A PROOFS

A.1 Regularized Nikaido-Isoda (RNI) Approach for FNEs

Example 1.

1. *Let $f(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) + \mathbf{x}^T Q\mathbf{y} - g(\mathbf{y})$. Then, consider the following cases:*
   (a) *(Constrained) Strongly-convex strongly-concave objective : $h(\mathbf{x})$ and $g(\mathbf{x})$ are strongly convex functions.*
   (b) *(Unconstrained) Bilinear objective : $h(\mathbf{x}) = g(\mathbf{y}) = 0$; $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$.*
   (c) *(Unconstrained) Strongly-convex strongly-convex objective : $h(\mathbf{x})$ is a strongly convex function, $g(\mathbf{y})$ is a strongly concave function; $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$.*

   For each of the above problem classes, the RNI reformulated objective (10) *is a convex function.*

2. *Consider an unconstrained objective $f(\mathbf{x},\mathbf{y}), \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ with $n = m$,*

$$\sigma^2_{min}(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})) \geq \lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \tag{28}$$

$$\sigma^2_{max}(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})) \leq \Lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m, \tag{29}$$

and for which the following conditions are satisfied:

$$a)\ -L_x I \preceq \nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y}) \preceq L_x I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \tag{30}$$

$$b)\ -L_y I \preceq \nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}) \preceq L_y I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \tag{31}$$

$$c)\ \lambda \geq \max\left\{ \frac{L+L_y}{L+L_x}\left(L_x^2 + 2LL_x\right), \frac{L+L_x}{L+L_y}\left(L_y^2 + 2LL_y\right) \right\}, \tag{32}$$

where $L > max\{L_x, L_y\}$ is the parameter of the RNI formulation (10)*. In addition, assume that $L_x = L_y = \widetilde{L}$, while we set $L = 2\widetilde{L}$ in* (10)*. Then, if*

$$\left[\lambda - 5\widetilde{L}^2\right]^2 - 144\widetilde{L}^2\Lambda > 0 \tag{33}$$

holds the RNI reformulated objective (10) *is a strongly convex function.*

Note that there exist nonconvex-nonconcave min-max games that belong in the above problem class. For instance, the (non-convex) quadratic function $f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T Q\mathbf{y} + \frac{1}{2}\mathbf{y}^T B\mathbf{y} + \mathbf{c}^T\mathbf{x} + \mathbf{d}^T\mathbf{y} + \mathbf{e}$, $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n$ with $-\widetilde{L}I \preceq A \preceq \widetilde{L}I$, $-\widetilde{L}I \preceq B \preceq \widetilde{L}I$, and for which the inequalities (33) *and $\lambda = \sigma^2_{min}(Q) \geq 5\widetilde{L}^2$ hold, satisfies the above conditions.*

Proof.
1. To begin with notice that

$$\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{x}} h(\mathbf{x}) + Q\mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) = -\nabla_{\mathbf{y}} g(\mathbf{y}) + Q^T\mathbf{x} \text{ and}$$
$$\nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y}) = \nabla^2_{\mathbf{xx}} h(\mathbf{x}), \nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}) = -\nabla^2_{\mathbf{yy}} g(\mathbf{y}), \nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y}) = Q, \nabla^2_{\mathbf{yx}} f(\mathbf{x},\mathbf{y}) = Q^T.$$

As a result

$$\nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) = L(\overline{\mathbf{x}} - \mathbf{x}) + \nabla_{\mathbf{x}} h(\mathbf{x}) + Q\overline{\mathbf{y}}$$
$$\nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) = L(\overline{\mathbf{y}} - \mathbf{y}) + \nabla_{\mathbf{y}} g(\mathbf{y}) - Q^T\overline{\mathbf{x}},$$

where $\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}$ and $\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}$.

(a) Let $(\mathbf{x}^*,\mathbf{y}^*) \in \mathcal{X}\times\mathcal{Y}$ be a stationary point of $P(\mathbf{x},\mathbf{y})$ and $\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x}^*,\mathbf{y}^*)$, $\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x}^*,\mathbf{y}^*)$. Then, from the optimality conditions of P we get

$$\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) + Q\overline{\mathbf{y}} + L(\overline{\mathbf{x}} - \mathbf{x}^*), \mathbf{z} - \mathbf{x}^*\rangle \geq 0, \forall \mathbf{z} \in \mathcal{X} \tag{34}$$

$$\left\langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - Q^T\overline{\mathbf{x}} + L(\overline{\mathbf{y}} - \mathbf{y}^*), \mathbf{z} - \mathbf{y}^*\right\rangle \geq 0, \forall \mathbf{z} \in \mathcal{Y} \tag{35}$$

Moreover, the optimality conditions of $\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*)$ and $\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*)$ imply that

$$\langle \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}) + Q\mathbf{y}^* + L(\overline{\mathbf{x}} - \mathbf{x}^*), \mathbf{z} - \overline{\mathbf{x}} \rangle \geq 0, \forall \mathbf{z} \in \mathcal{X} \tag{36}$$

$$\left\langle \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}) - Q^T\mathbf{x}^* + L(\overline{\mathbf{y}} - \mathbf{y}^*), \mathbf{z} - \overline{\mathbf{y}} \right\rangle \geq 0, \forall \mathbf{z} \in \mathcal{Y}. \tag{37}$$

Setting $\mathbf{z} = \overline{\mathbf{x}}$ in (34), $\mathbf{z} = \mathbf{x}^*$ in (36) and adding them together we get

$$\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}) + Q(\overline{\mathbf{y}} - \mathbf{y}^*), \overline{\mathbf{x}} - \mathbf{x}^* \rangle \geq 0.$$

Similarly, setting $\mathbf{z} = \overline{\mathbf{y}}$ in (35), $\mathbf{y} = \mathbf{y}^*$ in (37) and adding them together we get

$$\left\langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}) + Q^T(\mathbf{x}^* - \overline{\mathbf{x}}), \overline{\mathbf{y}} - \mathbf{y}^* \right\rangle \geq 0.$$

Adding the above two inequalities yields

$$\begin{aligned}
&\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}), \overline{\mathbf{x}} - \mathbf{x}^* \rangle + \langle Q(\overline{\mathbf{y}} - \mathbf{y}^*), \overline{\mathbf{x}} - \mathbf{x}^* \rangle + \\
&\langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}), \overline{\mathbf{y}} - \mathbf{y}^* \rangle + \left\langle Q^T(\mathbf{x}^* - \overline{\mathbf{x}}), \overline{\mathbf{y}} - \mathbf{y}^* \right\rangle \geq 0 \\
&\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}), \mathbf{x}^* - \overline{\mathbf{x}} \rangle \leq - \langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}), \mathbf{y}^* - \overline{\mathbf{y}} \rangle ,
\end{aligned} \tag{38}$$

where in the second line we used $\langle Q(\overline{\mathbf{y}} - \mathbf{y}^*), \overline{\mathbf{x}} - \mathbf{x}^* \rangle = \left\langle \overline{\mathbf{y}} - \mathbf{y}^*, Q^T(\overline{\mathbf{x}} - \mathbf{x}^*) \right\rangle = - \left\langle \overline{\mathbf{y}} - \mathbf{y}^*, Q^T(\mathbf{x}^* - \overline{\mathbf{x}}) \right\rangle$.

If h and g are strongly convex function then we have

$$\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}), \mathbf{x}^* - \overline{\mathbf{x}} \rangle \geq \mu_x \|\mathbf{x}^* - \overline{\mathbf{x}}\|^2$$

$$- \langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}), \mathbf{y}^* - \overline{\mathbf{y}} \rangle \leq -\mu_y \|\mathbf{y}^* - \overline{\mathbf{y}}\|^2,$$

for some $\mu_x > 0, \mu_y > 0$.

Plugging the above expressions into (38) gives

$$0 < \mu_x \|\overline{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}), \mathbf{x}^* - \overline{\mathbf{x}} \rangle \leq - \langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}), \mathbf{y}^* - \overline{\mathbf{y}} \rangle \leq -\mu_y \|\overline{\mathbf{y}} - \mathbf{y}^*\|^2 < 0$$

which implies that

$$\langle \nabla_{\mathbf{x}} h(\mathbf{x}^*) - \nabla_{\mathbf{x}} h(\overline{\mathbf{x}}), \mathbf{x}^* - \overline{\mathbf{x}} \rangle = \langle \nabla_{\mathbf{y}} g(\mathbf{y}^*) - \nabla_{\mathbf{y}} g(\overline{\mathbf{y}}), \mathbf{y}^* - \overline{\mathbf{y}} \rangle = 0.$$

Then, we can deduce that $\overline{\mathbf{x}} = \mathbf{x}^*$ and $\overline{\mathbf{y}} = \mathbf{y}^*$; the strong convexity of h and g exclude the possibility of finding points $\mathbf{x}^* \neq \overline{\mathbf{x}}$, $\mathbf{y}^* \neq \overline{\mathbf{y}}$ such that $\nabla_{\mathbf{x}} h(\mathbf{x}^*) = \nabla_{\mathbf{x}} h(\overline{\mathbf{x}})$ and $\nabla_{\mathbf{y}} g(\mathbf{y}^*) = \nabla_{\mathbf{y}} g(\overline{\mathbf{y}})$. Note that if $\mathbf{x}^* = \overline{\mathbf{x}}(\mathbf{x}^*, \mathbf{y}^*), \mathbf{y}^* = \overline{\mathbf{y}}(\mathbf{x}^*, \mathbf{y}^*)$ then $(\mathbf{x}, \mathbf{y})$ is a global minimum of P. Consequently, every stationary point $(\mathbf{x}^*, \mathbf{y}^*)$ of $P(\mathbf{x}, \mathbf{y})$ is also global min (of $P(\mathbf{x}, \mathbf{y})$) and thus $P(\mathbf{x}, \mathbf{y})$ is a convex function. ■

(b) To begin with, we have that

$$\Phi_x(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathbb{R}^n} \{\mathbf{z}^T Q\mathbf{y} + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|^2\}.$$

The above problem is strongly convex and thus its minimum satisfies the following equation

$$Q\mathbf{y} + L(\mathbf{z} - \mathbf{x}) = 0 \Rightarrow \mathbf{z} = \mathbf{x} - \frac{1}{L} Q\mathbf{y}.$$

Then,

$$\Phi_x(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \frac{1}{L} Q\mathbf{y})^T Q\mathbf{y} + \frac{L}{2}\|\frac{1}{L} Q\mathbf{y}\|^2 = \mathbf{x}^T Q\mathbf{y} - \frac{1}{L}\|Q\mathbf{y}\|^2 + \frac{1}{2L}\|Q\mathbf{y}\|^2 = \mathbf{x}^T Q\mathbf{y} - \frac{1}{2L}\|Q\mathbf{y}\|^2.$$

Similarly, w.r.t variable $\mathbf{y}$ we have the expression

$$\Phi_y(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{y} \in \mathbb{R}^m} \{-\mathbf{x}^T Q\mathbf{z} + \frac{L}{2}\|\mathbf{z} - \mathbf{y}\|^2\}$$

whose minimum satisfies the following equation,

$$-Q^T\mathbf{x} + L(\mathbf{z}-\mathbf{y}) = 0 \Rightarrow \mathbf{z} = \mathbf{y} + \frac{1}{L}Q^T\mathbf{x}.$$

Then, $\Phi_y(\mathbf{x},\mathbf{y})$ admits the following form

$$\Phi_y(\mathbf{x},\mathbf{y}) = -\mathbf{x}^TQ(\mathbf{y}+\frac{1}{L}Q^T\mathbf{x}) + \frac{L}{2}\|\frac{1}{L}Q^T\mathbf{x}\|^2 = -\mathbf{x}^TQ\mathbf{y} - \frac{1}{2L}\|Q^T\mathbf{x}\|^2.$$

