APPENDIX

In section A, we establish technical lemmas based on the assumptions considered in the paper. These lemmas characterize important properties of problem 1. Notably, Lemma 1 is instrumental in understanding the properties associated with problem 1. Moving on to section B, we present a series of lemmas essential for deriving the rate results of the proposed algorithm. Among them, Lemma 2 quantifies the error between the approximated direction F_k and $\nabla \ell(\mathbf{x}_k)$. This quantification plays a crucial role in establishing the one-step improvement lemma (see Lemma 7). Next, we provide the proofs of Theorem 1 and Corollary 1 in sections C and D, respectively, that support the results presented in the paper for the convex scenario. Finally, in sections E and F we provide the proofs for Theorem 2 along with Corollary 2 for the nonconvex scenario.

A SUPPORTING LEMMAS

In this section, we provide detailed explanations and proofs for the lemmas supporting the main results of the paper.

A.1 PROOF OF LEMMA 1

(I) Recall that $\mathbf{y}^*(\mathbf{x})$ is the minimizer of the lower-level problem whose objective function is strongly convex, therefore,

$$\begin{aligned}\mu_g \left\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\right\|^2 &\leq \langle \nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_y g(\mathbf{x}, \mathbf{y}^*(\bar{\mathbf{x}})), \mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\rangle \\ &= \langle \nabla_y g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}})) - \nabla_y g(\mathbf{x}, \mathbf{y}^*(\bar{\mathbf{x}})), \mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\rangle\end{aligned}$$

Note that $\nabla_y g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) = \nabla_y g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}})) = 0$. Using the Cauchy-Schwartz inequality we have

$$\begin{aligned}\mu_g \left\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\right\|^2 &\leq \|\nabla_y g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}})) - \nabla_y g(\mathbf{x}, \mathbf{y}^*(\bar{\mathbf{x}}))\| \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\| \\ &\leq C_{yx}^g \|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\|\end{aligned}$$

where the last inequality is obtained by using the Assumption 2. Therefore, we conclude that $\mu_g \left\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\right\| \leq C_{yx}^g \|\mathbf{x} - \bar{\mathbf{x}}\|$ which leads to the desired result in part (I).

(II) We first show that the function $\mathbf{x} \mapsto \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ is Lipschitz continuous. To see this, note that for any $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$, we have

$$\begin{aligned}\|\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_y f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\| &\leq L_{yx}^f \|\mathbf{x} - \bar{\mathbf{x}}\| + L_{yy}^f \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\| \\ &\leq \Big(L_{yx}^f + \frac{L_{yy}^f C_{yx}^g}{\mu_g}\Big) \|\mathbf{x} - \bar{\mathbf{x}}\|,\end{aligned}$$

where in the last inequality we used Lemma 1-(I). Since $\mathcal{X}$ is bounded, we also have $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq D_{\mathcal{X}}$. Therefore, letting $\bar{\mathbf{x}} = \mathbf{x}^*$ in the above inequality and using the triangle inequality, we have

$$\|\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \Big(L_{yx}^f + \frac{L_{yy}^f C_{yx}^g}{\mu_g}\Big) D_{\mathcal{X}} + \|\nabla_y f(\mathbf{x}^*, \mathbf{y}^*(\mathbf{x}^*))\|.$$

Thus, we complete the proof by letting $C_y^f = \Big(L_{yx}^f + \frac{L_{yy}^f C_{yx}^g}{\mu_g}\Big) D_{\mathcal{X}} + \|\nabla_y f(\mathbf{x}^*, \mathbf{y}^*(\mathbf{x}^*))\|$.

Before proceeding to show the result of part (III) of Lemma 1, we first establish an auxiliary lemma stated next.

Lemma 3. *Under the premises of Lemma 1, we have that for any* $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$, $\|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\bar{\mathbf{x}})\| \leq \mathbf{C}_\mathbf{v} \|\mathbf{x} - \bar{\mathbf{x}}\|$ *for some* $\mathbf{C}_\mathbf{v} \geq 0$.

Proof. We start the proof by recalling that $\mathbf{v}(\mathbf{x}) = \nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))^{-1} \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$. Next, adding and subtracting $\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \nabla_y f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))$ followed by a triangle inequality leads to,

$$
\begin{aligned}
&\|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\bar{\mathbf{x}})\| \\
&= \|[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1} \nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - [\nabla^2_{yy} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))]^{-1} \nabla_y f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\| \\
&\le \|[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1} \big(\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_y f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\big)\| + \|\big([\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1} \\
&- [\nabla^2_{yy} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))]^{-1}\big) \nabla_y f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\| \\
&\le \frac{1}{\mu_g}\big(L^f_{yx}\|\mathbf{x} - \bar{\mathbf{x}}\| + L^f_{yy}\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\|\big) + C^f_y \|[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1} - [\nabla^2_{yy} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))]^{-1}\|,
\end{aligned}
\tag{13}
$$

where in the last inequality we used Assumptions 1 and 2-(iii) along with the premises of Lemma 1-(II). Moreover, for any invertible matrices H_1 and H_2, we have that

$$
\|H_2^{-1} - H_1^{-1}\| = \|H_1^{-1}\big(H_1 - H_2\big)H_2^{-1}\| \le \|H_1^{-1}\|\|H_2^{-1}\|\|H_1 - H_2\|. \tag{14}
$$

Therefore, using the result of Lemma 1-(I) and (14) we can further bound inequality (13) as follows,

$$
\begin{aligned}
&\|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\bar{\mathbf{x}})\| \\
&\le \frac{1}{\mu_g}\big(L^f_{yx}\|\mathbf{x} - \bar{\mathbf{x}}\| + L^f_{yy}\mathbf{L}_{\mathbf{y}}\|\mathbf{x} - \bar{\mathbf{x}}\|\big) + C^f_y \|[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1} - [\nabla^2_{yy} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))]^{-1}\| \\
&\le \frac{1}{\mu_g}\big(L^f_{yx} + L^f_{yy}\mathbf{L}_{\mathbf{y}}\big)\|\mathbf{x} - \bar{\mathbf{x}}\| + \frac{C^f_y}{\mu_g^2} L^g_{yy}\big(\|\mathbf{x} - \bar{\mathbf{x}}\| + \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\|\big) \\
&= \big(\frac{L^f_{yx} + L^f_{yy}\mathbf{L}_{\mathbf{y}}}{\mu_g} + \frac{C^f_y L^g_{yy}}{\mu_g^2}(1 + \mathbf{L}_{\mathbf{y}})\big)\|\mathbf{x} - \bar{\mathbf{x}}\|.
\end{aligned}
$$

The result follows by letting $\mathbf{C}_{\mathbf{v}} = \frac{L^f_{yx} + L^f_{yy}\mathbf{L}_{\mathbf{y}}}{\mu_g} + \frac{C^f_y L^g_{yy}}{\mu_g^2}(1 + \mathbf{L}_{\mathbf{y}})$. □

