Appendix

In Section A, we present a set of general technical lemmas that are essential for proving the results in the paper. In Section B, we present the necessary lemmas to establish the proof of the result for Algorithm 1, which focuses on the utilization of the LMO for both variables. Subsequently, in section C, we analyze and demonstrate the convergence rate of Algorithm 1 in Theorem 4.2 and Corollary 4.3 for nonconvex-concave (NC-C) scenario and in Theorem 4.4 for nonconvex-strongly concave (NC-SC) scenario. Moving forward to Section D, we introduce the lemmas essential for verifying the correctness of Algorithm 2, which involves employing the LMO for the minimization variables and the PO for the maximization variable. Furthermore, in Section E, we investigate and establish the convergence rate of Algorithm 2 in Theorem 5.1 and Corollary 5.2 for NC-C scenario and in Theorem 5.3 for NC-SC scenario. Finally, in Sections F and G, details of our numerical experiment and supplementary plots are provided. To simplify the notations, we will drop the associated space from the norms unless it is not clear from the context. For instance $\|x\|_{\mathcal{X}}$ and $\|x\|_{\mathcal{X}^*}$ will be replaced by $\|x\|$ and $\|x\|_*$, respectively.

Definition .1. Let $f_\mu : X \to \mathbb{R}$ be a function such that $f_\mu(x) \triangleq \max_{y\in Y} \mathcal{L}(x,y) - \frac{\mu}{2}\|y-y_0\|^2$. Moreover, we define $y^*_\mu(x) \triangleq \operatorname{argmin}_{y\in Y} \mathcal{L}(x,y) - \frac{\mu}{2}\|y-y_0\|^2$.

A Technical Lemmas

We will now present technical lemmas that will be utilized in the proofs.

Lemma A.1. *[29] The solution map $y^*_\mu : X \to Y$ is Lipschitz continuous. In particular, for any $x, \bar{x} \in X$*

$$\left\|y^*_\mu(\bar{x}) - y^*_\mu(x)\right\| \le \frac{L_{yx}}{\mu}\|\bar{x}-x\|.$$

Proof. First note that since $\mathcal{L}_\mu(x,\cdot)$ is strongly concave for any $x \in X$, we have that

$$(y^*_\mu(\bar{x}) - y^*_\mu(x))^\top(\nabla_y\mathcal{L}_\mu(x,y^*_\mu(\bar{x})) - \nabla_y\mathcal{L}_\mu(x,y^*_\mu(x)) + \mu\left\|y^*_\mu(\bar{x}) - y^*_\mu(x)\right\|^2 \le 0. \tag{6}$$

Moreover, the optimality of $y^*_\mu(\bar{x})$ and $y^*_\mu(x)$ given that $\mathcal{L}_\mu(x,y) = \mathcal{L}(x,y) - \frac{\mu}{2}\|y-y_0\|^2$ implies that for any $y \in Y$,

$$(y - y^*_\mu(\bar{x}))^\top\nabla_y\mathcal{L}_\mu(\bar{x},y^*_\mu(\bar{x})) \le 0, \tag{7}$$

$$(y - y^*_\mu(x))^\top\nabla_y\mathcal{L}_\mu(x,y^*_\mu(x)) \le 0. \tag{8}$$

Let $y = y^*_\mu(x)$ in 7 and $y = y^*_\mu(\bar{x})$ in 8 and summing up two inequalities, we obtain

$$(y^*_\mu(x) - y^*_\mu(\bar{x}))^\top(\nabla_y\mathcal{L}_\mu(\bar{x},y^*_\mu(\bar{x})) - \nabla_y\mathcal{L}_\mu(x,y^*_\mu(x))) \le 0. \tag{9}$$

By combining 9 and 6 we have

$$\begin{aligned}\mu\left\|y^*_\mu(\bar{x}) - y^*_\mu(x)\right\|^2 &\le (y^*_\mu(\bar{x}) - y^*_\mu(x))^\top(\nabla_y\mathcal{L}_\mu(\bar{x},y^*_\mu(\bar{x})) - \nabla_y\mathcal{L}_\mu(x,y^*_\mu(\bar{x})))\\ &\overset{(a)}{\le} L_{yx}\left\|y^*_\mu(\bar{x}) - y^*_\mu(x)\right\|\|\bar{x}-x\|,\end{aligned}$$

where (a) follows from Assumption 2.6. The result follows immediately from the above inequality. □

Lemma A.2. *[29] The function $f_\mu(\cdot)$ is differentiable on an open set containing X and $\nabla f_\mu(x) = \nabla_x\mathcal{L}(x,y^*_\mu(x))$ where $y^*_\mu(x) \triangleq \operatorname{argmin}_{y\in Y}\mathcal{L}_\mu(x,y)$. Moreover, f_μ has a Lipschitz continuous gradient with constant $L_{f_\mu} \triangleq L_{xx} + L^2_{yx}/\mu$.*

Proof. From Danskins's theorem [5] one can obtain $f_\mu(\cdot)$ is differentiable and $\nabla f_\mu(x) = \nabla_x\mathcal{L}(x,y^*_\mu(x))$. Therefore, we have

$$\begin{aligned}\|\nabla f_\mu(x) - \nabla f_\mu(x')\| &= \left\|\nabla_x\mathcal{L}(x,y^*_\mu(x)) - \nabla_x\mathcal{L}(x',y^*_\mu(x'))\right\|\\ &\le L_{xx}\|x-x'\| + L_{yx}\left\|y^*_\mu(x) - y^*_\mu(x')\right\|\\ &\le L_{xx}\|x-x'\| + \frac{L^2_{yx}}{\mu}\|x-x'\|,\end{aligned}$$

where the last inequality holds by Lemma A.1.

□

B Required Lemmas for Theorems 4.2 and 4.4

Lemma B.1. *Let $\{a_k\}_{k\geq 0}$ be a sequence of non-negative real numbers such that $a_{k+1} \leq \max\{1/2, 1 - M_1\sqrt{a_k}\}a_k + M_2$ for some $M_1, M_2 > 0$ and any $k \geq 0$. Then,*

$$a_k \leq \frac{9}{(k+2)^2}\max\left\{a_0, \frac{2}{M_1^2}\right\} + \left(\frac{M_2}{M_1}\right)^{2/3} + M_2, \quad \forall k \geq 1. \tag{10}$$

Proof. We use induction to show the result. Indeed, for $k = 1$ we have that $a_1 \leq \max\{1/2, 1 - M_1\sqrt{a_0}\}a_0 + M_2 \leq a_0 + M_2$ which clearly satisfies (10). Now, suppose (10) holds for $k \geq 1$, and we show the inequality for $k + 1$. We begin by examining the recursive relation $a_{k+1} \leq \max\{1/2, 1 - M_1\sqrt{a_k}\}a_k + M_2$ and analyzing the different cases in which the maximum occurs on different terms.