As a result,

$$\begin{aligned} P(\mathbf{x},\mathbf{y}) &= -\mathbf{x}^TQ\mathbf{y} + \frac{1}{2L}\|Q\mathbf{y}\|^2 + \mathbf{x}^TQ\mathbf{y} + \frac{1}{2L}\|Q^T\mathbf{x}\|^2 = \frac{1}{2L}\|Q\mathbf{y}\|^2 + \frac{1}{2L}\|Q^T\mathbf{x}\|^2 \\ &= \frac{1}{2L}\mathbf{y}^TQ^TQ\mathbf{y} + \frac{1}{2L}\mathbf{x}^TQQ^T\mathbf{x} = \frac{1}{2L}\begin{bmatrix}\mathbf{x}^T\mathbf{y}^T\end{bmatrix}\begin{bmatrix}QQ^T & 0\\ 0 & Q^TQ\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}. \end{aligned}$$

The above matrix is positive semidefinite and thus $P(\mathbf{x},\mathbf{y})$ is a convex quadratic function. ■

(c) Let $(\mathbf{x}^*,\mathbf{y}^*) \in \mathcal{X}\times\mathcal{Y}$ be a stationary point of $P(\mathbf{x},\mathbf{y})$ and $\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x}^*,\mathbf{y}^*)$, $\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x}^*,\mathbf{y}^*)$. Using a similar reasoning as in (a), consider the expressions in (34), (35), (36), (37). For unconstrained problems these admit the form

$$\nabla_{\mathbf{x}}h(\mathbf{x}^*) + Q\overline{\mathbf{y}} + L(\overline{\mathbf{x}}-\mathbf{x}^*) = 0 \tag{39}$$

$$\nabla_{\mathbf{y}}g(\mathbf{y}^*) - Q^T\overline{\mathbf{x}} + L(\overline{\mathbf{y}}-\mathbf{y}^*) = 0 \tag{40}$$

and

$$\nabla_{\mathbf{x}}h(\overline{\mathbf{x}}) + Q\mathbf{y}^* + L(\overline{\mathbf{x}}-\mathbf{x}^*) = 0 \tag{41}$$

$$\nabla_{\mathbf{y}}g(\overline{\mathbf{y}}) - Q^T\mathbf{x}^* + L(\overline{\mathbf{y}}-\mathbf{y}^*) = 0. \tag{42}$$

Combining (39) with (41) and (40) with (42) we get

$$\nabla_{\mathbf{x}}h(\mathbf{x}^*) - \nabla_{\mathbf{x}}h(\overline{\mathbf{x}}) + Q(\overline{\mathbf{y}}-\mathbf{y}^*) = 0$$

$$\nabla_{\mathbf{y}}g(\mathbf{y}^*) - \nabla_{\mathbf{y}}g(\overline{\mathbf{y}}) + Q^T(\overline{\mathbf{x}}-\mathbf{x}^*) = 0.$$

The above two conditions imply that

$$\begin{aligned} &\langle\nabla_{\mathbf{x}}h(\mathbf{x}^*) - \nabla_{\mathbf{x}}h(\overline{\mathbf{x}}) + Q(\overline{\mathbf{y}}-\mathbf{y}^*), \mathbf{x}^*-\overline{\mathbf{x}}\rangle = \left\langle\nabla_{\mathbf{y}}g(\mathbf{y}^*) - \nabla_{\mathbf{y}}g(\overline{\mathbf{y}}) + Q^T(\overline{\mathbf{x}}-\mathbf{x}^*), \mathbf{y}^*-\overline{\mathbf{y}}\right\rangle \\ &\langle\nabla_{\mathbf{x}}h(\mathbf{x}^*) - \nabla_{\mathbf{x}}h(\overline{\mathbf{x}}), \mathbf{x}^*-\overline{\mathbf{x}}\rangle = \langle\nabla_{\mathbf{y}}g(\mathbf{y}^*) - \nabla_{\mathbf{y}}g(\overline{\mathbf{y}}), \mathbf{y}^*-\overline{\mathbf{y}}\rangle, \end{aligned}$$

since $\langle Q(\overline{\mathbf{y}}-\mathbf{y}^*), \mathbf{x}^*-\overline{\mathbf{x}}\rangle = \left\langle\overline{\mathbf{y}}-\mathbf{y}^*, Q^T(\mathbf{x}^*-\overline{\mathbf{x}})\right\rangle = \left\langle\mathbf{y}^*-\overline{\mathbf{y}}, Q^T(\overline{\mathbf{x}}-\mathbf{x}^*)\right\rangle$. Also, h is a strongly convex function and g is a strongly concave function and so we get

$$\mu_x\|\mathbf{x}^*-\overline{\mathbf{x}}\|^2 \leq \langle\nabla_{\mathbf{x}}h(\mathbf{x}^*) - \nabla_{\mathbf{x}}h(\overline{\mathbf{x}}), \mathbf{x}^*-\overline{\mathbf{x}}\rangle = \langle\nabla_{\mathbf{y}}g(\mathbf{y}^*) - \nabla_{\mathbf{y}}g(\overline{\mathbf{y}}), \mathbf{y}^*-\overline{\mathbf{y}}\rangle \leq -\mu_y\|\mathbf{y}^*-\overline{\mathbf{y}}\|^2.$$

The above expression implies that

$$\langle\nabla_{\mathbf{x}}h(\mathbf{x}^*) - \nabla_{\mathbf{x}}h(\overline{\mathbf{x}}), \mathbf{x}^*-\overline{\mathbf{x}}\rangle = 0 \Rightarrow \mathbf{x}^* = \overline{\mathbf{x}}$$

and

$$\langle\nabla_{\mathbf{y}}g(\mathbf{y}^*) - \nabla_{\mathbf{y}}g(\overline{\mathbf{y}}), \mathbf{y}^*-\overline{\mathbf{y}}\rangle = 0 \Rightarrow \mathbf{y}^* = \overline{\mathbf{y}}.$$

Using the same reasoning as in a) we conclude that, $P(\mathbf{x},\mathbf{y})$ is a convex function. ■

2. We want to show that under the conditions specified in (30)-(33) the objective P is strongly convex. We are going to do that by showing that the Hessian of P is positive definite, that is $\nabla^2P(\mathbf{x},\mathbf{y}) \succ 0, \forall\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m$ (Step 2). Therefore, the first step towards that goal is to compute the Hessian of $P(\mathbf{x},\mathbf{y})$ (Step 1). Also, assume that throughout this proof it holds that $n = m$.

Step 1 - Hessian Computation

First, we need the expressions of the gradients of the argmin functions $\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})$ and $\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})$. Specifically, we have that

$$\begin{aligned}\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) &= \arg\min_{\mathbf{z}\in\mathbb{R}^n}\{f(\mathbf{z},\mathbf{y}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}\\ \Rightarrow &\nabla_{\mathbf{x}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + L(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})-\mathbf{x}) = 0, \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Taking the gradient w.r.t $\mathbf{x}$ of the above expression yields

$$\begin{aligned}&\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + L(\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) - I) = 0\\ &\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})\left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right) = LI\\ &\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = L\left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right)^{-1} = LH(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m,\end{aligned}$$

where in the last equality we used the fact that $L > L_x$, which makes the matrix $\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI$ invertible, and we introduced the notation $H(\mathbf{x},\mathbf{y}) = \left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right)^{-1}$. Next, consider the gradient w.r.t $\mathbf{y}$ of the same expression,

$$\begin{aligned}&\nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + (\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + L\nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = 0\\ &\nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})\left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right) = -(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\\ &\nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = -(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right)^{-1}\\ &\nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = -(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})H(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Moreover, we have that

$$\begin{aligned}\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) &= \arg\min_{\mathbf{z}\in\mathbb{R}^m}\{-f(\mathbf{x},\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}\\ &\Rightarrow -\nabla_{\mathbf{y}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + L(\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})-\mathbf{y}) = 0, \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Taking the gradient w.r.t $\mathbf{y}$ of the above expression yields

$$\begin{aligned}&-\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + L(\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) - I) = 0\\ &\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})[-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI] = LI\\ &\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = L\left(-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI\right)^{-1} = LG(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m,\end{aligned}$$

where we used the fact that $L > L_y$, which makes the matrix $\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI$ invertible, and we introduced the notation $G(\mathbf{x},\mathbf{y}) = \left(-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI\right)^{-1}$. Then, considering the gradient w.r.t $\mathbf{y}$ we get

$$\begin{aligned}&-(\nabla^2_{\mathbf{yx}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) - \nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + L\nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = 0\\ &\nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})\left(-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI\right) = (\nabla^2_{\mathbf{yx}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\\ &\nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = (\nabla^2_{\mathbf{yx}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\left(-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI\right)^{-1}\\ &\nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = (\nabla^2_{\mathbf{yx}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))G(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Next, from Lemma 1 we know that

$$\begin{aligned}\nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) &= L(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})-\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\\ \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) &= L(\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})-\mathbf{y}) - \nabla_{\mathbf{y}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}).\end{aligned}$$

Thus,

$$\begin{aligned}\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}) &= L(\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) - I) + \nabla^2_{\mathbf{xx}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + \nabla^T_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\\ &= L^2 H(\mathbf{x},\mathbf{y}) - LI + \nabla^2_{\mathbf{xx}} f(\mathbf{x},\overline{\mathbf{y}}) + \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}})G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m,\end{aligned}$$

where we have used the fact that $\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y}) = (\nabla^2_{\mathbf{yx}})^T f(\mathbf{x},\mathbf{y})$. For the $\nabla^2_{\mathbf{yy}}$ submatrix of the Hessian we have

$$\begin{aligned}\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) &= L(\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) - I) - \nabla^T_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{yx}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) - \nabla^2_{\mathbf{yy}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) \\ &= L^2 G(\mathbf{x},\mathbf{y}) - LI - \nabla^2_{\mathbf{yy}} f(\overline{\mathbf{x}},\mathbf{y}) + \nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{yx}})^T f(\overline{\mathbf{x}},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Also, for the non-diagonal submatrices of the Hessian of $P(\mathbf{x},\mathbf{y})$ we have that

$$\begin{aligned}\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y}) &= L\nabla_{\mathbf{y}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) + \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\nabla_{\mathbf{y}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) \\ &= -LH(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}}(\overline{\mathbf{x}},\mathbf{y}) + L\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m\end{aligned}$$

and

$$\begin{aligned}\nabla^2_{\mathbf{yx}} P(\mathbf{x},\mathbf{y}) &= L\nabla_{\mathbf{x}}\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) - \nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\nabla_{\mathbf{x}}\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) \\ &= LG(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{yx}} f(\mathbf{x},\overline{\mathbf{y}}) - L\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y}), \forall \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m.\end{aligned}$$

Therefore, the Hessian of P is

$$\nabla^2 P(\mathbf{x},\mathbf{y}) = \begin{bmatrix}\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y}) \\ \nabla^2_{\mathbf{yx}} P(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\end{bmatrix} = \begin{bmatrix}\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y}) \\ (\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\end{bmatrix}.$$

Note that it is a symmetric matrix.

