x}_N)_+}{\|H(\bar{x}_N)_+\|_2}$ and $\tilde{\lambda}=(\|\lambda^*\|_2+1)\frac{G(\bar{y}_N)_+}{\|G(\bar{y}_N)_+\|_2}$ and consider the reference point $(x^*,y^*,\tilde{\gamma},\tilde{\lambda})$, we obtain

$$\begin{aligned}&\frac{1}{N}\sum_{t=1}^{N}\Big(\mathcal{L}(x_t,y^*,\tilde{\gamma},\lambda_t)-\mathcal{L}(x^*,y_t,\gamma_t,\tilde{\lambda})\Big)\\&\geq\mathcal{L}(\bar{x}_N,y^*,\tilde{\gamma},\bar{\lambda}_N)-\mathcal{L}(x^*,\bar{y}_N,\bar{\gamma}_N,\tilde{\lambda})\\&=F(\bar{x}_N,y^*)+H(\bar{x}_N)^\top\tilde{\gamma}-G(y^*)^\top\bar{\lambda}_N-\Big[F(x^*,\bar{y}_N)+H(x^*)^\top\bar{\gamma}_N-G(\bar{y}_N)^\top\tilde{\lambda}\Big]\\&\geq F(\bar{x}_N,y^*)+H(\bar{x}_N)^\top\tilde{\gamma}-F(x^*,\bar{y}_N)+G(\bar{y}_N)^\top\tilde{\lambda}\\&=F(\bar{x}_N,y^*)+\big(\|\gamma^*\|_2+1\big)\|H(\bar{x}_N)_+\|_2-F(x^*,\bar{y}_N)+\big(\|\lambda^*\|_2+1\big)\|G(\bar{y}_N)_+\|_2,\end{aligned}\tag{B.19}$$

where the last equality follows from the facts that $H(\bar{x}_N)^\top H(\bar{x}_N)_+=\|H(\bar{x}_N)_+\|_2^2$ and $G(\bar{y}_N)^\top G(\bar{y}_N)_+=\|G(\bar{y}_N)_+\|_2^2$. Meanwhile, from the minimax relationship (2.2)

$$\mathcal{L}(\bar{x}_N,y^*,\gamma^*,\lambda^*)\geq\mathcal{L}(x^*,y^*,\gamma^*,\lambda^*)\geq\mathcal{L}(x^*,\bar{y}_N,\gamma^*,\lambda^*),$$

it holds that

$$\begin{aligned}0\leq\ &\mathcal{L}(\bar{x}_N,y^*,\gamma^*,\lambda^*)-\mathcal{L}(x^*,\bar{y}_N,\gamma^*,\lambda^*)\\&=F(\bar{x}_N,y^*)+H(\bar{x}_N)^\top\gamma^*-F(x^*,\bar{y}_N)+G(\bar{y}_N)^\top\lambda^*\\&\leq F(\bar{x}_N,y^*)+H(\bar{x}_N)_+^\top\gamma^*-F(x^*,\bar{y}_N)+G(\bar{y}_N)_+^\top\lambda^*\\&\leq F(\bar{x}_N,y^*)+\|\gamma^*\|_2\|H(\bar{x}_N)_+\|_2-F(x^*,\bar{y}_N)+\|\lambda^*\|_2\|G(\bar{y}_N)_+\|_2,\end{aligned}\tag{B.20}$$

where the second inequality holds since $\gamma^*\geq 0, H(\bar{x}_N)\leq H(\bar{x}_N)_+$, and $\lambda^*\geq 0, G(\bar{y}_N)\leq G(\bar{y}_N)_+$. Substituting (B.20) into (B.19) and then taking expectation, we have

$$\mathbb{E}[\|H(\bar{x}_N)_+\|_2]+\mathbb{E}[\|G(\bar{y}_N)_+\|_2]\leq\frac{1}{N}\mathbb{E}\Big[\sum_{t=1}^{N}\big(\mathcal{L}(x_t,y^*,\tilde{\gamma},\lambda_t)-\mathcal{L}(x^*,y_t,\gamma_t,\tilde{\lambda})\big)\Big].$$