(III) We start proving this part using the definition of $\nabla \ell(\mathbf{x})$ stated in (7a). Utilizing the triangle inequality we obtain

$$
\begin{aligned}
&\|\nabla \ell(\mathbf{x}) - \nabla \ell(\bar{\mathbf{x}})\| \\
&\quad = \|\nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla^2_{yx} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\mathbf{v}(\mathbf{x}) - \big(\nabla_x f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}})) - \nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\mathbf{v}(\bar{\mathbf{x}})\big)\| \\
&\quad \le \|\nabla_x f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - \nabla_x f(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\| + \|\big[\nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\mathbf{v}(\bar{\mathbf{x}}) - \nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\mathbf{v}(\mathbf{x})\big] \\
&\qquad + \big[\nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\mathbf{v}(\mathbf{x}) - \nabla^2_{yx} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\mathbf{v}(\mathbf{x})\big]\|
\end{aligned}
\tag{15}
$$

where the second term of the RHS follows from adding and subtracting the term $\nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}}))\mathbf{v}(\mathbf{x})$. Next, from Assumptions 1-(i) and 2-(v) together with the triangle inequality application we conclude that

$$
\begin{aligned}
\|\nabla \ell(\mathbf{x}) - \nabla \ell(\bar{\mathbf{x}})\| \le{}& L^f_{xx}\|\mathbf{x} - \bar{\mathbf{x}}\| + L^f_{xy}\|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\| + C^g_{yx}\|\mathbf{v}(\bar{\mathbf{x}}) - \mathbf{v}(\mathbf{x})\| \\
&+ \frac{C^f_y}{\mu_g}\|\nabla^2_{yx} g(\bar{\mathbf{x}}, \mathbf{y}^*(\bar{\mathbf{x}})) - \nabla^2_{yx} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\|
\end{aligned}
\tag{16}
$$

It should be that in the last inequality, we use the fact that $\|\mathbf{v}(\mathbf{x})\| = \|[\nabla^2_{yy} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1}\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \le \frac{C^f_y}{\mu_g}$. Combining the result of Lemma 1 part (I) and (II) with the Assumption 2-(iv) leads to

$$
\begin{aligned}
\|\nabla \ell(\mathbf{x}) - \nabla \ell(\bar{\mathbf{x}})\| &\le L^f_{xx}\|\mathbf{x} - \bar{\mathbf{x}}\| + L^f_{xy}\mathbf{L}_{\mathbf{y}}\|\mathbf{x} - \bar{\mathbf{x}}\| + C^g_{yx}\mathbf{C}_{\mathbf{v}}\|\mathbf{x} - \bar{\mathbf{x}}\| \\
&\quad + \frac{C^f_y}{\mu_g} L^g_{yx}\big(\|\mathbf{x} - \bar{\mathbf{x}}\| + \|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*(\bar{\mathbf{x}})\|\big) \\
&\le L^f_{xx}\|\mathbf{x} - \bar{\mathbf{x}}\| + L^f_{xy}\mathbf{L}_{\mathbf{y}}\|\mathbf{x} - \bar{\mathbf{x}}\| + C^g_{yx}\mathbf{C}_{\mathbf{v}}\|\mathbf{x} - \bar{\mathbf{x}}\| \\
&\quad + \frac{C^f_y}{\mu_g} L^g_{yx}\big(\|\mathbf{x} - \bar{\mathbf{x}}\| + \mathbf{L}_{\mathbf{y}}\|\mathbf{x} - \bar{\mathbf{x}}\|\big) \\
&\le \big(L^f_{xx} + L^f_{xy}\mathbf{L}_{\mathbf{y}} + C^g_{yx}\mathbf{C}_{\mathbf{v}} + \frac{C^f_y}{\mu_g} L^g_{yx}(1 + \mathbf{L}_{\mathbf{y}})\big)\|\mathbf{x} - \bar{\mathbf{x}}\|
\end{aligned}
\tag{17}
$$

The desired result can be obtained by letting $\mathbf{L}_\ell = L^f_{xx} + L^f_{xy}\mathbf{L}_\mathbf{y} + C^g_{yx}\mathbf{C}_\mathbf{v} + \frac{C^f_y}{\mu_g} L^g_{yx}(1 + \mathbf{L}_\mathbf{y})$. $\square$

B REQUIRED LEMMAS FOR THEOREMS 1 AND 2

Before we proceed to the proofs of Theorems 1 and 2, we present the following technical lemmas which quantify the error between the approximated solution $\mathbf{y}_k$ and $\mathbf{y}^*(\mathbf{x}_k)$, as well as between $\mathbf{w}_{k+1}$ and $\mathbf{v}(\mathbf{x}_k)$.

Lemma 4. *Suppose Assumption 2 holds. Let $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1, such that $\alpha = 2/(\mu_g + L_g)$. Then, for any $k \geq 0$*

$$\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| \leq \beta^k \|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{x}_0)\| + \mathbf{L}_\mathbf{y} D_\mathcal{X} \sum_{i=0}^{k-1} \gamma_i \beta^{k-i}, \tag{18}$$

where $\beta \triangleq (L_g - \mu_g)/(L_g + \mu_g)$.

Proof. We begin the proof by characterizing the one-step progress of the lower-level iterate sequence $\{\mathbf{y}_k\}_k$. Indeed, at iteration k we aim to approximate $\mathbf{y}^*(\mathbf{x}_{k+1}) = \operatorname{argmin}_\mathbf{y} g(\mathbf{x}_{k+1}, \mathbf{y})$. According to the update of $\mathbf{y}_{k+1}$ we observe that

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 &= \|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1}) - \alpha\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k)\|^2 \\ &= \|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 - 2\alpha\langle\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k), \mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\rangle \\ &\quad + \alpha^2\|\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k)\|^2. \end{aligned}\tag{19}$$

Moreover, from Assumption 2 and following Theorem 2.1.12 in (Nesterov, 2018), we have that

$$\langle\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k), \mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\rangle \geq \frac{\mu_g L_g}{\mu_g + L_g}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 + \frac{1}{\mu_g + L_g}\|\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k)\|^2 \tag{20}$$

The inequality in (19) together with (20) imply that

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 &\leq \|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 - \frac{2\alpha\mu_g L_g}{\mu_g + L_g}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 \\ &\quad + \Big(\alpha^2 - \frac{2\alpha}{\mu_g + L_g}\Big)\|\nabla_y g(\mathbf{x}_{k+1}, \mathbf{y}_k)\|^2. \end{aligned}\tag{21}$$

Setting the step-size $\alpha = \frac{2}{\mu_g + L_g}$ in (21) leads to

$$\|\mathbf{y}_{k+1} - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 \leq \Big(\frac{\mu_g - L_g}{\mu_g + L_g}\Big)^2\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\|^2 \tag{22}$$

Next, recall that $\beta = (L_g - \mu_g)/(L_g + \mu_g)$. Using the triangle inequality and Part (I) of Lemma 1 we conclude that

$$\begin{aligned}\|\mathbf{y}_{k+1} - \mathbf{y}^*(\mathbf{x}_{k+1})\| &\leq \beta\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_{k+1})\| \\ &\leq \beta\Big[\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| + \|\mathbf{y}^*(\mathbf{x}_k) - \mathbf{y}^*(\mathbf{x}_{k+1})\|\Big] \\ &\leq \beta\Big[\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| + \mathbf{L}_\mathbf{y}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|\Big]. \end{aligned}\tag{23}$$

Moreover, from the update of $\mathbf{x}_{k+1}$ in Algorithm 1 and boundedness of $\mathcal{X}$ we have that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \gamma_k D_\mathcal{X}$. Therefore, using this inequality within (23) leads to

$$\|\mathbf{y}_{k+1} - \mathbf{y}^*(\mathbf{x}_{k+1})\| \leq \beta\,\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| + \beta\gamma_k\mathbf{L}_\mathbf{y} D_\mathcal{X}.$$

Finally, the desired result can be deduced from the above inequality recursively. $\square$

Previously, in Lemma 4 we quantified how close the approximation $\mathbf{y}_k$ is from the optimal solution $\mathbf{y}^*(\mathbf{x}_k)$ of the inner problem. Now, in the following Lemma, we will find an upper bound for the error of approximating $\mathbf{v}(\mathbf{x}_k)$ via $\mathbf{w}_{k+1}$.