Step 2 - Hessian is positive definite

From Horn and Johnson (2012, Theorem 7.7.7, pg. 495) we know that

$$\nabla^2 P(\mathbf{x},\mathbf{y}) \succ 0 \Leftrightarrow \begin{cases}\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}) \succ 0 & (A) \\ \nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) - (\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y}) \left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y}) \succ 0 & (B)\end{cases}.$$

The plan is to establish the positive definiteness of the above matrices by computing lower bounds for their minimum eigenvalues and ensuring that these bounds are positive (over all $\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m$) under the conditions specified in (30)-(33). Specifically, for the expression (B) we have that

$$\begin{aligned}&\lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) - (\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y}) \left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) \\ &\overset{(a)}{\geq} \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) + \lambda_{min}\left(-(\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y}) \left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) \\ &\overset{(b)}{=} \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) + \theta_{min}(\mathbf{x},\mathbf{y})\lambda_{min}\left(-(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))^{-1}\right) \\ &\overset{(c)}{\geq} \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) + \sigma^2_{max}\left((\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})\right)\lambda_{min}\left(-(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))^{-1}\right) \\ &\overset{(d)}{=} \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) - \sigma^2_{max}\left((\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})\right)\lambda_{max}\left((\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))^{-1}\right),\end{aligned} \tag{43}$$

where in (a) we used Weyl's inequality (Horn and Johnson, 2012, Corollary 4.3.15, pg. 242); in (b),(c) we used a corollary of Ostrowskii's theorem (Horn and Johnson, 2012, Corollary 4.5.11, pg. 284) which establishes that for every $\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m$ there exists a $\theta_{min}(\mathbf{x},\mathbf{y})$ such that

$$\lambda_{min}\left(-(\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y}) \left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1} \nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) = \theta_{min}(\mathbf{x},\mathbf{y})\lambda_{min}\left(-\left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1}\right),$$

with $\sigma^2_{min}\left((\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})\right) \leq \theta_{min}(\mathbf{x},\mathbf{y}) \leq \sigma^2_{max}\left((\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})\right)$; in (d) we exploit the fact that for an eigenvalue λ of some matrix M, $-\lambda$ is an eigenvalue of $-M$. From the above expression it is apparent that we need to derive lower bounds for the minimum eigenvalues of $\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})$ and $\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})$, as well as an upper bound for the largest singular value of $\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})$.

Towards that goal, we start from the fundamental assumption about the Hessian w.r.t $\nabla^2_{\mathbf{xx}}$ (30) and we get,

$$
\begin{aligned}
&-L_x I \preceq \nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) \preceq L_x I \\
&\Rightarrow 0 \prec (L-L_x)I \preceq \nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI \preceq (L+L_x)I, \text{ for } L > L_x \\
&\Rightarrow \frac{1}{L+L_x} I \preceq \left(\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI\right)^{-1} \preceq \frac{1}{L-L_x} I \qquad (44) \\
&\Rightarrow \left(\frac{L^2}{L+L_x} - L - L_x\right) I \preceq L^2 H(\mathbf{x},\mathbf{y}) - LI + \nabla^2_{\mathbf{xx}} f(\mathbf{x},\overline{\mathbf{y}}) \preceq \left(\frac{L^2}{L-L_x} - L + L_x\right) I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \\
&\Rightarrow \lambda_{min}\left(L^2 H(\mathbf{x},\mathbf{y}) - LI + \nabla^2_{\mathbf{xx}} f(\mathbf{x},\overline{\mathbf{y}})\right) \geq \frac{-L_x^2 - 2LL_x}{L+L_x}, \qquad (45)
\end{aligned}
$$

where in the third line we utilized the positive definiteness of the matrix $\nabla^2_{\mathbf{xx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) + LI$ and the fact that for an eigenvalue λ of some matrix M we know that $\frac{1}{\lambda}$ is an eigenvalue of M^{-1}. Similarly, using the respective assumption w.r.t $\nabla^2_{\mathbf{yy}}$ (31) we have that

$$
\begin{aligned}
&-L_y I \preceq \nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) \preceq L_y I \\
&0 \prec (L-L_y)I \preceq -\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI \preceq (L+L_y)I, \text{ for } L > L_y \\
&\frac{1}{L+L_y} I \preceq \left(-\nabla^2_{\mathbf{yy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) + LI\right)^{-1} \preceq \frac{1}{L-L_y} I \qquad (46) \\
&\left(\frac{L^2}{L+L_y} - L - L_y\right) I \preceq L^2 G(\mathbf{x},\mathbf{y}) - LI - \nabla^2_{\mathbf{yy}} f(\overline{\mathbf{x}},\mathbf{y}) \preceq \left(\frac{L^2}{L-L_y} - L + L_y\right) I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \\
&\lambda_{min}\left(L^2 G(\mathbf{x},\mathbf{y}) - LI - \nabla^2_{\mathbf{yy}} f(\overline{\mathbf{x}},\mathbf{y})\right) \geq \frac{-L_y^2 - 2LL_y}{L+L_y}. \qquad (47)
\end{aligned}
$$

Next, we want to lower bound the minimum eigenvalue of $\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) G(\mathbf{x},\mathbf{y}) (\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))$. Specifically, we have that

$$
\begin{aligned}
&\lambda_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y})) G(\mathbf{x},\mathbf{y}) (\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\right) \\
&\overset{(a)}{=} \theta_{min}(\mathbf{x},\mathbf{y}) \lambda_{min}\left(G(\mathbf{x},\mathbf{y})\right) \\
&\overset{(b)}{\geq} \sigma^2_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\right) \lambda_{min}\left(G(\mathbf{x},\mathbf{y})\right) \\
&\overset{(c)}{\geq} \frac{\lambda}{L+L_y}, \qquad (48)
\end{aligned}
$$

where in (a), (b) we use Ostrowskii's theorem (Horn and Johnson, 2012, Corollary 4.5.11, pg. 284) and the respective inequality $\sigma^2_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\right) \leq \theta_{max}(\mathbf{x},\mathbf{y}) \leq \sigma_{max}\left(\nabla_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\right)$; in (c) we use the assumption $\sigma^2_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}(\mathbf{x},\mathbf{y}))\right) \geq \lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ and the bound in (46). Moreover, we want to bound $\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) H(\mathbf{x},\mathbf{y}) (\nabla^2_{\mathbf{yx}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})$ and in fact we have that

$$
\begin{aligned}
&\lambda_{min}\left(\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y}) H(\mathbf{x},\mathbf{y}) (\nabla^2_{\mathbf{yx}})^T f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\right) \\
&\overset{(a)}{=} \theta_{min}(\mathbf{x},\mathbf{y}) \lambda_{min}\left(H(\mathbf{x},\mathbf{y})\right) \\
&\overset{(b)}{\geq} \sigma^2_{min}(\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})) \lambda_{min}\left(H(\mathbf{x},\mathbf{y})\right) \\
&\overset{(c)}{=} \sigma^2_{min}(\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})) \frac{1}{L+L_x} \\
&\overset{(d)}{\geq} \frac{\lambda}{L+L_x}, \qquad (49)
\end{aligned}
$$

where in (a), (b) we use Ostrowskii's theorem (Horn and Johnson, 2012, Corollary 4.5.11, pg. 284) and the inequality $\sigma^2_{min}\left(\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\right) \leq \theta_{min}(\mathbf{x},\mathbf{y}) \leq \sigma_{max}\left(\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})\right)$, in (c) we exploit the property $\sigma_{min}\left(\nabla^2_{\mathbf{yx}} f(\mathbf{x},\mathbf{y})\right) = \sigma_{min}\left((\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\mathbf{y})\right) = \sigma_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})\right), \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ and the bound in (44), and in (d) the assumption $\sigma^2_{min}(\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}}(\mathbf{x},\mathbf{y}),\mathbf{y})) \geq \lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$.

Combining the results in (45) and (48), and using a corollary of Weyl's theorem (Horn and Johnson, 2012, Corollary 4.3.15, pg. 242) we can derive bounds for the minimum eigenvalue of $\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})$, that is

$$\begin{aligned}\lambda_{min}\left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right) &\geq \lambda_{min}\left(L^2 H(\mathbf{x},\mathbf{y}) - LI + \nabla^2_{\mathbf{xx}} f(\mathbf{x},\overline{\mathbf{y}})\right) + \lambda_{min}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}})\right) \\ &\geq \frac{-L_x^2 - 2LL_x}{L+L_x} + \frac{\lambda}{L+L_y}. \end{aligned} \tag{50}$$

Notice that

$$\lambda \geq \frac{L+L_y}{L+L_x}\left(L_x^2 + 2LL_x\right) \Rightarrow \lambda_{min}\left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right) \geq 0. \tag{51}$$

Also, combining the results in (47) and (49), and using a corollary of Weyl's theorem (Horn and Johnson, 2012, Corollary 4.3.15, pg. 242) we get

$$\begin{aligned}\lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) &\geq \lambda_{min}\left(L^2 G(\mathbf{x},\mathbf{y}) - LI - \nabla^2_{\mathbf{yy}} f(\overline{\mathbf{x}},\mathbf{y})\right) + \lambda_{min}\left(\nabla^2_{\mathbf{yx}} f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{yx}})^T f(\overline{\mathbf{x}},\mathbf{y})\right) \\ &\geq \frac{-L_y^2 - 2LL_y}{L+L_y} + \frac{\lambda}{L+L_x}. \end{aligned} \tag{52}$$

Then, notice that

$$\lambda \geq \frac{L+L_x}{L+L_y}\left(L_y^2 + 2LL_y\right) \Rightarrow \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) \geq 0. \tag{53}$$

Next, we focus on the term $\sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right)$, that is

$$\begin{aligned}\sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) &= \sigma^2_{max}\left(-LH(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y}) + L\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})\right) \\ &= \lambda_{max}\Big(\left(-LH(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y}) + L\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})\right) \cdot \\ &\qquad \cdot \left(-LH(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y}) + L\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})\right)^T \Big) \\ &= \lambda_{max}\Big(L^2 H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y}) \\ &\qquad + L^2 \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}) \\ &\qquad - L^2 H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}) \\ &\qquad - L^2 \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})\Big) \\ &\overset{(a)}{\leq} \lambda_{max}\left(L^2 H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})\right) \\ &+ \lambda_{max}\left(L^2 \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}})\right) \\ &+ \lambda_{max}\big(- L^2 H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}) \\ &\qquad - L^2 \nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}}) G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})\big), \end{aligned}$$

where in (a) we used Weyl's inequality (Horn and Johnson, 2012, Corollary 4.3.15, pg. 242). Now let's work separately on the above three terms. Starting with the first one we have

$$\begin{aligned}\lambda_{max}\left(L^2 H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y}) H(\mathbf{x},\mathbf{y})\right) &= \sigma^2_{max}\left(LH(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})\right) \\ &\overset{(a)}{\leq} \sigma^2_{max}\left(LH(\mathbf{x},\mathbf{y})\right)\sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})\right) \\ &\overset{(b)}{\leq} L^2 \sigma^2_{max}\left(H(\mathbf{x},\mathbf{y})\right)\Lambda \\ &= L^2 \lambda_{max}\left(H(\mathbf{x},\mathbf{y})H^T(\mathbf{x},\mathbf{y})\right)\Lambda \\ &= L^2 \lambda^2_{max}\left(H(\mathbf{x},\mathbf{y})\right)\Lambda \\ &\overset{(c)}{\leq} \frac{L^2}{(L-L_x)^2}\Lambda, \end{aligned} \tag{54}$$

where in (a) we used the property $\sigma_{max}(AB) \leq \sigma_{max}(A)\sigma_{max}(B)$ (Hogben, 2013, property 7c, pg. 17-8), in (b) we utilize the assumption $\sigma^2_{max}(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})) \leq \Lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$, and in (c) we use (44). Similarly,

$$\begin{aligned}
\lambda_{max}\left(L^2\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}})G(\mathbf{x},\mathbf{y})G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}})\right) &= \sigma^2_{max}\left(L\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}})G(\mathbf{x},\mathbf{y})\right) \\
&\overset{(a)}{\leq} \sigma^2_{max}\left(LG(\mathbf{x},\mathbf{y})\right)\sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}})\right) \\
&\overset{(b)}{\leq} L^2\sigma^2_{max}\left(G(\mathbf{x},\mathbf{y})\right)\Lambda \\
&= L^2\lambda_{max}\left(G(\mathbf{x},\mathbf{y})G^T(\mathbf{x},\mathbf{y})\right)\Lambda \\
&= L^2\lambda^2_{max}\left(G(\mathbf{x},\mathbf{y})\right)\Lambda \\
&\overset{(c)}{\leq} \frac{L^2}{(L-L_y)^2}\Lambda, \qquad (55)
\end{aligned}$$

where in (a) we used the property $\sigma_{max}(AB) \leq \sigma_{max}(A)\sigma_{max}(B)$ (Hogben, 2013, property 7c, pg. 17-8) in (b) we utilize the assumption $\sigma^2_{max}(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})) \leq \Lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$, and in (c) we use (46). For the third term we have