Since $\|\tilde{\lambda}\| = \|\lambda^*\| + 1$ and $\|\tilde{\gamma}\| = \|\gamma^*\| + 1$, it holds from Theorem 3.2 that

$$\begin{aligned}
&\mathbb{E}[\|H(\bar{x}_N)_+\|_2] + \mathbb{E}[\|G(\bar{y}_N)_+\|_2] \\
\leq &\frac{1}{\sqrt{N}}\Big(2Re^2 + \frac{11C_f^2}{8C_h^2} + 4\|\gamma_0\|^2 + \frac{35}{8}(\|\lambda^*\|+1)^2 + (\|\lambda^*\|+1)\sigma_h + \frac{\sigma_h^2}{8} + 2C_h^2\|x_0 - x^*\|^2 + 2C_h^2\|x^*\|^2\Big) \\
&+ \frac{1}{\sqrt{N}}\Big(\frac{11C_f^2}{8C_g^2} + 4\|\lambda_0\|^2 + \frac{35}{8}(\|\lambda^*\|+1)^2 + (\|\lambda^*\|+1)\sigma_g + \frac{\sigma_g^2}{8} + 2C_g^2\|y_0 - y^*\|^2 + 2C_g^2\|y^*\|^2\Big),
\end{aligned}$$

where the last inequality follows from the fact that $\|\gamma_0 - \tilde{\gamma}\|^2 \leq 2\|\gamma_0\|^2 + 2\|\tilde{\gamma}\|^2 = 2\|\gamma_0\|^2 + 2(\|\gamma^*\|+1)^2$ and $\|\lambda_0 - \tilde{\lambda}\|^2 \leq 2\|\lambda_0\|^2 + 2(\|\lambda^*\|+1)^2$.

Lastly, by setting $(x,y) = (x^*, y^*)$ and combining the above inequality with (B.18) and the lower bound provided in Lemma 2.1, we conclude that there exist two constants $C_1, C_2 > 0$ such that

$$-\frac{C_1}{\sqrt{N}} \leq \mathbb{E}\Big[F(\bar{x}_N, y^*) - F(x^*, \bar{y}_N)\Big] \leq \frac{C_2}{\sqrt{N}}.$$

This completes the proof. ∎

B.7. Proof of Corollary 3.1

Proof: The first inequality can be easily obtained by combining Theorem 3.3 with the fact that

$$F(\bar{x}_N, y^\star(\bar{x}_N)) - F(x^\star(\bar{y}_N), \bar{y}_N) \geq F(\bar{x}_N, y^*) - F(x^*, \bar{y}_N),$$

and the second inequality can be derived by setting $(x,y) = (x^*(\bar{y}_N), y^*(\bar{x}_N))$ in Theorem 3.3 and using the boundedness of $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{Y}}$. ∎

C. Proof of Results in Section 4

C.1. Proof of Lemma 4.1

Proof: Recall the decompostion of Δ_x^{t+1} in (B.2). Under the update rule of x_{t+1} in Algorithm 2, for any $x \in \mathcal{X}$, we obtain from the three-point lemma B.1 that

$$\begin{aligned}
&\mathcal{H}(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2)^\top (x_{t+1} - x) \\
\leq &\frac{\eta_t}{2}\|x_t - x\|^2 - \frac{\eta_t}{2}\|x_t - x_{t+1}\|^2 - \frac{\eta_t + \rho_t}{2}\|x_{t+1} - x\|^2 - \frac{\rho_t}{2}\|x_{t+1} - x_0\|^2 + \frac{\rho_t}{2}\|x - x_0\|^2.
\end{aligned}$$