(CASE I) $\max\{1/2, 1 - M_1\sqrt{a_k}\} = \frac{1}{2}$: In this case, clearly $a_{k+1} \leq \frac{1}{2}a_k \leq \frac{9}{2(k+2)^2}\max\{a_0, \frac{2}{M_1^2}\} + \frac{1}{2}((\frac{M_2}{M_1})^{2/3} + M_2) \leq \frac{9}{(k+3)^2}\max\{a_0, \frac{2}{M_1^2}\} + (\frac{M_2}{M_1})^{2/3} + M_2$ where we used the fact that $(k+3)^2 \leq 2(k+2)^2$ for any $k \geq 1$.

(CASE II.a) $\max\{1/2, 1 - M_1\sqrt{a_k}\} = 1 - M_1\sqrt{a_k}$ and $a_k \leq \frac{9}{2(k+2)^2}\max\{a_0, \frac{2}{M_1^2}\} + (\frac{M_2}{M_1})^{2/3}$: In this case, one can observe that from the recursive inequality together with the current assumption we have that $a_{k+1} \leq a_k + M_2 \leq \frac{9}{2(k+2)^2}\max\{a_0, \frac{2}{M_1^2}\} + (\frac{M_2}{M_1})^{2/3} + M_2 \leq \frac{9}{(k+3)^2}\max\{a_0, \frac{2}{M_1^2}\} + (\frac{M_2}{M_1})^{2/3} + M_2$.

(CASE II.b) $\max\{1/2, 1 - M_1\sqrt{a_k}\} = 1 - M_1\sqrt{a_k}$ and $a_k > \frac{9}{2(k+2)^2}\max\left\{a_0, \frac{2}{M_1^2}\right\} + (\frac{M_2}{M_1})^{2/3}$: Let $\Gamma_k \triangleq \frac{9}{2(k+2)^2}\max\left\{a_0, \frac{2}{M_1^2}\right\} + (\frac{M_2}{M_1})^{2/3}$. From the recursive inequality and (10) we conclude that

$$\begin{aligned} a_{k+1} &\leq (1 - M_1\sqrt{\Gamma_k})\left(\frac{9}{(k+2)^2}\max\left\{a_0, \frac{2}{M_1^2}\right\} + M_2 + (\frac{M_2}{M_1})^{2/3}\right) + M_2 \\ &= \max\left\{a_0, \frac{2}{M_1^2}\right\}\frac{9}{(k+3)^2}\left(1 + \frac{3}{k+2}\right)(1 - M_1\sqrt{\Gamma_k}) \\ &\quad + (1 - M_1\sqrt{\Gamma_k})\left(M_2 + \left(\frac{M_2}{M_1}\right)^{2/3}\right) + M_2. \end{aligned} \tag{11}$$

Next, we simplify the first two terms on the right-hand side of the above inequality by providing some upper bounds. In fact, a simple calculation reveals that $(1 - M_1\sqrt{\Gamma_k}) \leq 1 - \frac{3}{k+2}$ holds if and only if $M_1^2\Gamma_k \geq \frac{9}{(k+2)^2}$ which is true for any $k \geq 1$ since $\frac{M_1^2}{2}\frac{9}{(k+2)^2}\max\{a_0, \frac{2}{M_1^2}\} \geq \frac{9}{(k+2)^2}$. Moreover, from the fact that $\Gamma_k \geq (\frac{M_2}{M_1})^{2/3}$ for any $k \geq 1$, one can easily verify that $M_2 \leq (\frac{M_2}{M_1})^{2/3}M_1\sqrt{\Gamma_k} \leq (M_2 + (\frac{M_2}{M_1})^{2/3})M_1\sqrt{\Gamma_k}$, therefore, $(1 - M_1\sqrt{\Gamma_k})(M_2 + (\frac{M_2}{M_1})^{2/3}) \leq (\frac{M_2}{M_1})^{2/3}$ for any $k \geq 1$. Using these two inequalities within (11) and the fact that $(1 + \frac{3}{k+2})(1 - \frac{3}{k+2}) \leq 1$, we conclude that $a_{k+1} \leq \max\{a_0, \frac{2}{M_1^2}\}\frac{9}{(k+3)^2} + (\frac{M_2}{M_1})^{2/3} + M_2$ which completes the induction and henceforth the result of the lemma. □

In the following, we provide the proof of Lemma 4.1 which offers an upper bound on the decrease of $\mathcal{L}_\mu(x, y_\mu^*(x)) - \mathcal{L}_\mu(x, y)$ based on the consecutive iterates.

Proof of Lemma 4.1 Let $u_k \triangleq \frac{1}{2}(y_k + p_k) + \frac{\alpha}{8}\|y_k - p_k\|^2 v_k$ where $v_k \in \operatorname{argmax}_{\|v\|\leq 1}\langle\nabla_y\mathcal{L}_\mu(x_k, y_k), v\rangle$. From the definition of the conjugate norm, one can verify that $\langle\nabla_y\mathcal{L}_\mu(x_k, y_k), v_k\rangle = \|\nabla_y\mathcal{L}_\mu(x_k, y_k)\|_*$. Moreover, we note that since $y_k, p_k \in Y$ and Y is α-strongly convex we have that $u_k \in Y$. Recalling that $p_k = \operatorname{argmax}_{y\in Y}\langle\nabla_y\mathcal{L}_\mu(x_k, y_k), y\rangle$ we

conclude that

$$\begin{aligned}\langle y_k - p_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle &\leq \langle y_k - u_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle \\ &= \frac{1}{2}\langle y_k - p_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*.\end{aligned} \tag{12}$$

Next, with a similar argument and using concavity of $\mathcal{L}_\mu(x,\cdot)$ for any $y \in Y$, we have that $\langle y_k - p_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle \leq \mathcal{L}_\mu(x_k, y_k) - \mathcal{L}_\mu(x_k, y^*_\mu(x_k))$. Now, recall that $H_k = \mathcal{L}_\mu(x_k, y^*_\mu(x_k)) - \mathcal{L}_\mu(x_k, y_k)$, then from (12) we obtain

$$\begin{aligned}\langle y_k - p_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle &\leq \frac{1}{2}\langle y_k - p_k, \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_* \\ &\leq \frac{1}{2}\langle y_k - y^*_\mu(x_k), \nabla_y \mathcal{L}_\mu(x_k, y_k)\rangle - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_* \\ &\leq -\frac{1}{2}H_k - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*.\end{aligned} \tag{13}$$

Now, we will show one-step progress for the update of y_{k+1}. Indeed, from Lipschitz continuity of $\nabla_y \mathcal{L}_\mu(x,\cdot)$ we have that

$$\mathcal{L}_\mu(x_k, y_k) \leq \mathcal{L}_\mu(x_k, y_{k+1}) + \sigma_k \langle \nabla_y \mathcal{L}_\mu(x_k, y_k), y_k - p_k\rangle + \frac{(L_{yy}+\mu)}{2}\sigma_k^2 \|y_k - p_k\|^2.$$

Adding $\mathcal{L}_\mu(x_k, y^*_\mu(x_k))$ to both sides of the above inequality, using (13), and rearranging the terms lead to

$$\begin{aligned}\mathcal{L}_\mu(x_k, y^*_\mu(x_k)) - \mathcal{L}_\mu(x_k, y_{k+1}) &\leq \left(1 - \frac{\sigma_k}{2}\right) H_k - \sigma_k \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_* \\ &\quad + \frac{(L_{yy}+\mu)}{2}\sigma_k^2 \|y_k - p_k\|^2 \\ &\leq \max\left\{\frac{1}{2}, 1 - \frac{\alpha}{8(L_{yy}+\mu)}\|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*\right\} H_k,\end{aligned} \tag{14}$$

where the last inequality follows from the choice of step-size $\sigma_k = \min\{1, \frac{\alpha}{4(L_{yy}+\mu)}\|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*\}$.