$$\begin{aligned}
&\lambda_{max}\left(-L^2H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}}) - L^2\nabla^2_{\mathbf{xy}} f(\mathbf{x},\overline{\mathbf{y}})G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\overline{\mathbf{x}},\mathbf{y})H(\mathbf{x},\mathbf{y})\right) \\
&\overset{(a)}{\leq} 2\sigma_{max}\left(L^2H(\mathbf{x},\mathbf{y})\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})G(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}})\right) \\
&\overset{(b)}{\leq} 2\sigma_{max}\left(LH(\mathbf{x},\mathbf{y})\right)\sigma_{max}\left(LG(\mathbf{x},\mathbf{y})\right)\sigma_{max}\left(\nabla^2_{\mathbf{xy}} f(\overline{\mathbf{x}},\mathbf{y})\right)\sigma_{max}\left((\nabla^2_{\mathbf{xy}})^T f(\mathbf{x},\overline{\mathbf{y}})\right) \\
&\overset{(c)}{\leq} 2\sigma_{max}\left(LH(\mathbf{x},\mathbf{y})\right)\sigma_{max}\left(LG(\mathbf{x},\mathbf{y})\right)\Lambda \\
&= 2\lambda^{1/2}_{max}\left(L^2H(\mathbf{x},\mathbf{y})H^T(\mathbf{x},\mathbf{y})\right)\lambda^{1/2}_{max}\left(L^2G(\mathbf{x},\mathbf{y})G^T(\mathbf{x},\mathbf{y})\right)\Lambda \\
&= 2L^2\lambda_{max}\left(H(\mathbf{x},\mathbf{y})\right)\lambda_{max}\left(G(\mathbf{x},\mathbf{y})\right)\Lambda \\
&\overset{(d)}{\leq} \frac{2L^2}{(L-L_x)(L-L_y)}\Lambda, \qquad (56)
\end{aligned}$$

where in (a) we used the properties $\lambda_{max}(M+M^T) \leq 2\sigma_{max}(M)$ (Hogben, 2013, property 10, pg. 17-9) and $\sigma_{max}(-M) = \sigma_{max}(M)$, in (b) we used the property $\sigma_{max}(AB) \leq \sigma_{max}(A)\sigma_{max}(B)$ (Hogben, 2013, property 7c, pg. 17-8), in (c) we utilize the assumption $\sigma^2_{max}(\nabla^2_{\mathbf{xy}} f(\mathbf{x},\mathbf{y})) \leq \Lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$, and in (d) we use the bounds in (44), (46). As a result, combining the above three terms we have

$$\sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) \leq \frac{L^2}{(L-L_x)^2}\Lambda + \frac{L^2}{(L-L_y)^2}\Lambda + \frac{2L^2}{(L-L_x)(L-L_y)}\Lambda. \qquad (57)$$

Then, by noting that for an eigenvalue λ of some matrix M we have that $\frac{1}{\lambda}$ is an eigenvalue of M^{-1}, and by considering the fact that $\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})$ is a positive definite matrix (under the condition (51)), inequality (43) becomes

$$\begin{aligned}
&\lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) - (\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})\left(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y})\right)^{-1}\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) \\
&\geq \lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y})\right) - \sigma^2_{max}\left(\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right)\frac{1}{\lambda_{min}(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))}. \qquad (58)
\end{aligned}$$

Combining the results in (50), (52), (57),(58), and assuming that condition (51) holds, we get

$$\begin{aligned}
&\lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) - (\nabla^2_{\mathbf{xy}})^T P(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))^{-1}\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) \\
&\geq \frac{-L_y^2 - 2LL_y}{L+L_y} + \frac{\lambda}{L+L_x} - \frac{\frac{L^2}{(L-L_x)^2}\Lambda + \frac{L^2}{(L-L_y)^2}\Lambda + \frac{2L^2}{(L-L_x)(L-L_y)}\Lambda}{\frac{-L_x^2-2LL_x}{L+L_x} + \frac{\lambda}{L+L_y}}.
\end{aligned}$$

We want to ensure that

$$\lambda_{min}\left(\nabla^2_{\mathbf{yy}} P(\mathbf{x},\mathbf{y}) - (\nabla^2_{\mathbf{yx}})^T P(\mathbf{x},\mathbf{y})(\nabla^2_{\mathbf{xx}} P(\mathbf{x},\mathbf{y}))^{-1}\nabla^2_{\mathbf{xy}} P(\mathbf{x},\mathbf{y})\right) > 0.$$

So λ and Λ must satisfy the following condition

$$\begin{aligned}
&\frac{-L_y^2-2LL_y}{L+L_y}+\frac{\lambda}{L+L_x}-\frac{\frac{L^2}{(L-L_x)^2}\Lambda+\frac{L^2}{(L-L_y)^2}\Lambda+2\frac{L^2}{(L-L_x)(L-L_y)}\Lambda}{\frac{-L_x^2-2LL_x}{L+L_x}+\frac{\lambda}{L+L_y}}>0\\
&\Rightarrow\left[\frac{-L_y^2-2LL_y}{L+L_y}+\frac{\lambda}{L+L_x}\right]\left[\frac{-L_x^2-2LL_x}{L+L_x}+\frac{\lambda}{L+L_y}\right]-\left[\frac{L}{L-L_x}+\frac{L}{L-L_y}\right]^2\Lambda>0\\
&\Rightarrow\left[\frac{-L_y^2-2LL_y}{L+L_y}+\frac{\lambda}{L+L_x}\right]\left[\frac{-L_x^2-2LL_x}{L+L_x}+\frac{\lambda}{L+L_y}\right]-\left[\frac{2L^2-LL_x-LL_y}{(L-L_x)(L-L_y)}\right]^2\Lambda>0 \qquad (59)
\end{aligned}$$

where we have assumed in the above calculations that the conditions (51) and (53) hold.

Furthermore, assume that $L_x=L_y=\widetilde{L}$ and set $L=2\widetilde{L}>\max\{L_x,L_y\}$. Then, (59) becomes

$$\begin{aligned}
&\left[\frac{\lambda}{3\widetilde{L}}-\frac{5\widetilde{L}^2}{3\widetilde{L}}\right]\left[\frac{\lambda}{3\widetilde{L}}-\frac{5\widetilde{L}^2}{3\widetilde{L}}\right]-\frac{4\widetilde{L}^2\cdot 4\widetilde{L}^2}{\widetilde{L}^2\widetilde{L}^2}\Lambda>0\\
&\left[\lambda-5\widetilde{L}^2\right]^2-144\widetilde{L}^2\Lambda>0. \qquad (60)
\end{aligned}$$

In conclusion, under the conditions (30), (31), (51), (53), where the last two inequalities can be jointly expressed as $\lambda\geq\max\left\{\frac{L+L_y}{L+L_x}\left(L_x^2+2LL_x\right),\frac{L+L_x}{L+L_y}\left(L_y^2+2LL_y\right)\right\}$, as well as condition (60) (where we additionally selected $L=2\widetilde{L}>\max\{L_x,L_y\}$), we can ensure that $\nabla^2P(\mathbf{x},\mathbf{y})\succ 0,\forall\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m$.

□

Proposition 2. *Suppose that Assumption 1 holds. Then, provided that $L>\max\{L_x,L_y\}$, P has Lipschitz continuous gradients in $\mathbf{z}=(\mathbf{x},\mathbf{y})$, that is*

$$\|\nabla P(\mathbf{z}_1)-\nabla P(\mathbf{z}_2)\|\leq\overline{L}\|\mathbf{z}_1-\mathbf{z}_2\|,\ \forall\mathbf{z}_1,\mathbf{z}_2\in\mathcal{X}\times\mathcal{Y},$$

with constant $\overline{L}=\overline{L}_x+\overline{L}_y$, $\overline{L}_x=L+L_x+\frac{L^2+LL_y}{L-L_x}+\frac{L_xL_y+LL_y}{L-L_y}$, $\overline{L}_y=L+L_y+\frac{L^2+LL_x}{L-L_y}+\frac{L_xL_y+LL_x}{L-L_x}$.

Proof. For $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}$, $\mathbf{y}_1,\mathbf{y}_2\in\mathcal{Y}$ we have

$$\begin{aligned}
\|\nabla_{\mathbf{x}}P(\mathbf{x}_1,\mathbf{y}_1)-\nabla_{\mathbf{x}}P(\mathbf{x}_2,\mathbf{y}_2)\|&=\|L(\overline{\mathbf{x}}_1-\mathbf{x}_1)+\nabla_{\mathbf{x}}f(\mathbf{x}_1,\overline{\mathbf{y}}_1)-L(\overline{\mathbf{x}}_2-\mathbf{x}_2)-\nabla_{\mathbf{x}}f(\mathbf{x}_2,\overline{\mathbf{y}}_2)\|\\
&\leq L\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|+L\|\mathbf{x}_1-\mathbf{x}_2\|+\|\nabla_{\mathbf{x}}f(\mathbf{x}_1,\overline{\mathbf{y}}_1)-\nabla_{\mathbf{x}}f(\mathbf{x}_2,\overline{\mathbf{y}}_2)\|\\
&\leq(L+L_x)\|\mathbf{x}_1-\mathbf{x}_2\|+L\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|+L_x\|\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2\|, \qquad (61)
\end{aligned}$$

where in the first inequality we exploited the Lipschitz gradient property of f, and defined $\overline{\mathbf{x}}_1=\arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}_1)+\frac{L}{2}\|\mathbf{z}-\mathbf{x}_1\|^2\}$, $\overline{\mathbf{y}}_1=\arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x}_1,\mathbf{z})+\frac{L}{2}\|\mathbf{z}-\mathbf{y}_1\|^2\}$, $\overline{\mathbf{x}}_2=\arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}_2)+\frac{L}{2}\|\mathbf{z}-\mathbf{x}_2\|^2\}$, $\overline{\mathbf{y}}_2=\arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x}_2,\mathbf{z})+\frac{L}{2}\|\mathbf{z}-\mathbf{y}_2\|^2\}$. Then, the next step is to bound the terms $\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|$ and $\|\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2\|$. In order to derive a bound for the former term we consider the function

$$g(\mathbf{z},\mathbf{x},\mathbf{y})=f(\mathbf{z},\mathbf{y})+\frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2,$$

which is strongly convex in $\mathbf{z}$, for $L>\max\{L_x,L_y\}$, with modulus $\mu_1=L-L_x$. As a result, we have

$$\begin{aligned}
g(\overline{\mathbf{x}}_2,\mathbf{x}_1,\mathbf{y}_1)&\geq g(\overline{\mathbf{x}}_1,\mathbf{x}_1,\mathbf{y}_1)+\langle\nabla_{\mathbf{z}}g(\overline{\mathbf{x}}_1,\mathbf{x}_1,\mathbf{y}_1),\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_1\rangle+\frac{\mu_1}{2}\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|^2\\
g(\overline{\mathbf{x}}_1,\mathbf{x}_1,\mathbf{y}_1)&\geq g(\overline{\mathbf{x}}_2,\mathbf{x}_1,\mathbf{y}_1)+\langle\nabla_{\mathbf{z}}g(\overline{\mathbf{x}}_2,\mathbf{x}_1,\mathbf{y}_1),\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\rangle+\frac{\mu_1}{2}\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|^2.
\end{aligned}$$

Adding the above inequalities yields

$$\langle\nabla_{\mathbf{z}}g(\overline{\mathbf{x}}_1,\mathbf{x}_1,\mathbf{y}_1)-\nabla_{\mathbf{z}}g(\overline{\mathbf{x}}_2,\mathbf{x}_1,\mathbf{y}_1),\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_1\rangle+\mu_1\|\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\|^2\leq 0.$$

Then, combining the above with the respective optimality condition at $\overline{\mathbf{x}}_1$, that is $\langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_1, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1 \rangle \geq 0$ we obtain

$$\mu_1 \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2 \leq \langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1 \rangle .$$

Furthermore, adding to the above inequality the optimality condition of $\overline{\mathbf{x}}_2$, that is $-\langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_2, \mathbf{y}_2), \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1 \rangle \geq 0$, we get

$$\begin{aligned}
&\mu_1 \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2 \leq \langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_2, \mathbf{y}_2), \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1 \rangle \\
&\mu_1 \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2 \leq \|\nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{z}} g(\overline{\mathbf{x}}_2, \mathbf{x}_2, \mathbf{y}_2)\| \|\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1\| \\
&\mu_1 \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\| \leq \|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}}_2, \mathbf{y}_1) + L(\overline{\mathbf{x}}_2 - \mathbf{x}_1) - \nabla_{\mathbf{z}} f(\overline{\mathbf{x}}_2, \mathbf{y}_2) - L(\overline{\mathbf{x}}_2 - \mathbf{x}_2)\| \\
&\mu_1 \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\| \leq \|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}}_2, \mathbf{y}_1) - \nabla_{\mathbf{z}} f(\overline{\mathbf{x}}_2, \mathbf{y}_2)\| + L\|\mathbf{x}_2 - \mathbf{x}_1\| \\
&\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\| \leq \frac{L_x}{\mu_1} \|\mathbf{y}_1 - \mathbf{y}_2\| + \frac{L}{\mu_1} \|\mathbf{x}_1 - \mathbf{x}_2\|,
\end{aligned} \tag{62}$$

where the second inequality follows from Cauchy-Schwarz inequality, in the third inequality we compute the gradient of g w.r.t $\mathbf{z}$, and in the last inequality the Lipschitz gradient property of f is utilized.