Similar to (B.5), it holds for all $x \in \mathcal{X}$ that

$$\begin{aligned}
\Delta_x^{t+1} + \frac{\eta_t}{3}\|x_t - x_{t+1}\|^2 \leq &\frac{\eta_t}{2}\|x_t - x\|^2 - \frac{\eta_t + \rho_t}{2}\|x_{t+1} - x\|^2 + \frac{\rho_t}{2}\|x - x_0\|^2 \\
&+ \big(\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2) - \widetilde{\nabla}_x \mathcal{L}(x_t, y_t, \gamma_{t+1})\big)^\top (x - x_t) \\
&+ \frac{3\|\widetilde{\nabla}_x L(x_t, y_t, \gamma_{t+1}, \omega_t^1, \xi_t^2)\|^2}{2\eta_t}.
\end{aligned}$$

Summing the above inequality over $t=0,1,\dots,K-1$ and noting that $\eta_{t+1}\le\eta_t+\rho_t$, we know that

$$\begin{aligned}
&\sum_{t=0}^{K-1}\big(\Delta_x^{t+1}+\frac{\eta_t}{3}\|x_t-x_{t+1}\|^2\big)+\frac{\eta_K}{2}\|x_K-x\|^2\\
\le&(\frac{\eta_0}{2}+\sum_{t=0}^{K-1}\frac{\rho_t}{2})\|x_0-x\|^2+\sum_{t=0}^{K-1}\big(\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)-\widetilde{\nabla}_x\mathcal{L}(x_t,y_t,\gamma_{t+1})\big)^\top(x-x_t)\\
&+\sum_{t=0}^{K-1}\frac{3\|\widetilde{\nabla}_x L(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)\|^2}{2\eta_t},\quad\forall x\in\mathcal{X}.
\end{aligned}\tag{C.1}$$

Recall the decompostion of Δ_γ^{t+1} in (B.6). From the update rule of γ_t in (4.1), we know from Lemma B.1 that for all $\gamma\in\mathbb{R}_+^{m_1}$,

$$\begin{aligned}
&\Delta_{\gamma,1}^{t+1}=-h(x_t,\xi_t^1)^\top(\gamma_{t+1}-\gamma)\\
\le&\frac{\beta_t}{2}\|\gamma_t-\gamma\|^2-\frac{\beta_t}{2}\|\gamma_t-\gamma_{t+1}\|^2-\frac{\beta_t+\tau_t}{2}\|\gamma_{t+1}-\gamma\|^2-\frac{\tau_t}{2}\|\gamma_{t+1}-\gamma_0\|^2+\frac{\tau_t}{2}\|\gamma-\gamma_0\|^2.
\end{aligned}$$

Substituting the above inequality into (B.6), and using (B.8) and (B.9), we see that

$$\begin{aligned}
\Delta_\gamma^{t+1}\le&\frac{\beta_t}{2}\|\gamma_t-\gamma\|^2-\frac{\beta_t+\tau_t}{2}\|\gamma_{t+1}-\gamma\|^2-\frac{\tau_t}{2}\|\gamma_{t+1}-\gamma_0\|^2+\frac{\tau_t}{2}\|\gamma-\gamma_0\|^2\\
&+\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma-\gamma_t)+\frac{\|H(x_t)-h(x_t,\xi_t^1)\|^2}{2\beta_t}\\
&+\frac{3\|\gamma\|^2C_h^2}{2\eta_t}+\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2.
\end{aligned}$$

Summing the above inequality over $t=0,1,\dots,K-1$ and noting that $\beta_{t+1}\le\beta_t+\tau_t$, we have that

$$\begin{aligned}
&\sum_{t=0}^{K-1}\big(\Delta_\gamma^{t+1}-\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2\big)+\frac{\beta_K}{2}\|\gamma_K-\gamma\|^2\\
\le&(\frac{\beta_0}{2}+\sum_{t=0}^{K-1}\frac{\tau_t}{2})\|\gamma_0-\gamma\|^2-\sum_{t=0}^{K-1}\frac{\tau_t}{2}\|\gamma_{t+1}-\gamma_0\|^2+\sum_{t=0}^{K-1}\big(H(x_t)-h(x_t,\xi_t^1)\big)^\top(\gamma-\gamma_t)\\
&+\sum_{t=0}^{K-1}\frac{\|H(x_t)-h(x_t,\xi_t^1)\|^2}{2\beta_t}+\sum_{t=0}^{K-1}\frac{3\|\gamma\|^2C_h^2}{2\eta_t},\quad\forall\gamma\in\mathbb{R}_+^{m_1}.
\end{aligned}\tag{C.2}$$