Lemma 5. *Let* $\{(\mathbf{x}_k, \mathbf{w}_k)\}_{k\geq 0}$ *be the sequence generated by Algorithm 1, such that* $\gamma_k = \gamma$. *Define* $\rho_k \triangleq (1 - \eta_k \mu_g)$ *and* $\mathbf{C}_1 \triangleq L^g_{yy}\frac{C^f_y}{\mu_g} + L^f_{yy}$. *Under Assumptions 1 and 2 we have that for any* $k \geq 0$,

$$\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| \leq \rho_k \|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_{k-1})\| + \rho_k \mathbf{C}_\mathbf{v}\gamma D_\mathcal{X} + \eta_k \mathbf{C}_1\big(\beta^k D^y_0 + \mathbf{L}_\mathbf{y}\gamma\frac{\beta}{1-\beta}D_\mathcal{X}\big). \tag{24}$$

Proof. From the optimality condition of (8) one can easily verify that $\mathbf{v}(\mathbf{x}_k) = \mathbf{v}(\mathbf{x}_k) - \eta_k\big(\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\mathbf{v}(\mathbf{x}_k) - \nabla_y f(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\big)$. Now using definition of $\mathbf{w}_{k+1}$ we can write

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| &= \Big\|\Big(\mathbf{w}_k - \eta_k(\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}_k)\mathbf{w}_k - \nabla_y f(\mathbf{x}_k, \mathbf{y}_k))\Big) - \Big(\mathbf{v}(\mathbf{x}_k) \\ &\quad - \eta_k\big(\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\mathbf{v}(\mathbf{x}_k) - \nabla_y f(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\big)\Big)\Big\| \\ &= \Big\|\Big(I - \eta_k\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}_k)\Big)(\mathbf{w}_k - \mathbf{v}(\mathbf{x}_k)) - \eta_k\Big(\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}_k) \\ &\quad - \nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\Big)\mathbf{v}(\mathbf{x}_k) + \eta_k\Big(\nabla_y f(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k)) - \nabla_y f(\mathbf{x}_k, \mathbf{y}_k)\Big)\Big\|,\end{aligned} \tag{25}$$

where the last equality is obtained by adding and subtracting the term $(I - \eta_k\nabla^2_{yy}g(\mathbf{x}_k, \mathbf{y}_k))\mathbf{v}(\mathbf{x}_k)$. Next, using Assumptions 1 and 2 along with the application of the triangle inequality we obtain

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| &\leq (1 - \eta_k\mu_g)\|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_k)\| + \eta_k L^g_{yy}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\|\|\mathbf{v}(\mathbf{x}_k)\| \\ &\quad + \eta_k L^f_{yy}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\|.\end{aligned} \tag{26}$$

Note that $\|\mathbf{v}(\mathbf{x}_k)\| = \|[\nabla^2_{yy}g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))]^{-1}\nabla_y f(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))\| \leq \frac{C^f_y}{\mu_g}$. Now, by adding and subtracting $\mathbf{v}(\mathbf{x}_{k-1})$ to the term $\|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_k)\|$ followed by triangle inequality application we can conclude that

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| &\leq (1 - \eta_k\mu_g)\|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_{k-1})\| + (1 - \eta_k\mu_g)\|\mathbf{v}(\mathbf{x}_{k-1}) - \mathbf{v}(\mathbf{x}_k)\| \\ &\quad + \eta_k\Big(L^g_{yy}\frac{C^f_y}{\mu_g} + L^f_{yy}\Big)\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\|.\end{aligned} \tag{27}$$

Therefore, using the result of Lemma 4, we can further bound inequality (27) as follows

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| &\leq (1 - \eta_k\mu_g)\|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_{k-1})\| + (1 - \eta_k\mu_g)\mathbf{C}_\mathbf{v}\|\mathbf{x}_{k-1} - \mathbf{x}_k\| \\ &\quad + \eta_k\mathbf{C}_1\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| \\ &\leq \rho_k\|\mathbf{w}_k - \mathbf{v}(\mathbf{x}_{k-1})\| + \rho_k\mathbf{C}_\mathbf{v}\gamma D_\mathcal{X} + \eta_k\mathbf{C}_1\big(\beta^k D^y_0 + \mathbf{L}_\mathbf{y}\gamma\frac{\beta}{1-\beta}D_\mathcal{X}\big)\end{aligned} \tag{28}$$

where the last inequality follows from the boundedness assumption of set $\mathcal{X}$, recalling that $D^y_0 = \|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{x}_0)\|$, and the fact that $\sum_{i=0}^{k-1}\beta^{k-i} \leq \frac{\beta}{1-\beta}$. ☐

Lemma 6. *Let* $\{(\mathbf{x}_k, \mathbf{w}_k)\}_{k\geq 0}$ *be the sequence generated by Algorithm 1 with step-size* $\eta_k = \eta < \frac{1-\beta}{\mu_g}$ *where* β *is defined in Lemma 4. Suppose that Assumption 2 holds and* $\mathbf{v}(\mathbf{x}_{-1}) = \mathbf{v}(\mathbf{x}_0)$*, then for any* $K \geq 1$,

$$\|\mathbf{w}_K - \mathbf{v}(\mathbf{x}_{K-1})\| \leq \rho^K\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| + \frac{\gamma\rho\mathbf{C}_\mathbf{v}D_\mathcal{X}}{1-\rho} + \frac{\eta\mathbf{C}_1 D^y_0\rho^{K+1}}{\rho-\beta} + \frac{\gamma\eta\beta\mathbf{C}_1\mathbf{L}_\mathbf{y}D_\mathcal{X}}{(1-\rho)(1-\beta)}, \tag{29}$$

where $\rho \triangleq 1 - \eta\mu_g$.

Proof. Applying the result of Lemma 5 recursively for $k = 0$ to $K - 1$, one can conclude that

$$\begin{aligned}\|\mathbf{w}_K - \mathbf{v}(\mathbf{x}_{K-1})\| &\leq \rho^K\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| + \mathbf{C}_\mathbf{v}\gamma D_\mathcal{X}\sum_{i=1}^K\rho^i + \eta\mathbf{C}_1\sum_{i=0}^K\big(\beta^i D^y_0 + \gamma\mathbf{L}_\mathbf{y}D_\mathcal{X}\frac{\beta}{1-\beta}\big)\rho^{K-i} \\ &\leq \rho^K\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| + \frac{\rho}{1-\rho}\mathbf{C}_\mathbf{v}\gamma D_\mathcal{X} + \eta\mathbf{C}_1 D^y_0\big(\sum_{i=0}^K\beta^i\rho^{K-i}\big) \\ &\quad + \frac{\gamma\eta\beta\mathbf{C}_1\mathbf{L}_\mathbf{y}D_\mathcal{X}}{1-\beta}\sum_{i=0}^K\rho^{K-i},\end{aligned} \tag{30}$$

where the last inequality is obtained by noting that $\sum_{i=1}^{K}\rho^i \leq \frac{\rho}{1-\rho}$. Finally, the choice $\eta < \frac{1-\beta}{\mu_g}$ implies that $\beta < \rho$, hence, $\sum_{i=0}^{K}(\frac{\beta}{\rho})^i \leq \frac{\rho}{\rho-\beta}$ which leads to the desired result. $\square$