Let us define $\gamma_k \triangleq \max\left\{\frac{1}{2}, 1 - \frac{\alpha}{8(L_{yy}+\mu)}\|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*\right\}$. We will provide an upper bound for γ_k, by lower bounding $\|\nabla_y \mathcal{L}_\mu(x_k, y_k)\|_*$ in terms of the function value using strong concavity of $\mathcal{L}_\mu$. In fact, since for any $x \in X$, we have that $y^*_\mu(x) = \operatorname{argmin}_{y\in Y} \mathcal{L}_\mu(x,y)$ then one can conclude that for any $y \in Y$, $\mathcal{L}_\mu(x, y^*_\mu(x)) - \mathcal{L}_\mu(x,y) \geq \frac{\mu}{2}\left\|y - y^*_\mu(x)\right\|^2$. Then, using concavity of $\mathcal{L}_\mu(x,\cdot)$ we obtain that

$$\begin{aligned}\mathcal{L}_\mu(x, y^*_\mu(x)) - \mathcal{L}_\mu(x,y) &\leq \left\langle \nabla_y \mathcal{L}_\mu(x,y), y^*_\mu(x) - y\right\rangle \\ &\leq \|\nabla_y \mathcal{L}_\mu(x,y)\|_* \left\|y^*_\mu(x) - y\right\| \\ &\leq \|\nabla_y \mathcal{L}_\mu(x,y)\|_* \sqrt{\frac{2}{\mu}(\mathcal{L}_\mu(x, y^*_\mu(x)) - \mathcal{L}_\mu(x,y))}.\end{aligned}$$

Therefore, $\|\nabla_y \mathcal{L}_\mu(x,y)\|_* \geq \sqrt{\frac{\mu}{2}(\mathcal{L}_\mu(x, y^*_\mu(x)) - \mathcal{L}_\mu(x,y))}$. This immediately implies that $\gamma_k \leq \max\{\frac{1}{2}, 1 - \frac{\alpha\sqrt{\mu}}{8\sqrt{2}(L_{yy}+\mu)}\sqrt{H_k}\}$. Now using this lower bound within (14) we obtain the following result.

$$\mathcal{L}_\mu(x_k, y^*_\mu(x_k)) - \mathcal{L}_\mu(x_k, y_{k+1}) \leq \max\left\{\frac{1}{2}, 1 - \frac{\alpha\sqrt{\mu}}{8\sqrt{2}(L_{yy}+\mu)}\sqrt{H_k}\right\} H_k. \tag{15}$$

The next step is to lower bound the left-hand side of the above inequality in terms of H_{k+1}. This is indeed possible by invoking Lipschitz continuity of $\nabla_x \mathcal{L}$ and the fact that $y^*_\mu(x_k) = \operatorname{argmax}_{y\in Y} \mathcal{L}_\mu(x_k, y)$. In particular, one can easily verify that $\mathcal{L}_\mu(x_k, y^*_\mu(x_k)) - \mathcal{L}_\mu(x_k, y_{k+1}) \geq$

$\mathcal{L}_\mu(x_k, y_\mu^*(x_{k+1})) - \mathcal{L}_\mu(x_k, y_{k+1})$, therefore, using Lipschitz continuity of $\nabla_x \mathcal{L}_\mu(\cdot, y)$ for any $y \in Y$, we obtain

$$\begin{aligned}
\mathcal{L}_\mu(x_k, y_\mu^*(x_k)) - \mathcal{L}_\mu(x_k, y_{k+1}) &\geq \mathcal{L}_\mu(x_{k+1}, y_\mu^*(x_{k+1})) - \mathcal{L}_\mu(x_{k+1}, y_{k+1}) \\
&\quad + \left\langle \nabla_x \mathcal{L}_\mu(x_k, y_\mu^*(x_{k+1})) - \nabla_x \mathcal{L}_\mu(x_{k+1}, y_{k+1}), x_k - x_{k+1} \right\rangle \\
&\quad - L_{xx} \left\| x_{k+1} - x_k \right\|^2 \\
&\geq \mathcal{L}_\mu(x_{k+1}, y_\mu^*(x_k)) - \mathcal{L}_\mu(x_{k+1}, y_{k+1}) \\
&\quad - \left(L_{xx} \left\| x_{k+1} - x_k \right\| + L_{yx} \left\| y_{k+1} - y_\mu^*(x_{k+1}) \right\| \right) \left\| x_{k+1} - x_k \right\| \\
&\quad - L_{xx} \left\| x_{k+1} - x_k \right\|^2 \\
&\geq H_{k+1} - L_{yx} \tau D_Y D_X - 2 L_{xx} \tau^2 D_X^2, \qquad (16)
\end{aligned}$$

where the penultimate inequality follows from Cauchy-Schwarz inequality and Lipschitz continuity of $\nabla_x \mathcal{L}_\mu(x, \cdot)$ for any $x \in X$, and the last inequality follows from the update of x_{k+1} as well as boundedness of X and Y. Finally, using the above lower bound within (15) leads to the desired result. □

C Convergence Analysis for Algorithm 1

In this section, we prove the convergence result for Algorithm 1 which includes NC-C and NC-SC scenarios.

C.1 Proof of Theorem 4.2

To show the convergence rate result, we consider implementing the result of Lemma B.1 on (5) by letting $a_k = H_k$, $M_1 = \frac{\alpha\sqrt{\mu}}{8\sqrt{2}(L_{yy}+\mu)}$, and $M_2 = \mathcal{E}(\tau)$. Therefore,

$$H_k \leq \frac{9}{(k+2)^2} \max\left\{ H_0, \frac{256(L_{yy}+\mu)^2}{\alpha^2 \mu} \right\} + \mathcal{E}(\tau) + \left(\frac{8\sqrt{2}(L_{yy}+\mu)\mathcal{E}(\tau)}{\alpha\sqrt{\mu}} \right)^{2/3}. \qquad (17)$$

Based on this inequality, we can obtain an upper bound on the distance between iterate y_k and the regularized solution $y_\mu^*(x_k)$. Subsequently, we will show the convergence results in terms of dual and primal gap functions.