It then holds from (C.1) and (C.2) that

$$\begin{aligned}
&\sum_{t=0}^{K-1}\big(\Delta_x^{t+1}+\Delta_\gamma^{t+1}+\frac{\eta_t}{6}\|x_{t+1}-x_t\|^2\big)+\frac{\eta_K}{2}\|x_K-x\|^2+\frac{\beta_K}{2}\|\gamma_K-\gamma\|^2\\
\le&(\frac{\eta_0}{2}+\sum_{t=0}^{K-1}\frac{\rho_t}{2})\|x_t-x\|^2+(\frac{\beta_0}{2}+\sum_{t=0}^{K-1}\frac{\tau_t}{2})\|\gamma_0-\gamma\|^2+\sum_{t=0}^{K-1}\big(\frac{3\|\gamma\|^2C_h^2}{2\eta_t}-\frac{\tau_t}{2}\|\gamma_{t+1}-\gamma_0\|^2\big)\\
&+\sum_{t=0}^{K-1}U_t(x,\gamma),\quad\forall(x,\gamma)\in\mathcal{X}\times\mathbb{R}_+^{m_1},
\end{aligned}\tag{C.3}$$

where $U_t(x,\gamma)$ is defined in (3.8).

Next, we perform similar analysis for Δ_y^{t+1} and Δ_γ^{t+1} and obtain

$$\begin{aligned}
&\sum_{t=0}^{K-1}\big(\Delta_y^{t+1}+\Delta_\lambda^{t+1}+\frac{\kappa_t}{6}\|y_{t+1}-y_t\|^2\big)+\frac{\kappa_K}{2}\|y_K-y\|^2+\frac{\alpha_K}{2}\|\lambda_K-\lambda\|^2\\
&\leq\big(\frac{\kappa_0}{2}+\sum_{t=0}^{K-1}\frac{\phi_t}{2}\big)\|y_t-y\|^2+\big(\frac{\alpha_0}{2}+\sum_{t=0}^{K-1}\frac{\nu_t}{2}\big)\|\lambda_0-\lambda\|^2+\sum_{t=0}^{K-1}\big(\frac{3\|\lambda\|^2C_g^2}{2\kappa_t}-\frac{\nu_t}{2}\|\lambda_{t+1}-\lambda_0\|^2\big)\\
&\quad+\sum_{t=0}^{K-1}V_t(y,\lambda),\quad\forall\,(y,\lambda)\in\mathcal{Y}\times\mathbb{R}_+^{m_2}
\end{aligned}\tag{C.4}$$

with $V_t(y,\lambda)$ defined in (3.10). Combining (C.3) and (C.4), we obtain the deried inequality and complete the proof. ∎

C.2. Proof of Lemma 4.2

Proof: The proof here is quite similar to the one for Lemma 3.4 in Section B.4. Similar to (B.12), for all $x\in\mathcal{X}$ satisfying $\mathbb{E}[\|x\|^2]<+\infty$, we know from Lemma B.2 (c) that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=0}^{K-1}\big(\widetilde{\nabla}_xL(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)-\widetilde{\nabla}_x\mathcal{L}(x_t,y_t,\gamma_{t+1})\big)^\top x\Big]\\
&\leq\big(\frac{\eta_0}{2}+\sum_{t=0}^{K}\frac{\rho_t}{2}\big)\mathbb{E}[\|x\|^2]+\sum_{t=0}^{K-1}\frac{1}{2\eta_t}\mathbb{E}[\|\widetilde{\nabla}_xL(x_t,y_t,\gamma_{t+1},\omega_t^1,\xi_t^2)-\widetilde{\nabla}_x\mathcal{L}(x_t,y_t,\gamma_{t+1})\|^2]\\
&\leq\big(\frac{\eta_0}{2}+\sum_{t=0}^{K}\frac{\rho_t}{2}\big)\mathbb{E}[\|x\|^2]+\sum_{t=0}^{K-1}\frac{1}{\eta_t}(C_f^2+C_h^2\mathbb{E}[\|\gamma_{t+1}\|^2]).
\end{aligned}\tag{C.5}$$