B.1 PROOF OF LEMMA 2

We begin the proof by considering the definition of $\nabla\ell(\mathbf{x}_k)$ and F_k followed by a triangle inequality to obtain

$$\begin{aligned}\|\nabla\ell(\mathbf{x}_k) - F_k\| &\leq \|\nabla_x f(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k)) - \nabla_x f(\mathbf{x}_k, \mathbf{y}_k)\| \\ &\quad + \|\nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}_k)\mathbf{w}_{k+1} - \nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\mathbf{v}(\mathbf{x}_k)\| \qquad (31)\end{aligned}$$

Combining Assumption 1-(i) together with adding and subtracting $\nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}_k)\mathbf{v}(\mathbf{x}_k)$ to the second term of RHS lead to

$$\begin{aligned}\|\nabla\ell(\mathbf{x}_k) - F_k\| &\leq L^f_{xy}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| + \|\nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}_k)\big(\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\big) + \big(\nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}_k) \\ &\qquad - \nabla^2_{yx} g(\mathbf{x}_k, \mathbf{y}^*(\mathbf{x}_k))\big)\mathbf{v}(\mathbf{x}_k)\| \\ &\leq L^f_{xy}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| + C^g_{yx}\|\mathbf{w}_{k+1} - \mathbf{v}(\mathbf{x}_k)\| + L^g_{yx}\frac{C^f_y}{\mu_g}\|\mathbf{y}_k - \mathbf{y}^*(\mathbf{x}_k)\| \qquad (32)\end{aligned}$$

where the last inequality is obtained using Assumption 2 and the triangle inequality. Next, utilizing Lemma 4 and 6 we can further provide upper-bounds for the term in RHS of (32) as follows

$$\begin{aligned}\|\nabla\ell(\mathbf{x}_k) - F_k\| &\leq \mathbf{C}_2\big(\beta^k D^y_0 + \frac{\gamma\beta\mathbf{L}_\mathbf{y}D_\mathcal{X}}{1-\beta}\big) + C^g_{yx}\Big(\rho^{k+1}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| + \frac{\gamma\rho\mathbf{C}_\mathbf{v}D_\mathcal{X}}{1-\rho} \\ &\qquad + \frac{\eta\mathbf{C}_1 D^y_0\rho^{k+2}}{\rho-\beta} + \frac{\gamma\eta\beta\mathbf{C}_1\mathbf{L}_\mathbf{y}D_\mathcal{X}}{(1-\rho)(1-\beta)}\Big).\end{aligned}$$

$\square$

B.2 IMPROVEMENT IN ONE STEP

In the following, we characterize the improvement of the objective function $\ell(\mathbf{x})$ after taking one step of Algorithm 1.

Lemma 7. *Let $\{\mathbf{x}_k\}_{k=0}^K$ be the sequence generated by Algorithm 1. Suppose Assumptions 1 and 2 hold and $\gamma_k = \gamma$, then for any $k \geq 0$ we have*

$$\begin{aligned}\ell(\mathbf{x}_{k+1}) &\leq \ell(\mathbf{x}_k) - \gamma\mathcal{G}(\mathbf{x}_k) + \gamma\mathbf{C}_2\beta^k D^y_0 D_\mathcal{X} + \frac{\gamma^2\mathbf{C}_2 D^2_\mathcal{X}\mathbf{L}_\mathbf{y}\beta}{1-\beta} + C^g_{yx}\Big[\gamma D_\mathcal{X}\rho^{k+1}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| \\ &\quad + \frac{\gamma^2 D^2_\mathcal{X}\rho\mathbf{C}_\mathbf{v}}{1-\rho} + \frac{\gamma D_\mathcal{X} D^y_0\mathbf{C}_1\eta\rho^{k+2}}{\rho-\beta} + \frac{\gamma^2 D^2_\mathcal{X}\mathbf{L}_\mathbf{y}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big] + \frac{1}{2}\mathbf{L}_\ell\gamma^2 D^2_\mathcal{X}. \qquad (33)\end{aligned}$$

Proof. Note that according to Lemma 1-(III), $\ell(\cdot)$ has a Lipschitz continuous gradient which implies that

$$\begin{aligned}\ell(\mathbf{x}_{k+1}) &\leq \ell(\mathbf{x}_k) + \gamma\langle\nabla\ell(\mathbf{x}_k), \mathbf{s}_k - \mathbf{x}_k\rangle + \frac{1}{2}\mathbf{L}_\ell\gamma^2\|\mathbf{s}_k - \mathbf{x}_k\|^2 \\ &= \ell(\mathbf{x}_k) + \gamma\langle F_k, \mathbf{s}_k - \mathbf{x}_k\rangle + \gamma\langle\nabla\ell(\mathbf{x}_k) - F_k, \mathbf{s}_k - \mathbf{x}_k\rangle + \frac{1}{2}\mathbf{L}_\ell\gamma^2\|\mathbf{s}_k - \mathbf{x}_k\|^2, \qquad (34)\end{aligned}$$

where the last inequality follows from adding and subtracting the term $\gamma\langle F_k, \mathbf{s}_k - \mathbf{x}_k\rangle$ to the RHS. Define $\mathbf{s}'_k = \operatorname{argmax}_{\mathbf{s}\in\mathcal{X}}\{\langle\nabla\ell(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}\rangle\}$ and observe that $\mathcal{G}(\mathbf{x}_k) = \langle\nabla\ell(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}'_k\rangle$ by Definition 1. Using the definition of $\mathbf{s}_k$, we can immediately observe that

$$\begin{aligned}\langle F_k, \mathbf{s}_k - \mathbf{x}_k\rangle &= \min_{\mathbf{s}\in\mathcal{X}}\langle F_k, \mathbf{s} - \mathbf{x}_k\rangle \\ &\leq \langle F_k, \mathbf{s}'_k - \mathbf{x}_k\rangle \\ &= \langle\nabla\ell(\mathbf{x}_k), \mathbf{s}'_k - \mathbf{x}_k\rangle + \langle F_k - \nabla\ell(\mathbf{x}_k), \mathbf{s}'_k - \mathbf{x}_k\rangle \\ &= -\mathcal{G}(\mathbf{x}_k) + \langle F_k - \nabla\ell(\mathbf{x}_k), \mathbf{s}'_k - \mathbf{x}_k\rangle. \qquad (35)\end{aligned}$$

Next, combining (34) with (35) followed by the Cauchy-Schwartz inequality leads to

$$\ell(\mathbf{x}_{k+1}) \leq \ell(\mathbf{x}_k) - \gamma\mathcal{G}(\mathbf{x}_k) + \gamma\|\nabla\ell(\mathbf{x}_k) - F_k\|\|\mathbf{s}_k - \mathbf{s}_k'\| + \frac{1}{2}\mathbf{L}_\ell\gamma^2\|\mathbf{s}_k - \mathbf{x}_k\|^2. \tag{36}$$

Finally, using the result of the Lemma 2 together with the boundedness assumption of set $\mathcal{X}$ we conclude the desired result. ☐