In particular, we note that using strong concavity of $\mathcal{L}_\mu(x, \cdot)$ for any $x \in X$, we have that $H_k \geq \frac{\mu}{2} \left\| y_k - y_\mu^*(x_k) \right\|^2$; therefore, from (17) one can deduce that for any $k \geq 1$,

$$\begin{aligned}
\frac{\mu}{2} \left\| y_k - y_\mu^*(x_k) \right\|^2 \leq \frac{9}{(k+2)^2} \max\left\{ H_0, \frac{256(L_{yy}+\mu)^2}{\alpha^2 \mu} \right\} + \mathcal{E}(\tau) \\
+ \left(\frac{8\sqrt{2}(L_{yy}+\mu)\mathcal{E}(\tau)}{\alpha\sqrt{\mu}} \right)^{2/3}. \qquad (18)
\end{aligned}$$

Now using the fact that $\mathcal{G}_Y(x_k, y_k) = \langle \nabla_y \mathcal{L}(x_k, y_k), p_k - y_k \rangle$ one can easily verify that $\langle \nabla_y \mathcal{L}_\mu(x_k, y_k), p_k - y_k \rangle \geq \mathcal{G}_Y(x_k, y_k) - \mu D_Y^2$. Moreover, from Lipschitz continuity of $\nabla_y \mathcal{L}_\mu(x, \cdot)$ we conclude that

$$\begin{aligned}
\mathcal{G}_Y(x_k, y_k) &\leq \langle \nabla_y \mathcal{L}_\mu(x_k, y_k), p_k - y_k \rangle + \mu D_Y^2 \\
&\leq \langle \nabla_y \mathcal{L}_\mu(x_k, y_k), y^*(x_k) - y_k \rangle + \mu D_Y^2 \\
&\leq H_k + \frac{L_{yy}+\mu}{2} \left\| y_k - y^*(x_k) \right\|^2 + \mu D_Y^2 \\
&\leq (2 + \frac{L_{yy}}{\mu}) H_k + \mu D_Y^2.
\end{aligned}$$

Therefore, using (17) we conclude that for any $k \geq 1$,

$$\mathcal{G}_Y(x_k, y_k) \leq \frac{9c_\mu}{(k+2)^2} \max\left\{H_0, \frac{256(L_{yy}+\mu)^2}{\alpha^2\mu}\right\} + c_\mu \mathcal{E}(\tau) + \left(\frac{8\sqrt{2}(L_{yy}+\mu)\mathcal{E}(\tau)}{\alpha\sqrt{\mu}}\right)^{2/3} c_\mu + \mu D_Y^2, \tag{19}$$

where $c_\mu \triangleq 2 + L_{yy}/\mu$. which proves the bound for the dual gap function.

Next, we show the convergence rate result in terms of the primal gap function. Recalling that $f_\mu(x) = \min_{y \in Y} \mathcal{L}_\mu(x, y)$, from Lipschitz continuity of ∇f_μ we obtain

$$\begin{aligned}
f_\mu(x_{k+1}) &\leq f_\mu(x_k) + \langle \nabla f_\mu(x_k), x_{k+1} - x_k \rangle + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2 \\
&= f_\mu(x_k) + \langle \nabla f_\mu(x_k) - \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle + \langle \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle \\
&\quad + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2 \\
&\leq f_\mu(x_k) + L_{yx} \left\|y_k - y_\mu^*(x_k)\right\| \|x_{k+1} - x_k\| + \langle \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle \\
&\quad + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2,
\end{aligned}$$

where in the last inequality we used Lipschitz continuity of $\nabla_x \mathcal{L}(x, \cdot)$ for any $x \in X$. Using (18) within the above inequality, recalling the update of $x_{k+1} = \tau s_k + (1-\tau)x_k$, boundedness of X, and rearranging the terms we obtain

$$\begin{aligned}
\tau \langle \nabla_x \mathcal{L}(x_k, y_k), x_k - s_k \rangle &\leq f_\mu(x_k) - f_\mu(x_{k+1}) + \frac{L_{f_\mu}}{2} \tau^2 D_X^2 \\
&\quad + L_{yx} \tau D_X \sqrt{\tfrac{2}{\mu}} \Big[\frac{3}{k+2} \max\left\{\sqrt{H_0}, \frac{16(L_{yy}+\mu)}{\alpha\sqrt{\mu}}\right\} \\
&\quad + \sqrt{\mathcal{E}(\tau)} + \left(\tfrac{8\sqrt{2}(L_{yy}+\mu)\mathcal{E}(\tau)}{\alpha\sqrt{\mu}}\right)^{1/3} \Big].
\end{aligned}$$

Summing the above inequality over $k \in \mathcal{K}$ where $\mathcal{K} \triangleq \{\lceil K/2 \rceil, \ldots, K-1\}$, dividing both sides by $\tau K/2$, and noting that $\mathcal{G}_X(x_k, y_k) = \langle \nabla_x \mathcal{L}(x_k, y_k), x_k - s_k \rangle$ imply that

$$\begin{aligned}
\frac{2}{K} \sum_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k) &\leq \frac{2(f_\mu(x_0) - f_\mu(x_K))}{\tau K} + \frac{L_{f_\mu}\tau}{K} D_X^2 \\
&\quad + \frac{2\sqrt{2}L_{yx}}{\sqrt{\mu}} D_X \Big[\frac{3\log(K+1)}{K} \max\left\{\sqrt{H_0}, \frac{16(L_{yy}+\mu)}{\alpha\sqrt{\mu}}\right\} \\
&\quad + \sqrt{\mathcal{E}(\tau)} + \left(\tfrac{8\sqrt{2}(L_{yy}+\mu)\mathcal{E}(\tau)}{\alpha\sqrt{\mu}}\right)^{1/3} \Big].
\end{aligned} \tag{20}$$

Moreover, we have that $f_\mu(x_0) - f_\mu(x_K) \leq f(x_0) - f(x_K) + \frac{\mu}{2} D_Y^2 \leq f(x_0) - f(x^*) + \frac{\mu}{2} D_Y^2$. and defining $t \triangleq \operatorname{argmin}_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k)$, implies that $\frac{2}{K} \sum_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k) \geq \mathcal{G}_X(x_t, y_t)$. Therefore, (20) together with the dual bound in (19) and noting that $t \geq K/2$ leads to the desired result. □

C.2 Proof of Corollary 4.3

Note that when $\tau < 1$, then we have that $\mathcal{E}(\tau) = \mathcal{O}(L_{yx}\tau)$, hence,

$$\begin{aligned}
\mathcal{G}_X(x_t, y_t) &\leq \mathcal{O}\Big(\frac{f(x_0) - f(x^*)}{\tau K} + \frac{(L_{xx} + L_{yx}^2/\mu)\tau}{K} + \frac{\log(K) L_{yx} L_{yy}}{\mu K} + \frac{\sqrt{L_{yx}\tau}}{\sqrt{\mu}} \\
&\quad + \frac{(L_{yx} L_{yy} \tau)^{1/3}}{\mu^{2/3}} \Big), \\
\mathcal{G}_Y(x_t, y_t) &\leq \mathcal{O}\left(\frac{L_{yy}^3}{\mu^2 K^2} + \frac{L_{yy}^{5/3} \tau^{2/3}}{\mu^{4/3}} + \frac{\tau L_{yy}}{\mu} + \mu \right).
\end{aligned}$$