In order to bound the $\|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|$ term we follow a similar path, that is we define the function

$$h(\mathbf{z}, \mathbf{x}, \mathbf{y}) = -f(\mathbf{x}, \mathbf{z}) + \frac{L}{2} \|\mathbf{z} - \mathbf{y}\|^2,$$

which is strongly convex in $\mathbf{z}$, for $L > \max\{L_x, L_y\}$, with modulus $\mu = L - L_y$. Then, we have

$$\begin{aligned}
h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1) &\geq h(\overline{\mathbf{y}}_1, \mathbf{x}_1, \mathbf{y}_1) + \langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_1, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1 \rangle + \frac{\mu_2}{2} \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|^2 \\
h(\overline{\mathbf{y}}_1, \mathbf{x}_1, \mathbf{y}_1) &\geq h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1) + \langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2 \rangle + \frac{\mu_2}{2} \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|^2.
\end{aligned}$$

The addition of the above inequalities yields

$$\langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_1, \mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1 \rangle + \mu_2 \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|^2 \leq 0$$

Then, the utilization of the optimality condition of $\overline{\mathbf{y}}_1$ and $\overline{\mathbf{y}}_2$ (following the same approach as above), that is $\langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_1, \mathbf{x}_1, \mathbf{y}_1), \overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1 \rangle \geq 0$ and $-\langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_2, \mathbf{y}_2), \overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1 \rangle \geq 0$, respectively, leads to

$$\begin{aligned}
&\mu_2 \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|^2 \leq \langle \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_2, \mathbf{y}_2), \overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1 \rangle \\
&\mu_2 \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\|^2 \leq \|\nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{z}} h(\overline{\mathbf{y}}_2, \mathbf{x}_2, \mathbf{y}_2)\| \|\overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1\| \\
&\mu_2 \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\| \leq \| - \nabla_{\mathbf{z}} f(\mathbf{x}_1, \overline{\mathbf{y}}_2) + L(\overline{\mathbf{y}}_2 - \mathbf{y}_1) + \nabla_{\mathbf{z}} f(\mathbf{x}_2, \overline{\mathbf{y}}_2) - L(\overline{\mathbf{y}}_2 - \mathbf{y}_2)\| \\
&\mu_2 \|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\| \leq \|\nabla_{\mathbf{z}} f(\mathbf{x}_1, \overline{\mathbf{y}}_2) - \nabla_{\mathbf{z}} f(\mathbf{x}_2, \overline{\mathbf{y}}_2)\| + L\|\mathbf{y}_2 - \mathbf{y}_1\| \\
&\|\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\| \leq \frac{L_y}{\mu_2} \|\mathbf{x}_1 - \mathbf{x}_2\| + \frac{L}{\mu_2} \|\mathbf{y}_1 - \mathbf{y}_2\|.
\end{aligned} \tag{63}$$

Combining (61), (62) and (63) we conclude that

$$\|\nabla_{\mathbf{x}} P(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{x}} P(\mathbf{x}_2, \mathbf{y}_2)\| \leq \left(L + L_x + \frac{L^2}{L - L_x} + \frac{L_x L_y}{L - L_y} \right) \|\mathbf{x}_1 - \mathbf{x}_2\| + \left(\frac{L L_x}{L - L_x} + \frac{L L_x}{L - L_y} \right) \|\mathbf{y}_1 - \mathbf{y}_2\|.$$

Using a similar reasoning we get

$$\|\nabla_{\mathbf{y}} P(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} P(\mathbf{x}_2, \mathbf{y}_2)\| \leq \left(L + L_y + \frac{L^2}{L - L_y} + \frac{L_x L_y}{L - L_x} \right) \|\mathbf{y}_1 - \mathbf{y}_2\| + \left(\frac{L L_y}{L - L_y} + \frac{L L_y}{L - L_x} \right) \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Combining the above results we notice that

$$\|\nabla P(\mathbf{z}_1) - \nabla P(\mathbf{z}_2)\| \leq \overline{L} \|\mathbf{z}_1 - \mathbf{z}_2\|, \ \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{X} \times \mathcal{Y},$$

with constant $\overline{L} = \overline{L}_x + \overline{L}_y$, where $\overline{L}_x = L + L_x + \frac{L^2 + L L_y}{L - L_x} + \frac{L_x L_y + L L_y}{L - L_y}$ and $\overline{L}_y = L + L_y + \frac{L^2 + L L_x}{L - L_y} + \frac{L_x L_y + L L_x}{L - L_x}$. $\square$

A.2 Proposed Algorithm and Theoretical Analysis

In order to facilitate the presentation we consider the following notation.

Notation 1. *We define the following notation for* $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$*:*

1. $\overline{\mathbf{x}} \equiv \overline{\mathbf{x}}(\mathbf{x}, \mathbf{y}) = \arg\min_{\mathbf{z} \in \mathcal{X}} \{f(\mathbf{z}, \mathbf{y}) + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|^2\}$
   $\overline{\mathbf{y}} \equiv \overline{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \arg\min_{\mathbf{z} \in \mathcal{Y}} \{-f(\mathbf{x}, \mathbf{z}) + \frac{L}{2}\|\mathbf{z} - \mathbf{y}\|^2\}$
   $\nabla_{\mathbf{x}} P(\mathbf{x}, \mathbf{y}) = L(\overline{\mathbf{x}} - \mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}, \overline{\mathbf{y}})$
   $\nabla_{\mathbf{y}} P(\mathbf{x}, \mathbf{y}) = L(\overline{\mathbf{y}} - \mathbf{y}) - \nabla_{\mathbf{y}} f(\overline{\mathbf{x}}, \mathbf{y})$
2. $\widetilde{\mathbf{x}} \equiv \widetilde{\mathbf{x}}(\mathbf{x}, \mathbf{y}, \mathbf{w}) = \arg\min_{\mathbf{z} \in \mathcal{X}} \{\widetilde{f}(\mathbf{z}, \mathbf{y}, \mathbf{w}) + \frac{L}{2}\|\mathbf{z} - \mathbf{x}\|^2\}$
   $\widetilde{\mathbf{y}} \equiv \widetilde{\mathbf{y}}(\mathbf{x}, \mathbf{y}, \mathbf{w}) = \arg\min_{\mathbf{z} \in \mathcal{Y}} \{-\widetilde{f}(\mathbf{x}, \mathbf{z}, \mathbf{w}) + \frac{L}{2}\|\mathbf{z} - \mathbf{y}\|^2\}$
   $\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x}, \mathbf{y}) = L(\widetilde{\mathbf{x}} - \mathbf{x}) + \widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \widetilde{\mathbf{y}}, \widetilde{\mathbf{w}})$
   $\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x}, \mathbf{y}) = L(\widetilde{\mathbf{y}} - \mathbf{y}) - \widetilde{\nabla}_{\mathbf{y}} f(\widetilde{\mathbf{x}}, \mathbf{y}, \widetilde{\mathbf{w}})$
3. $\widehat{\nabla}_{\mathbf{x}} P(\mathbf{x}, \mathbf{y}) = L(\widehat{\mathbf{x}} - \mathbf{x}) + \widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \widehat{\mathbf{y}}, \widetilde{\mathbf{w}})$
   $\widehat{\nabla}_{\mathbf{y}} P(\mathbf{x}, \mathbf{y}) = L(\widehat{\mathbf{y}} - \mathbf{y}) - \widetilde{\nabla}_{\mathbf{y}} f(\widehat{\mathbf{x}}, \mathbf{y}, \widetilde{\mathbf{w}})$,

where $\mathbf{w}, \widetilde{\mathbf{w}}$ *are random vectors whose elements are drawn from distribution* $\mathcal{W}$*, and* $\widehat{\mathbf{x}}, \widehat{\mathbf{y}}$ *are approximate solutions of the problems* $\widetilde{\mathbf{x}}(\mathbf{x}, \mathbf{y}, \mathbf{w})$*,* $\widetilde{\mathbf{y}}(\mathbf{x}, \mathbf{y}, \mathbf{w})$*, respectively, that will be specified later.*

Lemma 2. *For the mini-batch estimators of the objective* f *and its gradient* (22)-(24)*, and for any* $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$*,* $\mathbf{w} = (w_1, \ldots, w_n), w_i \sim \mathcal{W}, \forall i$*, it holds that*

$$\begin{aligned}
&\mathbb{E}[\widetilde{f}(\mathbf{x}, \mathbf{y}; \mathbf{w})] = f(\mathbf{x}, \mathbf{y}), \\
&\mathbb{E}[\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{w})] = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \\
&\mathbb{E}[\widetilde{\nabla}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{w})] = \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}).
\end{aligned} \tag{64}$$

Moreover, for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$*, it holds that*

$$\begin{aligned}
&\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{w}) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|^2] \leq \frac{\sigma^2}{n}, \\
&\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{w}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|^2] \leq \frac{\sigma^2}{n}.
\end{aligned} \tag{65}$$

Proof. Using Assumption 3 and taking expectation with respect to the random vector $\mathbf{w}$ we obtain

$$\mathbb{E}[\widetilde{f}(\mathbf{x}, \mathbf{y}; \mathbf{w})] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} F(\mathbf{x}, \mathbf{y}; w_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[F(\mathbf{x}, \mathbf{y}; w_i)] = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}),$$

and

$$\mathbb{E}[\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{w})] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; w_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; w_i)] = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}).$$

Using a similar argument we can prove that $\mathbb{E}[\widetilde{\nabla}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \mathbf{w})] = \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. Moreover, using the second part of Assumption 3 we get

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}; \mathbf{w}) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|^2] &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} (\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; w_i) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}))\|^2] \\
&= \frac{1}{n}\mathbb{E}[\|\nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}; w_1) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|^2] \leq \frac{\sigma^2}{n},
\end{aligned}$$

where we used the following elementary result from probability theory: $\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\|^2] = \frac{1}{n}\mathbb{E}[\|\mathbf{X}_1\|^2]$, with $\{\mathbf{X}_i\}_{i=1}^n$ zero-mean random vectors. Finally, the results with respect to $\mathbf{y}$ follow similarly. □

Proposition 3. *Suppose that Assumption 2 and 3 hold. For the stochastic gradient of P defined in* (25) *and for every $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ and $w \sim \mathcal{W}$, it holds that*

$$\mathbb{E}[\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y})] = \nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_x, \quad \textit{with } \|\mathbf{e}_x\| \leq \frac{L\sigma}{(L-L_x)\sqrt{n}} + \frac{L_x\sigma}{(L-L_y)\sqrt{n}}$$
$$\mathbb{E}[\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y})] = \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_y, \quad \textit{with } \|\mathbf{e}_y\| \leq \frac{L\sigma}{(L-L_y)\sqrt{n}} + \frac{L_y\sigma}{(L-L_x)\sqrt{n}}.$$

Moreover, for every $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$, we have that

$$\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y})\|^2] \leq \widetilde{\sigma}_x^2, \;\; \mathbb{E}[\|\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y})\|^2] \leq \widetilde{\sigma}_y^2, \tag{66}$$

where $\widetilde{\sigma}_x^2 = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_x)^2} + \frac{4L_x^2}{(L-L_y)^2} + 1\right)$ and $\widetilde{\sigma}_y^2 = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_y)^2} + \frac{4L_y^2}{(L-L_x)^2} + 1\right)$.