The desired inequality of U_t then follows from (C.5), (B.13), (B.14) and (B.15) and the definition in (3.8). The inequality of V_t can be proved in the same way. We thus complete the proof. ∎

C.3. Proof of Theorem 4.1

Proof: From Lemmas 4.1 and 3.4, by taking expectations and setting $\lambda_0=\mathbf{0},\gamma_0=\mathbf{0}$, we obtain that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=0}^{K-1}Q(z_{t+1},z)+\frac{\eta_K}{2}\|x_K-x\|^2+\frac{\beta_K}{2}\|\gamma_K-\gamma\|^2+\frac{\kappa_K}{2}\|y_K-y\|^2+\frac{\alpha_K}{2}\|\lambda_K-\lambda\|^2\Big]\\
&\leq\mathbb{E}\Big[\big(\frac{\eta_0}{2}+\sum_{t=0}^{K-1}\frac{\rho_t}{2}\big)(\|x_0-x\|^2+\|x\|^2)\Big]+\mathbb{E}\Big[\Big(\frac{\beta_0}{2}+\sum_{t=0}^{K-1}\big(\frac{\tau_t}{2}+\frac{3C_h^2}{2\eta_t}\big)\Big)\|\gamma\|^2\Big]+\sqrt{K}\mathbb{E}[\|\gamma\|]\sigma_h\\
&\quad+\mathbb{E}\Big[\sum_{t=0}^{K-1}\big(\frac{4C_h^2}{\eta_t}-\frac{\tau_t}{2}\big)\|\gamma_{t+1}\|^2\Big]+\sum_{t=0}^{K-1}\big(\frac{4C_f^2}{\eta_t}+\frac{\sigma_h^2}{2\beta_t}\big)\\
&\quad+\mathbb{E}\Big[\big(\frac{\kappa_0}{2}+\sum_{t=0}^{K-1}\frac{\phi_t}{2}\big)(\|y_0-y\|^2+\|y\|^2)\Big]+\mathbb{E}\Big[\Big(\frac{\alpha_0}{2}+\sum_{t=0}^{K-1}\big(\frac{\nu_t}{2}+\frac{3C_g^2}{2\kappa_t}\big)\|\lambda\|^2\Big)\Big]+\sqrt{K}\mathbb{E}[\|\lambda\|]\sigma_g\\
&\quad+\mathbb{E}\Big[\sum_{t=0}^{K-1}\big(\frac{4C_g^2}{\kappa_t}-\frac{\nu_t}{2}\big)\|\gamma_{t+1}\|^2\Big]+\sum_{t=0}^{K-1}\big(\frac{4C_f^2}{\kappa_t}+\frac{\sigma_g^2}{2\alpha_t}\big)
\end{aligned}\tag{C.6}$$

for all $(x,y,\gamma,\lambda)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{R}_+^{m_1}\times\mathbb{R}_+^{m_2}$ with bounded second moments. Meanwhile, it holds from (4.3) that

$$\begin{aligned}&\frac{\eta_0}{2}+\sum_{t=0}^{K-1}\frac{\rho_t}{2}=8\sqrt{K+2},\quad \frac{\beta_0}{2}+\sum_{t=0}^{K-1}\frac{\tau_t}{2}=\frac{C_h^2\sqrt{K+1}}{2},\quad \sum_{t=0}^{K-1}\frac{1}{\eta_t}=\sum_{t=0}^{K-1}\frac{1}{\kappa_t}\le\frac{\sqrt{K+2}}{8},\\ &\sum_{t=0}^{K-1}\frac{1}{\alpha_t}\le\frac{2\sqrt{K+1}}{C_g^2},\quad \sum_{t=0}^{K-1}\frac{1}{\beta_t}\le\frac{2\sqrt{K+1}}{C_h^2},\quad \frac{4C_h^2}{\eta_t}-\frac{\tau_t}{2}\le 0,\quad \frac{4C_g^2}{\kappa_t}-\frac{\nu_t}{2}\le 0.\end{aligned}\tag{C.7}$$