C PROOF OF THEOREM 1

Since ℓ is convex, from the definition of $\mathcal{G}(\mathbf{x}_k)$ in (4) we have

$$\mathcal{G}(\mathbf{x}_k) = \max_{\mathbf{s}\in\mathcal{X}}\{\langle\nabla\ell(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}\rangle\} \geq \langle\nabla\ell(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \geq \ell(\mathbf{x}_k) - \ell(\mathbf{x}^*). \tag{37}$$

We assume a fixed step-size in Theorem 1 and we set $\gamma_k = \gamma$. Combining the result of Lemma 7 with (37) leads to

$$\begin{aligned}\ell(\mathbf{x}_{k+1}) \leq \ell(\mathbf{x}_k) - \gamma(\ell(\mathbf{x}_k) - \ell(\mathbf{x}^*)) + \gamma\mathbf{C}_2\beta^k D_0^y D_\mathcal{X} + \frac{\gamma^2\mathbf{C}_2 D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\beta}{1-\beta} + C_{yx}^g\Big[\gamma D_\mathcal{X}\rho^{k+1}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| \\ + \frac{\gamma^2 D_\mathcal{X}^2\rho\mathbf{C}_\mathbf{v}}{1-\rho} + \frac{\gamma D_\mathcal{X} D_0^y\mathbf{C}_1\eta\rho^{k+2}}{\rho-\beta} + \frac{\gamma^2 D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big] + \frac{1}{2}\mathbf{L}_\ell\gamma^2 D_\mathcal{X}^2.\end{aligned} \tag{38}$$

Subtracting $\ell(\mathbf{x}^*)$ from both sides, we get

$$\ell(\mathbf{x}_{k+1}) - \ell(\mathbf{x}^*) \leq (1-\gamma)(\ell(\mathbf{x}_k) - \ell(\mathbf{x}^*)) + \mathcal{R}_k(\gamma), \tag{39}$$

where

$$\begin{aligned}\mathcal{R}_k(\gamma) \triangleq \gamma\mathbf{C}_2\beta^k D_0^y D_\mathcal{X} + \frac{\gamma^2\mathbf{C}_2 D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\beta}{1-\beta} + C_{yx}^g\Big[\gamma D_\mathcal{X}\rho^{k+1}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| \\ + \frac{\gamma^2 D_\mathcal{X}^2\rho\mathbf{C}_\mathbf{v}}{1-\rho} + \frac{\gamma D_\mathcal{X} D_0^y\mathbf{C}_1\eta\rho^{k+2}}{\rho-\beta} + \frac{\gamma^2 D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big] + \frac{1}{2}\mathbf{L}_\ell\gamma^2 D_\mathcal{X}^2.\end{aligned} \tag{40}$$

Continuing (39) recursively leads to the desired result. ☐

D PROOF OF COROLLARY 1

We start the proof by using the result of the Theorem 1, i.e.,

$$\ell(\mathbf{x}_K) - \ell(\mathbf{x}^*) \leq (1-\gamma)^K(\ell(\mathbf{x}_0) - \ell(\mathbf{x}^*)) + \sum_{k=0}^{K-1}(1-\gamma)^{K-k}\mathcal{R}_k(\gamma). \tag{41}$$

Note that

$$\begin{aligned}&\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\mathcal{R}_k(\gamma) \\ &= \mathbf{C}_2 D_0^y D_\mathcal{X}\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma\beta^k\Big] + \frac{\mathbf{C}_2 D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\beta}{1-\beta}\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma^2\Big] \\ &\quad + C_{yx}^g\Big(\rho D_\mathcal{X}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\|\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma\rho^k\Big] + \frac{D_\mathcal{X}^2\rho\mathbf{C}_\mathbf{v}}{1-\rho}\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma^2\Big] \\ &\quad + \frac{D_\mathcal{X} D_0^y\mathbf{C}_1\eta\rho^2}{\rho-\beta}\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma\rho^k\Big] + \frac{D_\mathcal{X}^2\mathbf{L}_\mathbf{y}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma^2\Big]\Big) \\ &\quad + \frac{1}{2}\mathbf{L}_\ell D_\mathcal{X}^2\Big[\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma^2\Big].\end{aligned}$$

Moreover, one can easily verify that $\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma^2 \leq \gamma(1-\gamma)$ and $\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\gamma\rho^k \leq \frac{\gamma(1-\gamma)}{|1-\gamma-\rho|}$ from which together with the above inequality we conclude that

$$\begin{aligned}
&\sum_{k=0}^{K-1}(1-\gamma)^{K-k}\mathcal{R}_k(\gamma) \\
&\leq \frac{\mathbf{C}_2 D_0^y D_{\mathcal{X}}\gamma(1-\gamma)}{|1-\gamma-\beta|} + \frac{\mathbf{C}_2 D_{\mathcal{X}}^2 \mathbf{L}_{\mathbf{y}}\beta\gamma(1-\gamma)}{1-\beta} + C_{yx}^g\Big(\frac{D_{\mathcal{X}}\rho\gamma(1-\gamma)}{|1-\gamma-\rho|}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| \\
&\quad + \frac{D_{\mathcal{X}}^2 \mathbf{C}_{\mathbf{v}}\rho\gamma(1-\gamma)}{1-\rho} + \frac{D_{\mathcal{X}} D_0^y \mathbf{C}_1\eta\rho^2\gamma(1-\gamma)}{(\rho-\beta)\,|1-\gamma-\rho|} + \frac{D_{\mathcal{X}}^2 \mathbf{L}_{\mathbf{y}}\mathbf{C}_1\eta\beta\gamma(1-\gamma)}{(1-\beta)(1-\rho)}\Big) + \frac{1}{2}\mathbf{L}_\ell D_{\mathcal{X}}^2\gamma(1-\gamma) \\
&= \mathcal{O}\left(\frac{\mathbf{C}_{\mathbf{v}}\rho}{1-\rho}\gamma + \frac{\mathbf{L}_{\mathbf{y}}\mathbf{C}_1\beta}{(1-\beta)(1-\rho)}\gamma\right). \qquad (42)
\end{aligned}$$

Using the above inequality within (41) we conclude that $\ell(\mathbf{x}_K) - \ell(\mathbf{x}^*) \leq (1-\gamma)^K(\ell(\mathbf{x}_0) - \ell(\mathbf{x}^*)) + \mathcal{O}(\frac{\mathbf{C}_{\mathbf{v}}\rho}{1-\rho}\gamma + \frac{\mathbf{L}_{\mathbf{y}}\mathbf{C}_1\beta}{(1-\beta)(1-\rho)}\gamma)$ where $\mathbf{C}_{\mathbf{v}} = \mathcal{O}(\kappa_g^3)$, $\mathbf{C}_1 = \mathcal{O}(\kappa_g^2)$, $\mathbf{L}_{\mathbf{y}} = \mathcal{O}(\kappa_g)$ as shown in Lemma 3 and $\min\{1-\rho, 1-\beta\} = \Omega(\frac{1}{\kappa_g})$ as shown in Lemma 2. Next, we show that by selecting $\gamma = \log(K)/K$ we have that $(1-\gamma)^K \leq 1/K$. In fact, for any $x > 0$, $\log(x) \geq 1 - \frac{1}{x}$ which implies that $\log(\frac{1}{1-\gamma}) \geq \gamma = \log(K)/K$, hence, $(\frac{1}{1-\gamma})^K \geq K$. Putting the pieces together we conclude that $\ell(\mathbf{x}_K) - \ell(\mathbf{x}^*) = \mathcal{O}((1-\gamma)^K(\ell(\mathbf{x}_0) - \ell(\mathbf{x}^*)) + \gamma\kappa_g^5) = \tilde{\mathcal{O}}(\kappa_g^5/K)$, which leads to an iteration complexity of $\tilde{\mathcal{O}}(\kappa_g^5\epsilon^{-1})$.