Proof. First of all, for $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$, we have that

$$\begin{aligned}\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) &= L[\arg\min_{\mathbf{z}\in\mathcal{X}}\{\widetilde{f}(\mathbf{z},\mathbf{y},\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\} - \mathbf{x}] + \widetilde{\nabla}_{\mathbf{x}} f\left(\mathbf{x}, \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-\widetilde{f}(\mathbf{x},\mathbf{z},\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}, \widetilde{\mathbf{w}}\right) = \\ &= L[\widetilde{\mathbf{x}} - \mathbf{x}] + \widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{\mathbf{w}}).\end{aligned}$$

Then, taking expectation over $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ yields

$$\begin{aligned}\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y})] &= -L\mathbf{x} + L\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\mathbf{x}}] + \mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{\mathbf{w}})] \\ &\overset{(a)}{=} -L\mathbf{x} + L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{x}}] + \mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}})] \\ &\overset{(b)}{=} -L\mathbf{x} + L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{x}} - \overline{\mathbf{x}}] + L\overline{\mathbf{x}} + \mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}}) - \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}})] + \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}}) = \\ &= \nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_{\mathbf{x}},\end{aligned} \tag{67}$$

where in (a) we used (64), in (b) we considered the fact that $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ are deterministic variables ($\overline{\mathbf{x}}, \overline{\mathbf{y}}, \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}$ are defined in Notation 1) and $\mathbf{e}_{\mathbf{x}} = L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{x}} - \overline{\mathbf{x}}] + \mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}}) - \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}})]$. Next, we obtain a bound for $\|\mathbf{e}_{\mathbf{x}}\|$, that is

$$\begin{aligned}\|\mathbf{e}_{\mathbf{x}}\| &\leq \|L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{x}} - \overline{\mathbf{x}}]\| + \|\mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}}) - \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}})]\| \\ &\leq L\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}} - \overline{\mathbf{x}}\|] + \mathbb{E}_{\mathbf{w}}[\|\nabla_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}}) - \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}})\|] \\ &\leq L\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}} - \overline{\mathbf{x}}\|] + L_x\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{y}} - \overline{\mathbf{y}}\|],\end{aligned}$$

where we used the Lipschitz gradient property of f.

In order to proceed we define the following functions :

$$g(\mathbf{z},\mathbf{x},\mathbf{y}) = f(\mathbf{z},\mathbf{y}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2$$
$$G(\mathbf{z},\mathbf{x},\mathbf{y},\mathbf{w}) = \widetilde{f}(\mathbf{z},\mathbf{y},\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2.$$

The above functions are strongly convex with respect to $\mathbf{z}$ with the same modulus $\mu_1 = L - L_y$, and their (global) minima are attained at $\mathbf{z} = \overline{\mathbf{x}}$ and $\mathbf{z} = \widetilde{\mathbf{x}}$, respectively. Therefore, we have that

$$\langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}},\mathbf{x},\mathbf{y}), \mathbf{z} - \overline{\mathbf{x}}\rangle \geq 0, \forall \mathbf{z} \in \mathcal{X} \Rightarrow \langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}},\mathbf{x},\mathbf{y}), \widetilde{\mathbf{x}} - \overline{\mathbf{x}}\rangle \geq 0$$
$$\langle \nabla_{\mathbf{z}} G(\widetilde{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \mathbf{z} - \widetilde{\mathbf{x}}\rangle \geq 0, \forall \mathbf{z} \in \mathcal{X} \Rightarrow \langle \nabla_{\mathbf{z}} G(\widetilde{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \overline{\mathbf{x}} - \widetilde{\mathbf{x}}\rangle \geq 0.$$

Also, the strong convexity of g implies that

$$\begin{aligned}
\mu_1\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2 &\leq \langle \nabla_{\mathbf{z}} G(\widetilde{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}) - \nabla_{\mathbf{z}} G(\overline{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\rangle \\
&\overset{(a)}{\leq} \langle \nabla_{\mathbf{z}} G(\overline{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \overline{\mathbf{x}}-\widetilde{\mathbf{x}}\rangle + \langle \nabla_{\mathbf{z}} G(\widetilde{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\rangle + \langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}},\mathbf{x},\mathbf{y}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\rangle \\
&\overset{(b)}{\leq} \langle \nabla_{\mathbf{z}} g(\overline{\mathbf{x}},\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{z}} G(\overline{\mathbf{x}},\mathbf{x},\mathbf{y},\mathbf{w}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\rangle \\
&\overset{(c)}{=} \left\langle \nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) + L(\overline{\mathbf{x}}-\mathbf{x}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w}) - L(\overline{\mathbf{x}}-\mathbf{x}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\right\rangle \\
&= \left\langle \nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w}), \widetilde{\mathbf{x}}-\overline{\mathbf{x}}\right\rangle \\
&\overset{(d)}{\leq} \|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|,
\end{aligned}$$

where in (a) and (b) we used the fact that $\overline{\mathbf{x}}$ and $\widetilde{\mathbf{x}}$ are the minima of g and G, respectively, in (c) we computed the gradients of g and G w.r.t $\mathbf{z}$, and in (d) the Cauchy-Schwarz inequality is utilized. The above inequality implies that

$$\mu_1\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\| \leq \|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|. \tag{68}$$

Taking expectation w.r.t $\mathbf{w}$ leads to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}}[\mu_1\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|] \leq \mathbb{E}_{\mathbf{w}}[\|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|] \overset{(a)}{\leq} \sqrt{\mathbb{E}_{\mathbf{w}}[\|\nabla_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y}) - \widetilde{\nabla}_{\mathbf{z}} f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|^2]} \overset{(b)}{\leq} \frac{\sigma}{\sqrt{n}} \\
&\Rightarrow \mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|] \leq \frac{\sigma}{\mu_1\sqrt{n}},
\end{aligned} \tag{69}$$

where the property in (a) follows from $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$ and in (b) the bound (65) is used. Following similar steps we can deduce that

$$\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{y}}-\overline{\mathbf{y}}\|] \leq \frac{\sigma}{\mu_2\sqrt{n}}, \tag{70}$$

where $\mu_2 = L - L_y$. Combining the above we get

$$\|\mathbf{e}_{\mathbf{x}}\| \leq \frac{L\sigma}{(L-L_x)\sqrt{n}} + \frac{L_x\sigma}{(L-L_y)\sqrt{n}}.$$

Next, w.r.t $\mathbf{y}$ we have that

$$\begin{aligned}
\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) &= L[\arg\min_{\mathbf{z}\in\mathcal{Y}}\{-\widetilde{f}(\mathbf{x},\mathbf{z},\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\} - \mathbf{y}] - \widetilde{\nabla}_{\mathbf{y}} f\left(\arg\min_{\mathbf{z}\in\mathcal{X}}\{\widetilde{f}(\mathbf{z},\mathbf{y},\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}, \mathbf{y}, \widetilde{\mathbf{w}}\right) = \\
&= L[\widetilde{\mathbf{y}}-\mathbf{y}] - \widetilde{\nabla}_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y},\widetilde{\mathbf{w}})
\end{aligned}$$

Then, taking expectation over $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ and utilizing the same arguments as in expression (67) we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y})] &= -L\mathbf{y} + L\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\mathbf{y}}] - \mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\widetilde{\nabla}_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y},\widetilde{\mathbf{w}})] \\
&= -L\mathbf{y} + L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{y}}] - \mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y})] \\
&= -L\mathbf{y} + L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{y}}-\overline{\mathbf{y}}] + L\overline{\mathbf{y}} + \mathbb{E}_{\mathbf{w}}[\nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y}) - \nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y})] - \nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y}) \\
&= \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_{\mathbf{y}},
\end{aligned}$$

where $\mathbf{e}_{\mathbf{y}} = L\mathbb{E}_{\mathbf{w}}[\widetilde{\mathbf{y}}-\overline{\mathbf{y}}] + \mathbb{E}[\nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y}) - \nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y})]$. Following a similar reasoning as above we deduce that

$$\|\mathbf{e}_y\| \leq \frac{L\sigma}{(L-L_y)\sqrt{n}} + \frac{L_y\sigma}{(L-L_x)\sqrt{n}}.$$

The first part of the proof is complete. So, we proceed by showing the bound in (66). To begin with, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\|\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{x}}P(\mathbf{x},\mathbf{y})\|^2] &= \mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\|L[\widetilde{\mathbf{x}}-\mathbf{x}]+\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{\mathbf{w}})-L[\overline{\mathbf{x}}-\mathbf{x}]-\nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\|^2] \\
&\overset{(a)}{=} \mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[2L^2\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2+2\|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{\mathbf{w}})-\nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\|^2] \\
&\overset{(b)}{=} \mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[2L^2\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2+4\|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\widetilde{\mathbf{y}},\mathbf{w})-\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}},\mathbf{w})\|^2 \\
&+4\|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}},\widetilde{\mathbf{w}})-\nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\|^2] \\
&\leq 2L^2\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2]+4\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\widetilde{\mathbf{y}},\widetilde{\mathbf{w}})-\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}},\widetilde{\mathbf{w}})\|^2]+ \\
&+4\mathbb{E}_{\widetilde{\mathbf{w}}}[\|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}},\widetilde{\mathbf{w}})-\nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\|^2] \\
&\overset{(c)}{\leq} 2L^2\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2]+4L_x^2\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{y}}-\overline{\mathbf{y}}\|^2]+4\frac{\sigma^2}{n} \qquad (71)
\end{aligned}$$

where in (a) we used the property $(a+b)^2 \leq 2a^2+2b^2$; in (b) we added and subtracted the term $\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}},\widetilde{\mathbf{w}})$, and used the same property as in (a); in (c) we used the Lipschitz gradient property of F and the bound in (65). Moreover, from the inequality in (68) we can obtain

$$\begin{aligned}
&\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2 \leq \frac{1}{\mu_1^2}\|\nabla_{\mathbf{z}}f(\overline{\mathbf{x}},\mathbf{y})-\widetilde{\nabla}_{\mathbf{z}}f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|^2 \\
&\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{x}}-\overline{\mathbf{x}}\|^2] \leq \frac{1}{\mu_1^2}\mathbb{E}_{\mathbf{w}}[\|\nabla_{\mathbf{z}}f(\overline{\mathbf{x}},\mathbf{y})-\widetilde{\nabla}_{\mathbf{z}}f(\overline{\mathbf{x}},\mathbf{y},\mathbf{w})\|^2] \leq \frac{\sigma^2}{\mu_1^2 n}, \qquad (72)
\end{aligned}$$

where the bound in (65) is used. Similarly, it holds that

$$\begin{aligned}
&\|\widetilde{\mathbf{y}}-\overline{\mathbf{y}}\|^2 \leq \frac{1}{\mu_2^2}\|\nabla_{\mathbf{z}}f(\mathbf{x},\overline{\mathbf{y}})-\widetilde{\nabla}_{\mathbf{z}}f(\mathbf{x},\overline{\mathbf{y}},\mathbf{w})\|^2 \Rightarrow \\
&\mathbb{E}_{\mathbf{w}}[\|\widetilde{\mathbf{y}}-\overline{\mathbf{y}}\|^2] \leq \frac{1}{\mu_2^2}\mathbb{E}_{\mathbf{w}}[\|\nabla_{\mathbf{z}}f(\mathbf{x},\overline{\mathbf{y}})-\widetilde{\nabla}_{\mathbf{z}}f(\mathbf{x},\overline{\mathbf{y}},\mathbf{w})\|^2] \leq \frac{\sigma^2}{\mu_2^2 n}. \qquad (73)
\end{aligned}$$