The desired inequality then follows by noting (4.3), and substituting (C.7) into (C.6). ∎

C.4. Proof of Theorem 4.2

Proof: To establish the bound for the objective optimality gap, we set $\gamma=\mathbf{0}$ and $\lambda=\mathbf{0}$ in Theorem 4.1 and adopt similar analysis as in Theorem 3.3 to obtain

$$\begin{aligned}&\mathbb{E}[F(\bar{x}_N,y)-F(x,\bar{y}_N)]\\ &\le\frac{1}{N}\mathbb{E}\Big[\sum_{t=0}^{N-1}\big(\mathcal{L}(x_{t+1},y,0,\lambda_{t+1})-\mathcal{L}(x,y_{t+1},\gamma_{t+1},0)\big)\Big]\\ &\le\frac{\sqrt{N+1}}{N}\Big(\frac{\sigma_h^2}{C_h^2}+\frac{\sigma_g^2}{C_g^2}\Big)+\frac{\sqrt{N+2}}{N}\Big(8\mathbb{E}[\|x-x_0\|^2+\|x\|^2]+\frac{11C_f^2}{16}+8\mathbb{E}[\|y-y_0\|^2+\|y\|^2]+\frac{11C_f^2}{16}\Big)\end{aligned}$$

for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ with bounded second moments.

For the bound of feasibility residuals, we choose $\tilde{\gamma}=(\|\gamma^*\|_2+1)\frac{H(\bar{x}_N)_+}{\|H(\bar{x}_N)_+\|_2}$ and $\tilde{\lambda}=(\|\lambda^*\|_2+1)\frac{G(\bar{y}_N)_+}{\|G(\bar{y}_N)_+\|_2}$. Again by adopting a similar analysis to Theorem 3.3, we have from Theorem 4.1 that

$$\begin{aligned}&\mathbb{E}[\|H(\bar{x}_N)_+\|_2]+\mathbb{E}[\|G(\bar{y}_N)_+\|_2]\\ &\le\frac{1}{N}\mathbb{E}\Big[\sum_{t=0}^{N-1}\big(\mathcal{L}(x_{t+1},y^*,\tilde{\gamma},\lambda_{t+1})-\mathcal{L}(x^*,y_{t+1},\gamma_{t+1},\tilde{\lambda})\big)\Big]\\ &\le\frac{\sqrt{N+2}}{N}\Big(\frac{11C_h^2}{16}\mathbb{E}[\|\tilde{\gamma}\|^2]+\frac{11C_g^2}{16}\mathbb{E}[\|\tilde{\lambda}\|^2]\Big)+\frac{1}{\sqrt{N}}\Big(\mathbb{E}[\|\tilde{\gamma}\|]\sigma_h+\mathbb{E}[\|\tilde{\lambda}\|]\sigma_g\Big)+\frac{\sqrt{N+1}}{N}\Big(\frac{\sigma_h^2}{C_h^2}+\frac{\sigma_g^2}{C_g^2}\Big)\\ &\quad+\frac{\sqrt{N+2}}{N}\Big(8\|x^*-x_0\|^2+8\|x^*\|^2+\frac{11C_f^2}{16}+8\|y^*-y_0\|^2+8\|y^*\|^2+\frac{11C_f^2}{16}\Big).\end{aligned}$$

Noting that $\|\tilde{\gamma}\|=\|\gamma^*\|+1$ and $\|\tilde{\lambda}\|=\|\lambda^*\|+1$, we arrive at the desired inequality.