Minimizing the above upper bounds simultaneously in τ by considering μ as a parameter implies that $\tau = \mathcal{O}(\mu^5/L_{yx}^3)$. Then replacing τ, we can minimize the upper bounds in terms of μ which implies that $\mu = \mathcal{O}(\frac{\sqrt{L_{yx}}}{K^{1/6}})$. Therefore, we conclude that $\tau = \mathcal{O}(\frac{1}{K^{5/6}\sqrt{L_{yx}}})$ and $\mathcal{G}_Z(x_t, y_t) = \mathcal{G}_X(x_t, y_t) + \mathcal{G}_Y(x_t, y_t) \leq \mathcal{O}(1/K^{1/6})$. Therefore, an ϵ-gap solution can be computed within $\mathcal{O}(\epsilon^{-6})$ iterations by setting $\tau = \mathcal{O}(\epsilon^5)$ and $\mu = \mathcal{O}(\epsilon)$. □

C.3 Proof of Theorem 4.4

Recall that we assume $\mathcal{L}(x, \cdot)$ is $\tilde{\mu}$-strongly concave for any $x \in X$ and we set $\mu = 0$ in Algorithm 1. Following similar steps as in Lemma 4.1 one can readily obtain

$$\begin{aligned}\langle y_k - p_k, \nabla_y \mathcal{L}(x_k, y_k)\rangle &\leq \langle y_k - u_k, \nabla_y \mathcal{L}(x_k, y_k)\rangle \\ &= \frac{1}{2}\langle y_k - p_k, \nabla_y \mathcal{L}(x_k, y_k)\rangle - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}(x_k, y_k)\|_* \\ &\leq -\frac{1}{2}H_k - \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}(x_k, y_k)\|_* . \end{aligned} \tag{21}$$

Note that in this setting $y^*(x) \triangleq \operatorname{argmax}_{y\in Y} \mathcal{L}(x, y)$ is uniquely defined as satisfies the result of Lemma A.1. Moreover, from the update of y_{k+1} and Lipschitz continuity of $\nabla_y \mathcal{L}(x, \cdot)$, one can obtain

$$\mathcal{L}(x_k, y_k) \leq \mathcal{L}(x_k, y_{k+1}) + \sigma_k \langle \nabla_y \mathcal{L}(x_k, y_k), y_k - p_k\rangle + \frac{L_{yy}}{2}\sigma_k^2 \|y_k - p_k\|^2 .$$

Adding $\mathcal{L}(x_k, y^*(x_k))$ to both sides of the above inequality and rearranging the terms lead to

$$\begin{aligned}\mathcal{L}(x_k, y^*(x_k)) - \mathcal{L}(x_k, y_{k+1}) &\leq \left(1 - \frac{\sigma_k}{2}\right) H_k - \sigma_k \frac{\alpha}{8}\|y_k - p_k\|^2 \|\nabla_y \mathcal{L}(x_k, y_k)\|_* \\ &\quad + \frac{L_{yy}}{2}\sigma_k^2 \|y_k - p_k\|^2 \\ &\leq \max\left\{\frac{1}{2}, 1 - \frac{\alpha}{8L_{yy}}\|\nabla_y \mathcal{L}(x_k, y_k)\|_*\right\} H_k,\end{aligned}$$

where the last inequality follows from the choice of step-size $\sigma_k = \min\{1, \frac{\alpha}{4(L_{yy})}\|\nabla_y \mathcal{L}(x_k, y_k)\|_*\}$. Then defining $\widetilde{H}_k \triangleq \mathcal{L}(x_k, y^*(x_k)) - \mathcal{L}(x_k, y_k)$ for any $k \geq 0$, and following the same steps for proving (15), we obtain the following one-step improvement bound

$$\mathcal{L}(x_k, y^*(x_k)) - \mathcal{L}(x_k, y_{k+1}) \leq \max\left\{\frac{1}{2}, 1 - \frac{\alpha\sqrt{\tilde{\mu}}}{8\sqrt{2L_{yy}}}\sqrt{\widetilde{H}_k}\right\}\widetilde{H}_k. \tag{22}$$

Next, we can find a lower-bound for the left-hand side of the above inequality similar to (16) to conclude that

$$\widetilde{H}_{k+1} \leq \max\left\{\frac{1}{2}, 1 - \frac{\alpha\sqrt{\tilde{\mu}}}{8\sqrt{2L_{yy}}}\sqrt{\widetilde{H}_k}\right\}\widetilde{H}_k + \mathcal{E}(\tau). \tag{23}$$

Now, to show the convergence rate result, we consider implementing the result of Lemma B.1 on (23) by letting $a_k = \widetilde{H}_k$, $M_1 = \frac{\alpha\sqrt{\tilde{\mu}}}{8\sqrt{2L_{yy}}}$, and $M_2 = \mathcal{E}(\tau)$ which implies that

$$\widetilde{H}_k \leq \frac{9}{(k+2)^2}\max\left\{\widetilde{H}_0, \frac{256(L_{yy})^2}{\alpha^2\tilde{\mu}}\right\} + \mathcal{E}(\tau) + \left(\frac{8\sqrt{2}L_{yy}\mathcal{E}(\tau)}{\alpha\sqrt{\tilde{\mu}}}\right)^{2/3}. \tag{24}$$

Following similar steps as in the proof of inequality (19) and using (24) instead of (17) lead to the following bound for the dual gap function

$$\mathcal{G}_Y(x_k, y_k) \leq \frac{9c_{\tilde{\mu}}}{(k+2)^2}\max\left\{\widetilde{H}_0, \frac{256L_{yy}^2}{\alpha^2\tilde{\mu}}\right\} + c_{\tilde{\mu}}\mathcal{E}(\tau) + \left(\frac{8\sqrt{2}L_{yy}\mathcal{E}(\tau)}{\alpha\sqrt{\tilde{\mu}}}\right)^{2/3} c_{\tilde{\mu}}. \tag{25}$$

The upper bound on the primal gap function can be obtained by following the same lines as in the proof of Theorem 4.2 to obtain (20). In particular, one can show that

$$\mathcal{G}_X(x_t, y_t) \leq \frac{2}{K}\sum_{k\in\mathcal{K}} \mathcal{G}_X(x_k, y_k) \leq \frac{2(f(x_0) - f(x_K))}{\tau K} + \frac{L_f \tau}{K} D_X^2 + \frac{2\sqrt{2}L_{yx}}{\sqrt{\tilde{\mu}}} D_X \Big[\frac{3\log(K+1)}{K} \max\left\{\sqrt{\widetilde{H}_0}, \frac{16 L_{yy}}{\alpha\sqrt{\tilde{\mu}}}\right\} + \sqrt{\mathcal{E}(\tau)} + \left(\frac{8\sqrt{2}L_{yy}\mathcal{E}(\tau)}{\alpha\sqrt{\tilde{\mu}}}\right)^{1/3}\Big]. \tag{26}$$

Letting $\tau = \mathcal{O}(\frac{1}{K^{3/4}\sqrt{L_{yx}}})$ in (25) and (26) imply that $\mathcal{G}_X(x_t, y_t) \leq \mathcal{O}(1/K^{1/4})$ and $\mathcal{G}_Y(x_t, y_t) \leq \mathcal{O}(1/K^{1/2})$. Therefore, to achieve $\mathcal{G}_X(x_t, y_t) \leq \epsilon$ Algorithm 1 with $\mu = 0$ requires $\mathcal{O}(\epsilon^{-4})$ iterations while achieving $\mathcal{G}_Y(x_t, y_t) \leq \epsilon$ requires $\mathcal{O}(\epsilon^{-2})$ iterations. □

D Required Lemmas for Theorems 5.1 and 5.3

Lemma D.1. *Suppose Assumptions 2.6 and 2.7 hold and $\{(x_k, y_k)\}_{k\geq 0}$ be the sequence generated by Algorithm 2. Let $\sigma_k = \sigma \leq \frac{2}{L_{yy}+2\mu}$. Then for any $k \geq 0$,*

$$\left\|y_k - y_\mu^*(x_k)\right\| \leq \rho \left\|y_{k-1} - y_\mu^*(x_k)\right\| \tag{27}$$

where $\rho \triangleq \max\{|1 - \sigma(L_{yy} + \mu)|, |1 - \sigma\mu|\}$.