Furthermore, assuming that $\nabla_y f(\mathbf{x},\cdot)$ is uniformly bounded for any $\mathbf{x} \in \mathcal{X}$, we conclude that $C_y^f = \mathcal{O}(1)$, hence, $\mathbf{C}_1 = \mathcal{O}(\kappa_g)$ from which we have that $\ell(\mathbf{x}_K) - \ell(\mathbf{x}^*) = \mathcal{O}((1-\gamma)^K(\ell(\mathbf{x}_0) - \ell(\mathbf{x}^*)) + \gamma\kappa_g^4)$. Therefore, selecting $\gamma = \log(K)/K$ implies that $\ell(\mathbf{x}_K) - \ell(\mathbf{x}^*) = \mathcal{O}(\kappa_g^4/K)$ which leads to an iteration complexity of $\mathcal{O}(\kappa_g^4\epsilon^{-1})$. □

E PROOF OF THEOREM 2

Recall that from Lemma 7 we have

$$\begin{aligned}
\mathcal{G}(\mathbf{x}_k) \leq \frac{\ell(\mathbf{x}_k) - \ell(\mathbf{x}_{k+1})}{\gamma} &+ \mathbf{C}_2\beta^k D_0^y D_{\mathcal{X}} + \frac{\gamma\mathbf{C}_2 D_{\mathcal{X}}^2\mathbf{L}_{\mathbf{y}}\beta}{1-\beta} + C_{yx}^g\Big[D_{\mathcal{X}}\rho^{k+1}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\| \\
&+ \frac{\gamma D_{\mathcal{X}}^2\rho\mathbf{C}_{\mathbf{v}}}{1-\rho} + \frac{D_{\mathcal{X}} D_0^y\mathbf{C}_1\eta\rho^{k+2}}{\rho-\beta} + \frac{\gamma D_{\mathcal{X}}^2\mathbf{L}_{\mathbf{y}}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big] + \frac{1}{2}\mathbf{L}_\ell\gamma D_{\mathcal{X}}^2.
\end{aligned}$$

Summing both sides of the above inequality from $k = 0$ to $K-1$, we get

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathcal{G}(\mathbf{x}_k) \leq \frac{\ell(\mathbf{x}_0) - \ell(\mathbf{x}_K)}{\gamma} &+ \frac{\mathbf{C}_2 D_0^y D_{\mathcal{X}}}{1-\beta} + K\frac{\gamma\mathbf{C}_2 D_{\mathcal{X}}^2\mathbf{L}_{\mathbf{y}}\beta}{1-\beta} + C_{yx}^g\Big[\frac{\rho D_{\mathcal{X}}\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\|}{1-\rho} \\
&+ K\frac{\gamma D_{\mathcal{X}}^2\rho\mathbf{C}_{\mathbf{v}}}{1-\rho} + \frac{D_{\mathcal{X}} D_0^y\mathbf{C}_1\eta\rho^2}{(1-\rho)(\rho-\beta)} + K\frac{\gamma D_{\mathcal{X}}^2\mathbf{L}_{\mathbf{y}}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)}\Big] + \frac{K}{2}\mathbf{L}_\ell\gamma D_{\mathcal{X}}^2,
\end{aligned}$$

where in the above inequality we use the fact that $\sum_{i=0}^{K}\beta^i \leq \frac{1}{1-\beta}$. Next, dividing both sides of the above inequality by K and denoting the smallest gap function over the iterations from $k=0$ to $K-1$, i.e.,

$$\mathcal{G}_{k^*} \triangleq \min_{0\leq k\leq K-1}\mathcal{G}(\mathbf{x}_k) \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathcal{G}(\mathbf{x}_k),$$

imply that

$$\begin{aligned}
\mathcal{G}_{k^*} \leq \frac{\ell(\mathbf{x}_0) - \ell(\mathbf{x}_K)}{K\gamma} &+ \frac{\gamma\mathbf{C}_2 D_{\mathcal{X}}\mathbf{L}_{\mathbf{y}}\beta}{1-\beta} + \frac{\gamma D_{\mathcal{X}}^2\rho\mathbf{C}_{\mathbf{v}}C_{yx}^g\rho}{1-\rho} + \frac{\gamma D_{\mathcal{X}}^2 C_{yx}^g\mathbf{L}_{\mathbf{y}}\mathbf{C}_1\beta\eta}{(1-\beta)(1-\rho)} + \frac{1}{2}\mathbf{L}_\ell\gamma D_{\mathcal{X}}^2 \\
&+ \frac{\mathbf{C}_2 D_0^y D_{\mathcal{X}}\beta}{K(1-\beta)} + \frac{D_{\mathcal{X}} C_{yx}^g\rho\|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\|}{K(1-\rho)} + \frac{D_{\mathcal{X}} D_0^y C_{yx}^g\mathbf{C}_1\eta\rho^2}{K(1-\beta)(1-\rho)}. \qquad (43)
\end{aligned}$$

□

F PROOF OF COROLLARY 2

We begin the proof by using the result of the Theorem 2.

$$\begin{aligned}\mathcal{G}_{k^*} \le\ & \frac{\ell(\mathbf{x}_0)-\ell(\mathbf{x}_K)}{K\gamma} + \frac{\gamma \mathbf{C}_2 D_{\mathcal{X}} \mathbf{L}_{\mathbf{y}}\beta}{1-\beta} + \frac{\gamma D_{\mathcal{X}}^2 \rho \mathbf{C}_{\mathbf{v}} C_{yx}^g \rho}{1-\rho} + \frac{\gamma D_{\mathcal{X}}^2 C_{yx}^g \mathbf{L}_{\mathbf{y}} \mathbf{C}_1 \beta\eta}{(1-\beta)(1-\rho)} + \frac{1}{2}\mathbf{L}_\ell \gamma D_{\mathcal{X}}^2 \\ & + \frac{\mathbf{C}_2 D_0^y D_{\mathcal{X}}\beta}{K(1-\beta)} + \frac{D_{\mathcal{X}} C_{yx}^g \rho \|\mathbf{w}_0 - \mathbf{v}(\mathbf{x}_0)\|}{K(1-\rho)} + \frac{D_{\mathcal{X}} D_0^y C_{yx}^g \mathbf{C}_1 \eta\rho^2}{K(1-\beta)(1-\rho)} \\ =\ & \mathcal{O}\left(\frac{1}{K\gamma} + \frac{\gamma \mathbf{C}_2 \mathbf{L}_{\mathbf{y}}\beta}{1-\beta} + \frac{\gamma \mathbf{L}_{\mathbf{y}} \mathbf{C}_1 \beta}{(1-\beta)(1-\rho)}\right)\end{aligned}$$

The desired result follows immediately from (43) and the fact that $\ell(\mathbf{x}^*) \le \ell(\mathbf{x}_K)$. Moreover, similar to the proof of Corollary 1 we have that $\mathbf{C}_{\mathbf{v}} = \mathcal{O}(\kappa_g^3)$, $\mathbf{C}_1 = \mathcal{O}(\kappa_g^2)$, $\mathbf{L}_{\mathbf{y}} = \mathcal{O}(\kappa_g)$, and $\min\{1-\rho, 1-\beta\} = \Omega(\frac{1}{\kappa_g})$. Hence, by choosing $\gamma = 1/(\kappa_g^{2.5}\sqrt{K})$, we obtain that $\mathcal{G}_k^* = \mathcal{O}(\frac{1}{K\gamma} + \gamma\kappa_g^5) = \mathcal{O}(\kappa_g^{2.5}/\sqrt{K})$, which leads to an iteration complexity of $\mathcal{O}(\kappa_g^5 \epsilon^{-2})$.