The above two inequalities, together with Lemma 2.1, imply that there exist constants $C_1,C_2>0$ such that

$$-\frac{C_1}{\sqrt{N}}\le\mathbb{E}\Big[F(\bar{x}_N,y^*)-F(x^*,\bar{y}_N)\Big]\le\frac{C_2}{\sqrt{N}}.$$

This completes the proof. ∎

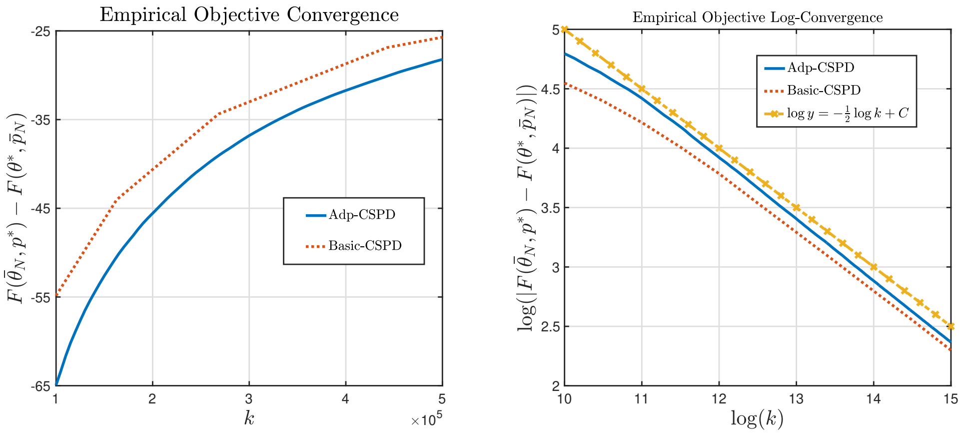

Figure 7 Empirical convergence rate of the objective gap $F(\bar{\theta}_N, p^*) - F(\theta^*, \bar{p}_N)$ for the robust optimal pricing under the normal design.

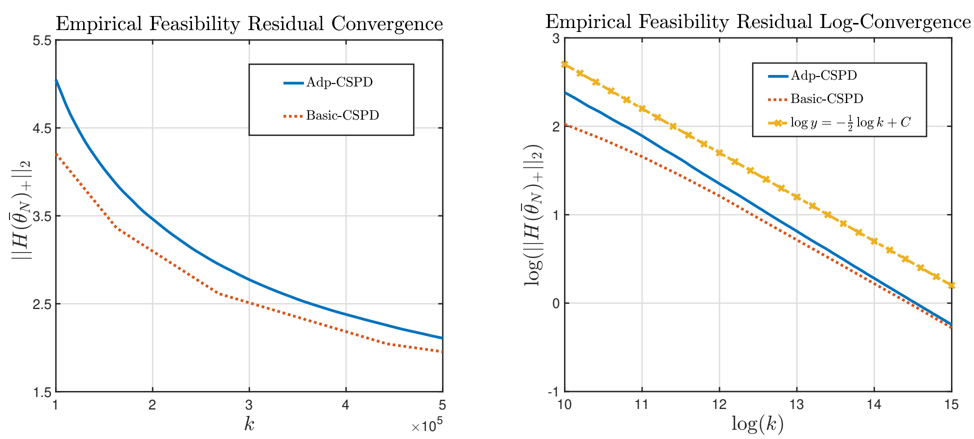

Figure 8 Empirical convergence rate of the feasibility residual $\|H(\bar{\theta}_N)_+\|_2$ for robust optimal pricing Normal Design.

D. Additional Numerical Results

Here, we conduct additional numerical experiments[1] for the robust optimal pricing problem in Section 5.2. Specifically, we consider two settings where the features $\tilde{s}_i$ are generated using normal and student distributions in the following way.

- Normal Design: For each $i = 1, \cdots, m$, each entry of the feature $\tilde{s}_i \in \mathbb{R}^d$ is independently generated from a normal distribution that $\tilde{s}_{i,j} \sim \mathcal{N}(2, 0.5)$ for $j = 1, \cdots, d$. Each entry of the feature $s \in \mathbb{R}^d$ in the objective is also independently generated under normal distribution $\mathcal{N}(2, 0.5)$.

[1] Codes and data are available at `https://github.com/MatOpt/DATA_DRIVEN_MINIMAX`.