Proof. Let $\mathcal{L}_\mu(x, y) \triangleq \mathcal{L}(x, y) - \frac{\mu}{2}\left\|y - y_0\right\|^2$ and recall that $y_\mu^*(x) = \operatorname{argmin}_{y\in Y} \mathcal{L}_\mu(x, y)$; therefore, from the optimality condition we have that $y_\mu^*(x) = \mathcal{P}_Y(y_\mu^*(x) + \sigma\nabla_y \mathcal{L}_\mu(x, y_\mu^*(x)))$ for any $x \in X$. From the update of y_k and the non-expansivity of the projection operator, we have that

$$\left\|y_k - y_\mu^*(x_k)\right\| \leq \left\|y_{k-1} + \sigma\nabla_y \mathcal{L}_\mu(x_k, y_{k-1}) - (y_\mu^*(x_k) + \sigma\nabla_y \mathcal{L}_\mu(x, y_\mu^*(x_k)))\right\|.$$

Now, let us define function $g_k : Y \to \mathbb{R}$ such that $g_k(y) \triangleq \frac{1}{2}\left\|y\right\|^2 + \sigma\mathcal{L}_\mu(x_k, y)$. Note that for any $k \geq 0$, $g_k(\cdot)$ is continuously differentiable and has a Lipschitz continuous gradient with parameter $\rho = \max\{|1 - \sigma(L_{yy} + \mu)|, |1 - \sigma\mu|\}$. Therefore, one can immediately conclude the result by noting that $\nabla g_k(y_k) = y_k + \sigma\nabla_y \mathcal{L}_\mu(x_k, y_k)$. □

Lemma D.2. *Under the premises of Lemma D.1, assume $\tau_k = \tau \geq 0$, we have*

$$\left\|y_k - y_\mu^*(x_k)\right\| \leq \rho^k \left\|y_0 - y_\mu^*(x_0)\right\| + \frac{L_{yx}\rho}{\mu(1-\rho)}\tau D_X.$$

Proof. From Lemma D.1 we have that $\left\|y_k - y_\mu^*(x_k)\right\| \leq \rho\left\|y_{k-1} - y_\mu^*(x_k)\right\|$. Using the triangle inequality we conclude that

$$\begin{aligned}
\left\|y_k - y_\mu^*(x_k)\right\| &\leq \rho\Big(\left\|y_{k-1} - y_\mu^*(x_{k-1})\right\| + \left\|y_\mu^*(x_k) - y_\mu^*(x_{k-1})\right\|\Big) \\
&\overset{(a)}{\leq} \rho\left(\left\|y_{k-1} - y_\mu^*(x_{k-1})\right\| + \frac{L_{yx}}{\mu}\left\|x_k - x_{k-1}\right\|\right) \\
&\overset{(b)}{\leq} \rho\left(\left\|y_{k-1} - y_\mu^*(x_{k-1})\right\| + \frac{L_{yx}}{\mu}\tau D_X\right) \\
&\overset{(c)}{\leq} \rho^k\left\|y_0 - y_\mu^*(x_0)\right\| + \frac{L_{yx}\rho}{\mu(1-\rho)}\tau D_X,
\end{aligned} \tag{28}$$

where (a) follows from Lemma A.1; (b) follows from the assumption that X is a bounded set with diameter $D \geq 0$, and (c) is derived from Lemma D.1. □

E Convergence Analysis for Algorithm 2

In this section, we prove the convergence result for Algorithm 2 which includes NC-C and NC-SC scenarios.

E.1 Proof of Theorem 5.1

From Lipschitz continuity of ∇f_μ we have that

$$\begin{aligned}
f_\mu(x_{k+1}) &\leq f_\mu(x_k) + \langle \nabla f_\mu(x_k), x_{k+1} - x_k \rangle + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2 \\
&= f_\mu(x_k) + \langle \nabla f_\mu(x_k) - \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle + \langle \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle \\
&\quad + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2 \\
&\leq f_\mu(x_k) + L_{yx} \left\|y_k - y_\mu^*(x_k)\right\| \|x_{k+1} - x_k\| + \langle \nabla_x \mathcal{L}(x_k, y_k), x_{k+1} - x_k \rangle \\
&\quad + \frac{L_{f_\mu}}{2} \|x_{k+1} - x_k\|^2,
\end{aligned}$$

where in the last inequality we used Lipschitz continuity of $\nabla_x \mathcal{L}(x, \cdot)$ for any $x \in X$. Next, using Lemma (D.2) in the above inequality, recalling the update of $x_{k+1} = \tau s_k + (1-\tau) x_k$, and rearranging the terms we obtain

$$\begin{aligned}
\tau \langle \nabla_x \mathcal{L}(x_k, y_k), x_k - s_k \rangle &\leq f_\mu(x_k) - f_\mu(x_{k+1}) + L_{yx} \tau D_X \rho^k \left\|y_0 - y_\mu^*(x_0)\right\| \\
&\quad + \left(\frac{L_{yx}^2 \rho}{\mu(1-\rho)} + \frac{L_{xx}}{2} + \frac{L_{yx}^2}{2\mu} \right) \tau^2 D_X^2.
\end{aligned}$$

Summing the above inequality over $k \in \mathcal{K}$ where $\mathcal{K} \triangleq \{\lceil K/2 \rceil, \ldots, K-1\}$, dividing both sides by $\tau K/2$, and defining $\mathcal{G}_X(x_k, y_k) \triangleq \langle \nabla_x \mathcal{L}(x_k, y_k), x_k - s_k \rangle$ imply that

$$\begin{aligned}
\frac{2}{K} \sum_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k) &\leq \frac{2(f_\mu(x_0) - f_\mu(x_K))}{\tau K} + \frac{2L_{yx}}{(1-\rho)K} D_X \left\|y_0 - y_\mu^*(x_0)\right\| \\
&\quad + \left(\frac{2L_{yx}^2 \rho}{\mu(1-\rho)} + L_{xx} + \frac{L_{yx}^2}{\mu} \right) \tau D_X^2. \qquad (29)
\end{aligned}$$

Note that $f_\mu(x_0) - f_\mu(x_K) \leq f(x_0) - f(x_K) + \frac{\mu}{2} D_Y^2 \leq f(x_0) - f(x^*) + \frac{\mu}{2} D_Y^2$. Finally, we let $t \triangleq \operatorname{argmin}_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k)$, then $\frac{2}{K} \sum_{k \in \mathcal{K}} \mathcal{G}_X(x_k, y_k) \geq \mathcal{G}_X(x_t, y_t)$. Therefore, (29) leads to the bound on the primal gap function.