Furthermore, assuming that $\nabla_y f(x,y)$ is uniformly bounded, we conclude that $C_y^f = \mathcal{O}(1)$, hence, $\mathbf{C}_1 = \mathcal{O}(\kappa_g)$ from which we have that $\mathcal{G}_{k^*} = \mathcal{O}(\frac{1}{K\gamma} + \gamma\kappa_g^4)$. Therefore, selecting $\gamma = 1/(\kappa_g^2\sqrt{K})$ implies that $\mathcal{G}_{k^*} = \mathcal{O}(\kappa_g^2/\sqrt{K})$ which leads to an iteration complexity of $\mathcal{O}(\kappa_g^4\epsilon^{-2})$. □

G ADDITIONAL EXPERIMENTS

In this section, we provide more details about the experiments conducted in section 5 as well as some additional experiments.

G.1 EXPERIMENT DETAILS

In this section, we include more details of the numerical experiments in Section 5. The MATLAB code is also included in the supplementary material.

For completeness, we briefly review the update rules of SBFW (Akhtar et al., 2022) and TTSA (Hong et al., 2020) for the setting considered in problem (1). In the following, we use $\mathcal{P}_{\mathcal{X}}(\cdot)$ to denote the Euclidean projection onto the set $\mathcal{X}$.

Each iteration of SBFW has the following updates:

$$\begin{aligned}&\mathbf{y}_k = \mathbf{y}_{k-1} - \delta_k \nabla_y g(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}),\\ &\mathbf{d}_k = (1-\rho_k)(\mathbf{d}_{k-1} - h(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})) + h(\mathbf{x}_k, \mathbf{y}_k),\\ &\mathbf{s}_k = \operatorname*{argmin}_{\mathbf{s}\in\mathcal{X}} \langle \mathbf{s}, \mathbf{d}_k\rangle,\\ &\mathbf{x}_{k+1} = (1-\eta_k)\mathbf{x}_k + \eta_k \mathbf{s}_k\end{aligned}$$

Based on the theoretical analysis in (Akhtar et al., 2022), $\rho_k = \frac{2}{k^{1/2}}$, $\eta_k = \frac{2}{(k+1)^{3/4}}$, and $\delta_k = \frac{a_0}{k^{1/2}}$ where $a_0 = \min\left\{\frac{2}{3\mu_g}, \frac{\mu_g}{2L_g^2}\right\}$. Moreover, $h(\mathbf{x}_k, \mathbf{y}_k)$ is a biased estimator of the surrogate $\ell(\mathbf{x}_k)$ which can be computed as follows

$$h(\mathbf{x}_k, \mathbf{y}_k) = \nabla_x f(\mathbf{x}_k, \mathbf{y}_k) - M(\mathbf{x}_k, \mathbf{y}_k)\nabla_y f(\mathbf{x}_k, \mathbf{y}_k),$$

where the term $M(\mathbf{x}_k, \mathbf{y}_k)$ is a biased estimation of $[\nabla_{yy}^2 g(\mathbf{x}_k, \mathbf{y}_k)]^{-1}$ with bounded variance whose explicit form is

$$M(\mathbf{x}_k, \mathbf{y}_k) = \nabla_{yx}^2 g(\mathbf{x}_k, \mathbf{y}_k) \times \left[\frac{k}{L_g}\Pi_{i=1}^l \Big(I - \frac{1}{L_g}\nabla_{yy}^2 g(x_k, y_k)\Big)\right],$$

and $l \in \{1, \ldots, k\}$ is an integer selected uniformly at random.

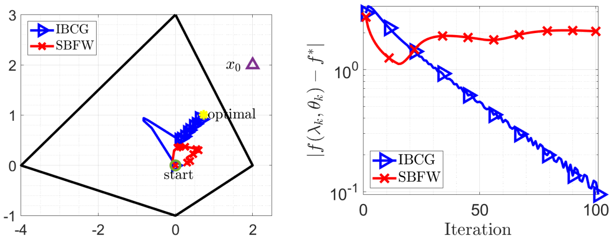

Figure 3: The performance of IBCG (blue) vs SBFW (red) on Problem (44) when $\mu_g = 1$. Plots from left to right are trajectories of θ_k and $f(\lambda_k, \theta_k) - f^*$.

The steps of TTSA algorithm are given by

$$
\begin{aligned}
&\mathbf{y}_{k+1} = \mathbf{y}_k - \beta h_k^g, \\
&\mathbf{x}_{k+1} = \mathcal{P}_{\mathcal{X}}(\mathbf{x}_k - \alpha h_k^f), \\
&h_k^g = \nabla_y g(\mathbf{x}_k, \mathbf{y}_k), \\
&h_k^f = \nabla_x f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{yx}^2 g(\mathbf{x}_k, \mathbf{y}_k) \times \left[\frac{t_{max}(k) c_h}{L_g} \Pi_{i=1}^p \Big(I - \frac{c_h}{L_g} \nabla_{yy}^2 g(\mathbf{x}_k, \mathbf{y}_k)\Big)\right] \nabla_y f(\mathbf{x}_k, \mathbf{y}_k),
\end{aligned}
$$

where based on the theory we define $L = L_x^f + \frac{L_y^f C_{yx}^g}{\mu_g} + C_y^f\left(\frac{L_{yx}^g}{\mu_g} + \frac{L_{yy}^g C_{yx}^g}{\mu_g^2}\right)$, and $L_y = \frac{C_{yx}^g}{\mu_g}$, then set $\alpha = \min\left\{\frac{\mu_g^2}{8 L_y L L_g^2}, \frac{1}{4 L_y L} K^{-3/5}\right\}$, $\beta = \min\left\{\frac{\mu_g}{L_g^2}, \frac{2}{\mu_g} K^{-2/5}\right\}$, $t_{max}(k) = \frac{L_g}{\mu_g} \log(k+1)$, $p \in \{0, \ldots, t_{max}(k) - 1\}$, and $c_h \in (0, 1]$.