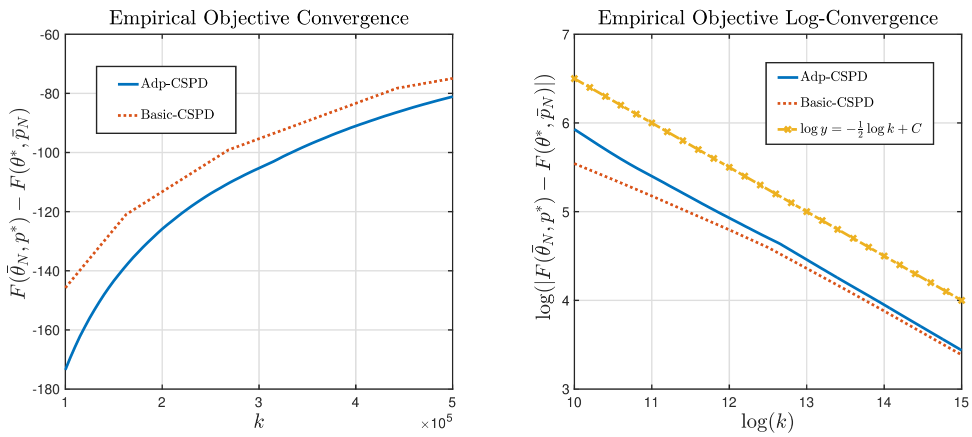

Figure 9 Empirical convergence rate of the objective gap $F(\bar{\theta}_N, p^*) - F(\theta^*, \bar{p}_N)$ for the robust optimal pricing under the Student design.

- Student Design: For each $i = 1, \cdots, m$, each entry of the feature $\tilde{s}_i \in \mathbb{R}^d$ is independently generated through $\tilde{s}_{i,j} = 2 + \nu_i$, where ν_i follows a heavy-tailed Student distribution t_4. Each entry of the feature $s \in \mathbb{R}^d$ in the objective is also generated under Student distribution similarly.

The rest parts of the simulation environment are set the same as the uniform setting in Section 5.2. To solve these problems, we run Algorithms 1 and 2 for 100 independent simulations, with the total number of iterations and stepsizes being the same as in Section 5.2. We report the numerical results for normal setup in Figures 7 and 8, and report the result for the student setup in Figures 9 and 10.

From these experiments, we observe that our Basic-CSPD and Adp-CSPD algorithms can efficiently solve the robust optimal pricing problem under various distribution settings. In addition, both the objective gap and feasibility residual converge to zero at the rate of $\mathcal{O}(1/\sqrt{N})$, matching our theoretical convergence rate claims.

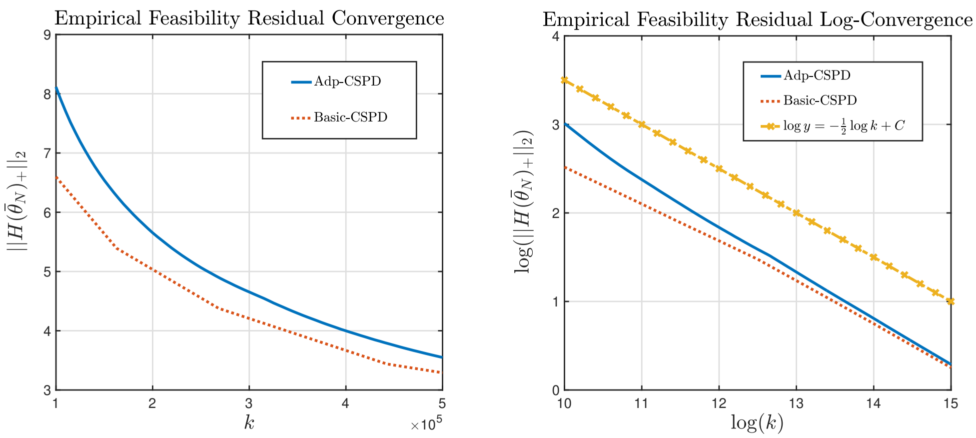

Figure 10 Empirical convergence rate of the feasibility residual $\|H(\bar{\theta}_N)_+\|_2$ for the robust optimal pricing under the Student design.