Next, we obtain an upper-bound for the dual gap function $\mathcal{G}_Y(x_k, y_k) = \frac{1}{\sigma} \|y_k - \mathcal{P}_Y(y_k + \nabla_y \mathcal{L}(x_k, y_k))\|$. In particular, using the triangle inequality we have that for any $k \geq 0$,

$$\begin{aligned}
\mathcal{G}_Y(x_k, y_k) &\leq \frac{1}{\sigma} \Big(\|y_k - \mathcal{P}_Y(y_k + \sigma \nabla_y \mathcal{L}(x_k, y_k))\| \\
&\quad + \|\mathcal{P}_Y(y_k + \sigma \nabla_y \mathcal{L}_\mu(x_k, y_k)) - \mathcal{P}_Y(y_k + \sigma \nabla_y \mathcal{L}(x_k, y_k))\| \Big) \\
&\leq \frac{1}{\sigma} \left(\|y_k - y_{k+1}\| + \sigma \|\nabla_y \mathcal{L}_\mu(x_k, y_k)) - \nabla_y \mathcal{L}(x_k, y_k)\| \right) \\
&\leq \frac{1}{\sigma} \|y_k - y_{k+1}\| + \mu D_Y, \qquad (30)
\end{aligned}$$

where the second inequality follows from non-expansivity of the projection mapping and the last inequality follows from the definition of $\nabla_y \mathcal{L}_\mu$ and boundedness of set Y.

On the other hand, using the triangle inequality one can observe that for any $k \geq 0$,

$$\begin{aligned}
\|y_{k+1} - y_k\| &\leq \left\|y_{k+1} - y_\mu^*(x_{k+1})\right\| + \left\|y_\mu^*(x_k) - y_k\right\| + \left\|y_\mu^*(x_{k+1}) - y_\mu^*(x_k)\right\| \\
&\leq 2\rho^k \left\|y_0 - y_\mu^*(x_0)\right\| + \frac{2L_{yx}\rho}{\mu(1-\rho)} \tau D_X + \frac{L_{yx}}{\mu} \tau D_X, \qquad (31)
\end{aligned}$$

where the last inequality follows from Lemma A.1 and D.2. Finally, the desired result follows from plugging (31) in (30) evaluated at $k = t$, and noting that $\rho^t \leq \rho^{K/2}$. □

E.2 Proof of Corollary 5.2

Theorem 5.1 implies that

$$\mathcal{G}_X(x_t, y_t) \leq \mathcal{O}\Big(\frac{1}{\tau K} + \frac{L_{yx}}{(1-\rho)K} + \Big(\frac{L_{yx}^2\rho}{\mu(1-\rho)} + L_{xx} + \frac{L_{yx}^2}{\mu}\Big)\tau\Big), \tag{32}$$

$$\mathcal{G}_Y(x_t, y_t) \leq \mathcal{O}\Big(\frac{\rho^K}{\sigma} + \frac{L_{yx}\rho}{\sigma\mu(1-\rho)}\tau + \frac{L_{yx}}{\sigma\mu}\tau + \mu\Big). \tag{33}$$

Selecting $\tau = \mathcal{O}(1/K^{3/4})$ and $\mu = \mathcal{O}(1/K^{1/4})$, together with the fact that $\rho^K \leq (1-\sigma\mu)^K \leq \exp(-\sigma\mu K) = \mathcal{O}(\exp(-K^{3/4}))$ implies that $\mathcal{G}_Z(x_t, y_t) = \mathcal{G}_X(x_t, y_t) + \mathcal{G}_Y(x_t, y_t) \leq \mathcal{O}(1/K^{1/4})$. Therefore, to achieve $\mathcal{G}_Z(x_t, y_t) \leq \epsilon$ Algorithm 2 requires $\mathcal{O}(\epsilon^{-4})$ iterations. □

E.3 Proof of Theorem 5.3

The proof follows the same steps as in the proof of Theorem 5.1. First, one needs to note that since $\mathcal{L}(x, \cdot)$ is $\tilde{\mu}$-strongly concave for any $x \in X$, $y^*(x) = \operatorname{argmax}_{y\in Y} \mathcal{L}(x, y)$ is uniquely defined for any $x \in X$. Therefore, the result in Lemma D.2 will be modified as follows

$$\|y_k - y^*(x_k)\| \leq \tilde{\rho}^k \|y_0 - y^*(x_0)\| + \frac{L_{yx}\tilde{\rho}}{\tilde{\mu}(1-\tilde{\rho})}\tau D_X,$$

where $\tilde{\rho} \triangleq \max\{|1 - \sigma L_{yy}|, |1 - \sigma\tilde{\mu}|\}$.

Next, following the same argument for showing equation (29) and (31) we conclude that

$$\begin{aligned}\mathcal{G}_X(x_t, y_t) \leq \frac{2}{K}\sum_{k\in\mathcal{K}} \mathcal{G}_X(x_k, y_k) \leq \frac{2(f(x_0) - f(x_K))}{\tau K} + \frac{2L_{yx}}{(1-\tilde{\rho})K} D_X \|y_0 - y^*(x_0)\| \\ + \left(\frac{2L_{yx}^2\tilde{\rho}}{\tilde{\mu}(1-\tilde{\rho})} + L_{xx} + \frac{L_{yx}^2}{\tilde{\mu}}\right)\tau D_X^2,\end{aligned} \tag{34}$$

and

$$\begin{aligned}\mathcal{G}_Y(x_t, y_t) = \frac{1}{\sigma}\|y_{t+1} - y_t\| &\leq \frac{2\tilde{\rho}^t}{\sigma}\|y_0 - y^*(x_0)\| + \frac{2L_{yx}\tilde{\rho}}{\sigma\tilde{\mu}(1-\tilde{\rho})}\tau D_X + \frac{L_{yx}}{\sigma\tilde{\mu}}\tau D_X \\ &\leq \frac{2\tilde{\rho}^{K/2}}{\sigma}\|y_0 - y^*(x_0)\| + \frac{2L_{yx}\tilde{\rho}}{\sigma\tilde{\mu}(1-\tilde{\rho})}\tau D_X + \frac{L_{yx}}{\sigma\tilde{\mu}}\tau D_X.\end{aligned} \tag{35}$$

Finally, selecting $\tau = \mathcal{O}(1/K^{1/2})$ implies that $\mathcal{G}_Z(x_t, y_t) = \mathcal{G}_X(x_t, y_t) + \mathcal{G}_Y(x_t, y_t) \leq \mathcal{O}(1/K^{1/2})$. Therefore, to achieve $\mathcal{G}_Z(x_t, y_t) \leq \epsilon$ Algorithm 2 requires $\mathcal{O}(\epsilon^{-2})$ iterations. □

F Experiment Details

In this section, we provide the details of the experiment to solve Dictionary Learning problem in Example 2. In particular, we consider solving the following SP problem

$$\min_{(\mathbf{D}', \mathbf{C}')\in X} \max_{y\in[0,B]} \tfrac{1}{2n'}\|\mathbf{A}' - \mathbf{D}'\mathbf{C}'\|_F^2 + y\left(\tfrac{1}{2n}\|\mathbf{A} - \mathbf{D}'\tilde{\mathbf{C}}\|_F^2 - \delta\right),$$

where $X = \{(\mathbf{D}', \mathbf{C}') \mid \|\mathbf{C}'\|_* \leq r,\ \|d_j'\|_2 \leq 1, \forall j \in \{1, \ldots, q\}\}$.