G.2 TOY EXAMPLE

Here we consider a variation of coreset problem in a two-dimensional space to illustrate the numerical stability of our proposed method. Given a point $x_0 \in \mathbb{R}^2$, the goal is to find the closest point to x_0 such that under a linear map it lies within the convex hull of given points $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^2$. Let $A \in \mathbb{R}^{2\times 2}$ represents the linear map, $X \triangleq [x_1, x_2, x_3, x_4] \in \mathbb{R}^{2\times 4}$, and $\Delta_4 \triangleq \{\lambda \in \mathbb{R}^4 | \langle \lambda, \mathbf{1} \rangle = 1, \lambda \geq 0\}$ be the standard simplex set. This problem can be formulated as the following bilevel optimization problem

$$
\min_{\lambda \in \Delta_4} \frac{1}{2} \|\theta(\lambda) - x_0\|^2 \quad \text{s.t.} \quad \theta(\lambda) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^2} \frac{1}{2} \|A\theta - X\lambda\|^2. \tag{44}
$$

We set the target $x_0 = (2, 2)$ and choose starting points as $\theta_0 = (0, 0)$ and $\lambda_0 = \mathbf{1}_4/4$. We implemented our proposed method and compared it with SBFW (Akhtar et al., 2022). It should be noted that in the SBFW method, they used a biased estimation for $[\nabla_{yy}^2 g(\lambda, \theta)]^{-1} = (A^\top A)^{-1}$ whose bias is upper bounded by $\frac{2}{\mu_g}$ (see (Ghadimi & Wang, 2018, Lemma 3.2)). Figure 3 illustrates the iteration trajectories of both methods for $\mu_g = 1$ and $K = 10^2$. The step-sizes for both methods are selected as suggested by their theoretical analysis. We observe that our method converges to the optimal solution while SBFW fails to converge. This situation for SBFW exacerbates for smaller values of μ_g.

Figure 4 illustrates the iteration trajectories of both methods for $\mu_g = 0.1$ and $K = 10^3$ in which we also included SBFW method whose Hessian inverse matrix is explicitly provided in the algorithm. The step-sizes for both methods are selected as suggested by their theoretical analysis. Despite incorporating the Hessian inverse matrix in the SBFW method, the algorithm's effectiveness is compromised by excessively conservative step-sizes, as dictated by the theoretical result. Consequently, the algorithm fails to converge to the optimal point effectively. Regarding this issue, we tune their step-sizes, i.e., scale the parameter δ and η in their method by a factor of 5 and 0.1, respectively. By tuning the parameters we can see in Figure 5 that the SBFW with Hessian inverse matrix algorithm has a better performance and converges to the optimal solution. In fact, using the Hessian inverse as well as tuning the step-sizes their method converges to the optimal solution while our method always shows a consistent and robust behavior.

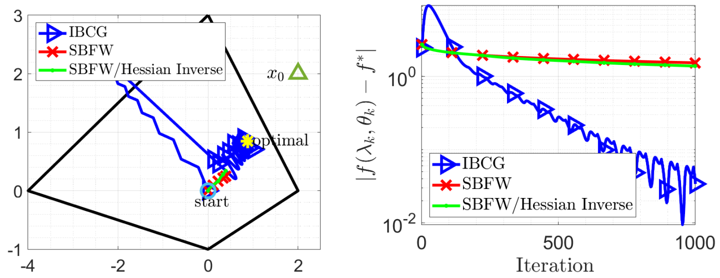

Figure 4: The performance of IBCG (blue) vs SBFW (red) and SBFW with Hessian inverse (green) on Problem (44) when $\mu_g = 0.1$. Plots from left to right are trajectories of θ_k and $f(\lambda_k, \theta_k) - f^*$.

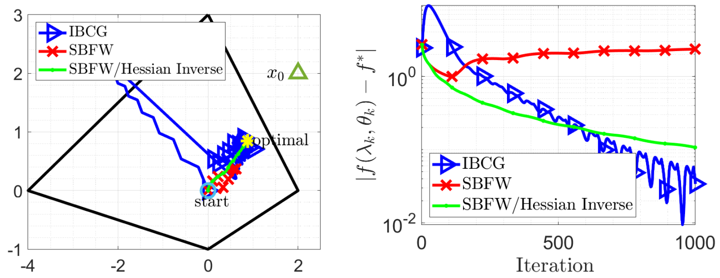

Figure 5: The performance of IBCG (blue) vs SBFW (red) and SBFW with Hessian inverse (green) on Problem (44) when $\mu_g = 0.1$ and the SBFW parameters are tuned. Plots from left to right are trajectories of θ_k and $f(\lambda_k, \theta_k) - f^*$.

G.3 MATRIX COMPLETION WITH DENOISING

G.3.1 SYNTHETIC DATASET

Dataset Generation. We create an observation matrix $M = \hat{X} + E$. In this setting $\hat{X} = WW^T$ where $W \in \mathbb{R}^{n\times r}$ containing normally distributed independent entries, and $E = \hat{n}(L + L^T)$ is a noise matrix where $L \in \mathbb{R}^{n\times n}$ containing normally distributed independent entries and $\hat{n} \in (0, 1)$ is the noise factor. During the simulation process, we set $n = 250$, $r = 10$, and $\alpha = \|\hat{X}\|_*$.

Initialization. All the methods start from the same initial point $\mathbf{x}_0$ and $\mathbf{y}_0$ which are generated randomly. We terminate the algorithms either when the maximum number of iterations $K_{\max} = 10^4$ or the maximum time limit $T_{\max} = 2 \times 10^2$ seconds are achieved.

Implementation Details. For our method IBCG, we choose the step-sizes as $\gamma = \frac{1}{4\sqrt{K}}$ to avoid instability due to large initial step-sizes, and set $\alpha = 2/(\mu_g + L_g)$ and $\eta = 0.9 \times \frac{1-\beta}{\mu_g}$. We tuned the step-size η_k in the SBFW method by multiplying it by a factor of 0.8, and for the TTSA method, we tuned their step-size β by multiplying it by a factor of 0.25.

G.3.2 REAL DATASET

In order to emphasize the importance of projection-free bilevel algorithms in practical applications, we conducted further experiments using a larger dataset known as MovieLens 1M. This dataset consists of 1 million ratings provided by 6000 individuals for a total of 4000 movies. In Figure 6 the inferior performance of TTSA algorithm in actual computation time, especially when dealing with large datasets becomes more evident. The observed difference can be attributed to the utilization of the projection operation in contrast to the projection-free algorithms. TTSA requires performing projections over nuclear norm at each iteration which is computationally expensive due to the computation of full singular value decomposition. In contrast, projection-free algorithms IBCG and SBFW solve a linear minimization at each iteration, which only requires the computation of singular vectors corresponding to the largest singular value. On the other hand, considering the slow

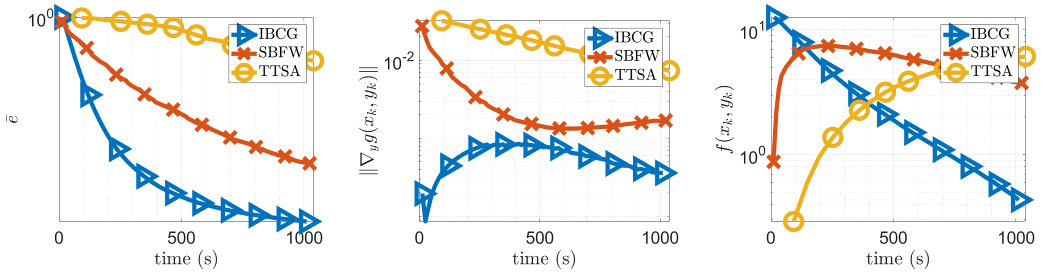

Figure 6: The performance of IBCG (blue) vs SBFW (red) and TTSA (yellow) on Problem (2) for real dataset. Plots from left to right are trajectories of normalized error ($\bar{e}$), $\|\nabla g_y(x_k, y_k)\|$, and $f(x_k, y_k)$ over time.

convergence rate of SBFW, when the size of the dataset increases, the improved performance of our proposed method becomes more evident compared to SBFW.