Combining (71), (72) and (73) yields

$$\mathbb{E}_{\mathbf{w},\widetilde{\mathbf{w}}}[\|\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{x}}P(\mathbf{x},\mathbf{y})\|^2] \leq 2L^2\frac{\sigma^2}{\mu_1^2 n}+4L_x^2\frac{\sigma^2}{\mu_2^2 n}+4\frac{\sigma^2}{n} = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_x)^2}+\frac{4L_x^2}{(L-L_y)^2}+4\right) := \widetilde{\sigma}_x^2.$$

Finally, following the same reasoning

$$\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}P(\mathbf{x},\mathbf{y})\|^2] \leq 2L^2\frac{\sigma^2}{\mu_2^2 n}+4L_y^2\frac{\sigma^2}{\mu_1^2 n}+4\frac{\sigma^2}{n} = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_y)^2}+\frac{4L_y^2}{(L-L_x)^2}+4\right) := \widetilde{\sigma}_y^2.$$

□

Theorem 1. *Suppose that Assumption 2 and 3 hold. In addition, assume that the gradients of f are bounded, that is $\|\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y})\| \leq c_x$ and $\|\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})\| \leq c_y$, for every $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$. We run Algorithm 1 for T iterations, with constant stepisize $0 \leq \alpha < \frac{2}{3\overline{L}}$, $L > \max\{L_x, L_y\}$ and for given parameters δ_x, δ_y. Then, we have*

$$\frac{1}{T}\sum_{r=0}^{T-1}\mathbb{E}[\|\mathbf{G}_\alpha^r\|^2] \leq \frac{\mathbb{E}[P^0]/\overline{a}}{T}+\frac{3\overline{L}\alpha^2}{2\overline{\alpha}}\delta^2+\frac{c\alpha}{\overline{\alpha}}\delta+\frac{c\alpha}{\overline{\alpha}}\hat{\sigma}+\frac{3\overline{L}\alpha^2}{\overline{\alpha}}\widetilde{\sigma}^2,$$

where $\overline{\alpha} = \alpha - \frac{3\overline{L}\alpha^2}{2}$, $\delta = 2(L+L_y)\delta_x + 2(L+L_x)\delta_y$, $\hat{\sigma} = \tilde{\sigma}_x + \tilde{\sigma}_y$, $\tilde{\sigma}^2 = 2(\tilde{\sigma}_x^2+\tilde{\sigma}_y^2)$ and $c = 4LD + c_x + c_y$.

Proof. Consider the following notation:

$$\begin{aligned}
&\mathbf{z}^r = (\mathbf{x}^r, \mathbf{y}^r) \\
&\nabla P(\mathbf{z}^r) = [\nabla_{\mathbf{x}}P(\mathbf{z}^r), \nabla_{\mathbf{y}}P(\mathbf{z}^r)]^T \\
&\widetilde{\nabla} P(\mathbf{z}^r) = [\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r), \widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)]^T \\
&\widehat{\nabla} P(\mathbf{z}^r) = [\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r), \widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)]^T
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{G}_a^r &= \frac{1}{\alpha}\left[\mathbf{z}^r - \begin{bmatrix}\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right)\\ \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\nabla_{\mathbf{y}}P(\mathbf{z}^r)\right)\end{bmatrix}\right],\\
\widetilde{\mathbf{G}}_a^r &= \frac{1}{\alpha}\left[\mathbf{z}^r - \begin{bmatrix}\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right)\\ \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right)\end{bmatrix}\right],\\
\widehat{\mathbf{G}}_a^r &= \frac{1}{\alpha}\left[\mathbf{z}^r - \begin{bmatrix}\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right)\\ \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right)\end{bmatrix}\right].
\end{aligned} \tag{74}$$

Moreover, one iteration of the Algorithm 1 can be written as

$$\mathbf{z}^{r+1} = \begin{bmatrix}\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right)\\ \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right)\end{bmatrix}. \tag{75}$$

Also, let denote with $\mathcal{F}^r$ the history up to iteration r, that is the iterates $\{(\mathbf{x}^r,\mathbf{y}^r),\dots,(\mathbf{x}^0,\mathbf{y}^0)\}$. Note that conditioning under $\mathcal{F}^r$ means that $\{(\mathbf{w}^{r-1},\widetilde{\mathbf{w}}^{r-1}),\dots,(\mathbf{w}^0,\widetilde{\mathbf{w}}^0)\}$ are not random variables; though $(\mathbf{w}^r,\widetilde{\mathbf{w}}^r)$ is still a random variable.

To begin with, we know that the function $P(\mathbf{z})$ has Lipschitz continuous gradient with constant $\overline{L}$. Thus, from the descent lemma we get

$$\begin{aligned}
P(\mathbf{z}^{r+1}) &\le P(\mathbf{z}^r) + \left\langle \nabla P(\mathbf{z}^r), \mathbf{z}^{r+1}-\mathbf{z}^r\right\rangle + \frac{\overline{L}}{2}\|\mathbf{z}^{r+1}-\mathbf{z}^r\|^2\\
P(\mathbf{z}^{r+1}) &\overset{(a)}{\le} P(\mathbf{z}^r) - \alpha\left\langle \nabla P(\mathbf{z}^r), \widehat{\mathbf{G}}_a^r\right\rangle + \frac{\overline{L}\alpha^2}{2}\|\widehat{\mathbf{G}}_a^r\|^2\\
P(\mathbf{z}^{r+1}) &\le P(\mathbf{z}^r) - \alpha\left\langle \nabla P(\mathbf{z}^r), \widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\right\rangle - \alpha\left\langle \nabla P(\mathbf{z}^r), \widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\right\rangle - \alpha\left\langle \nabla P(\mathbf{z}^r), \mathbf{G}_a^r\right\rangle +\\
&\quad + \frac{\overline{L}\alpha^2}{2}\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r + \widetilde{\mathbf{G}}_a^r + \mathbf{G}_a^r - \mathbf{G}_a^r\|^2\\
P(\mathbf{z}^{r+1}) &\overset{(b)}{\le} P(\mathbf{z}^r) - \alpha\left\langle \nabla P(\mathbf{z}^r), \widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\right\rangle - \alpha\left\langle \nabla P(\mathbf{z}^r), \widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\right\rangle - \alpha\left\langle \nabla P(\mathbf{z}^r), \mathbf{G}_a^r\right\rangle +\\
&\quad + \frac{3\overline{L}\alpha^2}{2}\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|^2 + \frac{3\overline{L}\alpha^2}{2}\|\widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\|^2 + \frac{3\overline{L}\alpha^2}{2}\|\mathbf{G}_a^r\|^2,
\end{aligned} \tag{76}$$

where in (a) we combine (74) and (75), and in (b) we used the inequality $(a+b+c)^3 \le 3a^3+3b^3+3c^3$. Then, the plan is to upper bound all the terms of the rhs of the above inequality.

First, our aim is to bound the quantities $\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|]$ and $\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\|]$. For the former bound we consider the following inequalities,

$$\begin{aligned}
\|\widehat{\nabla}P(\mathbf{z}^r) - \widetilde{\nabla}P(\mathbf{z}^r)\| &\le \|\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r) - \widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\| + \|\widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r) - \widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\| =\\
&= \|L(\widehat{\mathbf{x}}^r - \mathbf{x}^r) + \widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x}^r,\widehat{\mathbf{y}}^r,\widetilde{\mathbf{w}}^r) - L(\widetilde{\mathbf{x}}^r - \mathbf{x}^r) - \widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x}^r,\widetilde{\mathbf{y}}^r,\widetilde{\mathbf{w}}^r)\|\\
&\quad + \|L(\widehat{\mathbf{y}}^r - \mathbf{y}^r) - \widetilde{\nabla}_{\mathbf{y}}f(\widehat{\mathbf{x}}^r,\mathbf{y}^r,\widetilde{\mathbf{w}}^r) - L(\widetilde{\mathbf{y}}^r - \mathbf{y}^r) + \widetilde{\nabla}_{\mathbf{y}}f(\widetilde{\mathbf{x}}^r,\mathbf{y}^r,\widetilde{\mathbf{w}}^r)\|\\
&\le L\|\widehat{\mathbf{x}}^r - \widetilde{\mathbf{x}}^r\| + \|\widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x}^r,\widehat{\mathbf{y}}^r,\widetilde{\mathbf{w}}^r) - \widetilde{\nabla}_{\mathbf{x}}f(\mathbf{x}^r,\widetilde{\mathbf{y}}^r,\widetilde{\mathbf{w}}^r)\| +\\
&\quad + L\|\widehat{\mathbf{y}}^r - \widetilde{\mathbf{y}}^r\| + \|\widetilde{\nabla}_{\mathbf{y}}f(\widehat{\mathbf{x}}^r,\mathbf{y}^r,\widetilde{\mathbf{w}}^r) - \widetilde{\nabla}_{\mathbf{y}}f(\widetilde{\mathbf{x}}^r,\mathbf{y}^r,\widetilde{\mathbf{w}}^r)\|\\
&\le L\|\widehat{\mathbf{x}}^r - \widetilde{\mathbf{x}}^r\| + L_x\|\widehat{\mathbf{y}}^r - \widetilde{\mathbf{y}}^r\| + L\|\widehat{\mathbf{y}}^r - \widetilde{\mathbf{y}}^r\| + L_y\|\widehat{\mathbf{x}}^r - \widetilde{\mathbf{x}}^r\|\\
&\le (L+L_y)\delta_x + (L+L_x)\delta_y,
\end{aligned}$$

where in the third inequality the Lipschitz gradient property of F is used, while in the fourth one we defined

$\delta_x = \|\widehat{\mathbf{x}}^r - \widetilde{\mathbf{x}}^r\|$ and $\delta_y = \|\widehat{\mathbf{y}}^r - \widetilde{\mathbf{y}}^r\|$. As a result,

$$\begin{aligned}\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\| &\overset{(a)}{\leq} \left\|\frac{1}{\alpha}\left[\mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right)\right] - \frac{1}{\alpha}\left[\mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right)\right]\right\| \\ &+ \left\|\frac{1}{\alpha}\left[\mathbf{y}^r - \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right)\right] - \frac{1}{\alpha}\left[\mathbf{y}^r - \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right)\right]\right\| \\ &\overset{(b)}{\leq} \left\|\widehat{\nabla}_{\mathbf{x}}P(\mathbf{z}^r) - \widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r)\right\| + \left\|\widehat{\nabla}_{\mathbf{y}}P(\mathbf{z}^r) - \widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r)\right\| \\ &\leq 2\|\widehat{\nabla}P(\mathbf{z}^r) - \widetilde{\nabla}P(\mathbf{z}^r)\| \leq 2(L+L_y)\delta_x + 2(L+L_x)\delta_y, \end{aligned} \tag{77}$$

where in (a) we used the property $\|[\mathbf{x},\mathbf{y}]^T\| = \sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2} \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, and in (b) we used the non-expansiveness of the projection operator. Thus,

$$\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|] \leq 2(L+L_y)\delta_x + 2(L+L_x)\delta_y := \delta. \tag{78}$$

Next, we see that

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\|] &\overset{(a)}{\leq} \mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r) - \nabla_{\mathbf{x}}P(\mathbf{z}^r)\|] + \mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r) - \nabla_{\mathbf{y}}P(\mathbf{z}^r)\|] \\ &\overset{(b)}{\leq} \sqrt{\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r) - \nabla_{\mathbf{x}}P(\mathbf{z}^r)\|^2]} + \sqrt{\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r) - \nabla_{\mathbf{y}}P(\mathbf{z}^r)\|^2]} \\ &\overset{(c)}{\leq} \widetilde{\sigma}_x + \widetilde{\sigma}_y := \hat{\sigma}, \end{aligned} \tag{79}$$

where in (a) we applied the reasoning used in (77), in (b) the property $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$ and in (c) we used the bounds from (66).