Dataset Generation. We generate the old dataset matrix $\mathbf{A} = \mathbf{D}\mathbf{C} \in \mathbb{R}^{m\times n}$ where $\mathbf{D} \in \mathbb{R}^{m\times p}$ is generated randomly with elements drawn from the standard Gaussian distribution whose columns are scaled to have a unit ℓ_2-norm, and $\mathbf{C} = \frac{1}{\|U\|_2\|V\|_2}UV^\top$ where $U \in \mathbb{R}^{p\times l}$ and $V \in \mathbb{R}^{n\times l}$ are generated randomly with elements drawn from the standard Gaussian distribution. The matrix $\tilde{\mathbf{C}} \in \mathbb{R}^{q\times n}$ is generated by adding $q - p$ columns of zeros to $\mathbf{C}$. The new dataset $\mathbf{A}' \in \mathbb{R}^{m\times n'}$ is generated randomly with elements drawn from the standard Gaussian distribution.

Initialization. All the methods start from the same initial point $x_0 = (\mathbf{D}_0', \mathbf{C}_0')$ and $y_0 = 0$ where $\mathbf{D}'$ is generated randomly with elements drawn from the uniform distribution in $[0, 0.1]$ whose columns

are scaled to have a unit ℓ_2-norm, and $\mathbf{C}'_0 = \mathbf{0}_{q\times n'}$. For all the algorithms we set the maximum number of iterations $K = 10^3$.

Implementation Details. In this experiment, we let $n = 500$, $m = 100$, $p = 50$, $l = 5$, $q = 60$, $n' = 10^3$, $\delta = 10^{-4}$, $r = 5$, and $B = 1$. We compare the performance of our proposed methods R-PDCG (Algorithm 1) and CG-RPGA (Algorithm 2) with the Alternating Gradient Projection (AGP) algorithm presented by [49] and the Saddle Point Frank Wolfe (SPFW) algorithm proposed by [16]. Although the theoretical result for SPFW only holds in the convex-concave setting with certain assumptions, we have included it in our experiment to enable a comparison with another method that employs the LMO in both primal and dual updates. Moreover, we compare these methods in terms of the gap function defined in Definition 2.1 and infeasibility corresponding to the nonlinear constraint in (4). Since the algorithms use different oracles, to have a fair comparison we plot the performance metrics versus time (second). In Figure 3, we compared the gap versus iteration and running time of algorithms for a fixed duration.

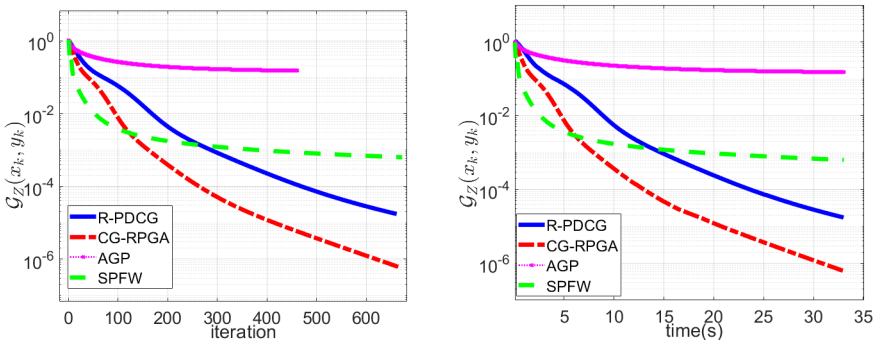
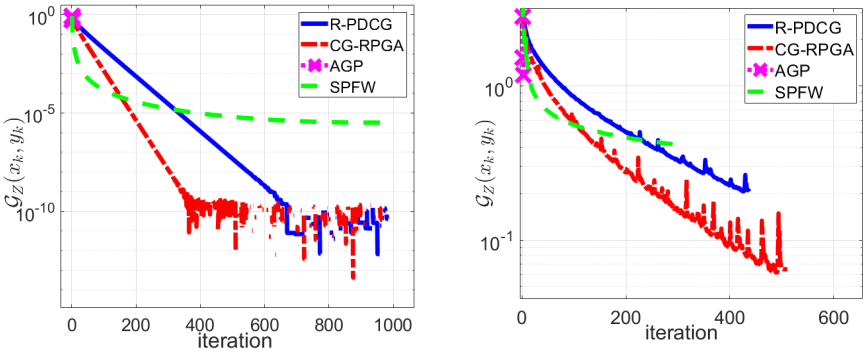


Figure 3: Comparing the performance of our proposed methods R-PDCG (blue) and CG-RPGA (red) with AGP (magenta) and SPFW (green) for a fixed amount of time in the Dictionary Learning problem.

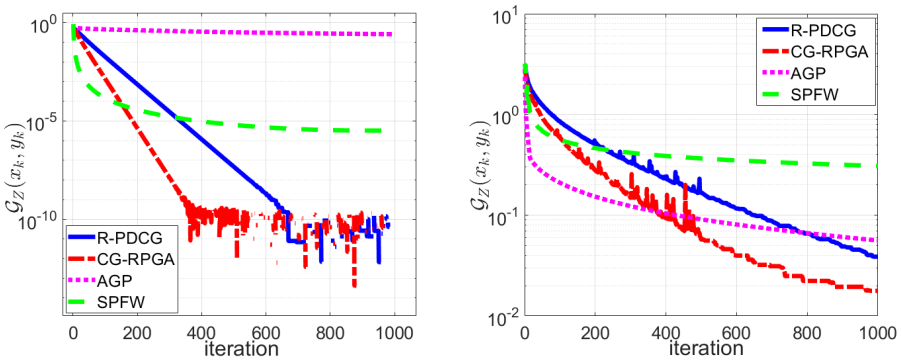
G Additional Experiments

To highlight the performance of our proposed methods for the example of Robust Multiclass Classification problem described in Example 1, we compared different methods in terms of the number of iterations. Figure 4 shows the performance of the methods in terms of the gap function versus the number of iterations within a fixed time. Notable, AGP takes only a few iterations due to its need for full SVD. In Figure 5, we conducted additional iterations of AGP to offer a more precise evaluation of its performance in comparison to other algorithms, considering a 1000-iteration count.



(a) rcv1 (b) news20

Figure 4: Comparing the performance of our proposed methods R-PDCG (blue) and CG-RPGA (red) with AGP (magenta) and SPFW (green) in the Robust Multiclass Classification problem.



(a) rcv1 (b) news20

Figure 5: Comparing the performance of our proposed methods R-PDCG (blue) and CG-RPGA (red) with AGP (magenta) and SPFW (green) in the Robust Multiclass Classification problem considering a fixed iteration count.