Moreover, for the bound of the gradient of P we have

$$\begin{aligned}\|\nabla P(\mathbf{z})\| &\leq \|L(\overline{\mathbf{x}} - \mathbf{x}) + \nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\| + \|L(\overline{\mathbf{y}} - \mathbf{y}) - \nabla_{\mathbf{y}}f(\overline{\mathbf{x}},\mathbf{y})\| \\ &\leq L[\|\overline{\mathbf{x}}\| + \|\mathbf{x}\|] + L[\|\overline{\mathbf{y}}\| + \|\mathbf{y}\|] + \|\nabla_{\mathbf{x}}f(\mathbf{x},\overline{\mathbf{y}})\| + \|\nabla_{\mathbf{y}}f(\overline{\mathbf{x}},\mathbf{y})\| \\ &\leq 2LD + 2LD + c_x + c_y := c. \end{aligned} \tag{80}$$

Next, notice that $\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\mathbf{G}_a^r\|^2] = \|\mathbf{G}_a^r\|^2$, (77) implies that $\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|^2] \leq \delta^2$ and

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r\|^2] &\overset{(a)}{\leq} 2\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}P(\mathbf{z}^r) - \nabla P(\mathbf{z}^r)\|^2] \\ &= 2\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{x}}P(\mathbf{z}^r) - \nabla_{\mathbf{x}}P(\mathbf{z}^r)\|^2] + 2\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\widetilde{\nabla}_{\mathbf{y}}P(\mathbf{z}^r) - \nabla_{\mathbf{y}}P(\mathbf{z}^r)\|^2] \\ &\overset{(b)}{\leq} 2\widetilde{\sigma}_x^2 + 2\widetilde{\sigma}_y^2 := \widetilde{\sigma}^2, \end{aligned} \tag{81}$$

where in (a) we used the inequality in (79), while in (b) we used the bounds in (66). Furthermore, using the Cauchy-Schwarz inequality, and the results in (78), (79), (80) we can see that

$$\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\left\langle \nabla P(\mathbf{z}^r), \widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r \right\rangle] \geq -\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\nabla P(\mathbf{z}^r)\|\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|] \geq -c\delta, \tag{82}$$

$$\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\left\langle \nabla P(\mathbf{z}^r), \widetilde{\mathbf{G}}_a^r - \mathbf{G}_a^r \right\rangle] \geq -\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\|\nabla P(\mathbf{z}^r)\|\|\widehat{\mathbf{G}}_a^r - \widetilde{\mathbf{G}}_a^r\|] \geq -c\hat{\sigma}. \tag{83}$$

Finally, in order to bound the expectation of $\langle \nabla P(\mathbf{z}^r), \mathbf{G}_a^r \rangle$ consider the projection operation (w.r.t $\mathbf{x}$) which can be written as

$$\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right) = \arg\min_{\mathbf{u}\in\mathcal{X}}\{\langle \nabla P(\mathbf{z}^r), \mathbf{u} - \mathbf{x}^r\rangle + \frac{1}{2\alpha}\|\mathbf{u} - \mathbf{x}^r\|^2\}.$$

The optimality condition of the above problem implies that

$$\left\langle \nabla_{\mathbf{x}}P(\mathbf{z}^r) + \frac{1}{\alpha}(\text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right) - \mathbf{x}^r), \mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right) \right\rangle \geq 0$$

$$\langle \nabla_{\mathbf{x}}P(\mathbf{z}^r), \mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right)\rangle \geq \frac{1}{\alpha}\|\mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{z}^r)\right)\|^2.$$

Deriving the respective expression w.r.t $\mathbf{y}$ and combining it with the one above yields

$$\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\langle \nabla P(\mathbf{z}^r), \mathbf{G}_a^r \rangle] = \langle \nabla P(\mathbf{z}^r), \mathbf{G}_a^r \rangle \geq \|\mathbf{G}_a^r\|^2. \tag{84}$$

Taking expectations on both sides of (76) conditioned on $\mathcal{F}^r$ and using the results from (81), (82), (83), (84) we get

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[P(\mathbf{z}^{r+1})] &\leq P(\mathbf{z}^r) + \alpha c(\delta + \hat{\sigma}) - \alpha \|\mathbf{G}_a^r\|^2 + \frac{3\overline{L}\alpha^2}{2}\delta^2 + \frac{3\overline{L}\alpha^2}{2}\widetilde{\sigma}^2 + \frac{3\overline{L}\alpha^2}{2}\|\mathbf{G}_a^r\|^2 \\
\left(\alpha - \frac{3\overline{L}\alpha^2}{2}\right)\|\mathbf{G}_a^r\|^2 &\leq P(\mathbf{z}^r) - \mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[P(\mathbf{z}^{r+1})] + \alpha c(\delta + \hat{\sigma}) + \frac{3\overline{L}\alpha^2}{2}(\delta^2 + \widetilde{\sigma}^2).
\end{aligned} \tag{85}$$

Taking expectation over $\mathcal{F}^r$ (we denote the total expectation $\mathbb{E}_{\mathcal{F}^r}[\mathbb{E}_{\mathbf{w}^r|\mathcal{F}^r}[\cdot]]$ as $\mathbb{E}[\cdot]$) and summing over $r = 0, \ldots, T-1$ we obtain

$$\begin{aligned}
\left(\alpha - \frac{3\overline{L}\alpha^2}{2}\right)\sum_{r=0}^{T-1}\mathbb{E}[\|\mathbf{G}_a^r\|^2] &\leq \mathbb{E}[P(\mathbf{z}^0)] - \mathbb{E}[P(\mathbf{z}^T)] + \sum_{r=0}^{T-1}\left[\alpha c(\delta + \hat{\sigma}) + \frac{3\overline{L}\alpha^2}{2}(\delta^2 + \widetilde{\sigma}^2)\right] \\
\frac{1}{T}\sum_{r=0}^{T-1}\mathbb{E}[\|\mathbf{G}_\alpha^r\|^2] &\leq \frac{\mathbb{E}[P^0]/\overline{a}}{T} + \frac{3\overline{L}\alpha^2}{2\overline{\alpha}}\delta^2 + \frac{c\alpha}{\overline{\alpha}}\delta + \frac{c\alpha}{\overline{\alpha}}\hat{\sigma} + \frac{3L\alpha^2}{\overline{\alpha}}\widetilde{\sigma}^2,
\end{aligned}$$

where we used the fact that $P(\mathbf{z}) \geq 0$ and we set $\overline{\alpha} = \alpha - \frac{3\overline{L}\alpha^2}{2}$. □

B EXTENSION OF THE HAMILTONIAN METHOD TO CONSTRAINED GAMES

The Hamiltonian method for finding FNEs of unconstrained games, or more precisely stationary points, has been analyzed in literature (Abernethy et al., 2019; Loizou et al., 2020). However, a major limitation of this approach is the fact that it cannot be directly utilized to min-max games with constraints. Below we propose a formulation which can be seen as an extension of the Hamiltonian method to the constrained case. Specifically, consider the following objective

$$\widetilde{H}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \text{proj}_{\mathcal{X}}\left(\mathbf{x} - \alpha\nabla_x f(\mathbf{x}, \mathbf{y})\right)\|^2 + \frac{1}{2}\|\mathbf{y} - \text{proj}_{\mathcal{Y}}\left(\mathbf{y} + \beta\nabla_y f(\mathbf{x}, \mathbf{y})\right)\|^2 \tag{86}$$

and the respective optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} \widetilde{H}(\mathbf{x}, \mathbf{y}). \tag{87}$$

Note that for $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$ (unconstrained problem) and $\alpha = 1$, $\beta = 1$ the objective (86) reduces to the Hamiltonian (7), that is

$$\begin{aligned}
\widetilde{H}(\mathbf{x}, \mathbf{y}) &= \frac{1}{2}\|\mathbf{x} - \mathbf{x} + \alpha\nabla_x f(\mathbf{x}, \mathbf{y})\|^2 + \frac{1}{2}\|\mathbf{y} - \mathbf{y} - \beta\nabla_y f(\mathbf{x}, \mathbf{y})\|^2 = \\
&= \frac{1}{2}\|\nabla_x f(\mathbf{x}, \mathbf{y})\|^2 + \frac{1}{2}\|\nabla_y f(\mathbf{x}, \mathbf{y})\|^2 = H(\mathbf{x}, \mathbf{y}).
\end{aligned}$$

For the sake of completeness we are going to establish formally the equivalence between the FNEs of (1) and the global minima of the (constrained) Hamiltonian $\widetilde{H}(\mathbf{x}, \mathbf{y})$ [eq. (86)].

Proposition 4. *Suppose that Assumption 1 holds. Then, the function $\widetilde{H}(\mathbf{x}, \mathbf{y})$ possesses the following properties:*

1. *The global minimum of $\widetilde{H}(\mathbf{x}, \mathbf{y})$ is 0.*
2. *A point $(\mathbf{x}^*, \mathbf{y}^*)$ is a FNE of (1) if and only if $(\mathbf{x}^*, \mathbf{y}^*)$ is a global minimum of $\widetilde{H}(\mathbf{x}, \mathbf{y})$.*

Proof. First of all, note that Assumption 1 ensures the existence of an FNE of the game (1) (Nouiehed et al., 2019, Theorem 2.2, pg.3). Secondly, consider the optimality conditions of the projection operators involved in the formulation of $\widetilde{H}(\mathbf{x},\mathbf{y})$:

$$\begin{aligned}&\hat{\mathbf{p}}_x(\mathbf{x},\mathbf{y}) = \text{proj}_{\mathcal{X}}\left(\mathbf{x} - \alpha\nabla_x f(\mathbf{x},\mathbf{y})\right) = \min_{\mathbf{p}\in\mathcal{X}}\{\frac{1}{2}\|\mathbf{p} - \mathbf{x} + \alpha\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|^2\}\\ &\Leftrightarrow \langle \hat{\mathbf{p}}_x(\mathbf{x},\mathbf{y}) - \mathbf{x} + \alpha\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}), \mathbf{p} - \hat{\mathbf{p}}_x(\mathbf{x},\mathbf{y})\rangle \geq 0, \forall \mathbf{p}\in\mathcal{X} \end{aligned} \tag{88}$$

and

$$\begin{aligned}&\hat{\mathbf{p}}_y(\mathbf{x},\mathbf{y}) = \text{proj}_{\mathcal{Y}}\left(\mathbf{y} + \beta\nabla_y f(\mathbf{x},\mathbf{y})\right) = \min_{\mathbf{p}\in\mathcal{Y}}\{\frac{1}{2}\|\mathbf{p} - \mathbf{y} - \beta\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|^2\}\\ &\Leftrightarrow \langle \hat{\mathbf{p}}_y(\mathbf{x},\mathbf{y}) - \mathbf{y} - \beta\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}), \mathbf{p} - \hat{\mathbf{p}}_y(\mathbf{x},\mathbf{y})\rangle \geq 0, \forall \mathbf{p}\in\mathcal{Y}. \end{aligned} \tag{89}$$

Then, assume that $(\mathbf{x}^*,\mathbf{y}^*)$ is an FNE of (1). Equivalently, it holds that

$$\begin{aligned}&\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq 0, \forall \mathbf{x}\in\mathcal{X} \Leftrightarrow \langle \mathbf{x}^* - \mathbf{x}^* + \alpha\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq 0, \forall \mathbf{x}\in\mathcal{X}\\ &\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{y} - \mathbf{y}^*\rangle \leq 0, \forall \mathbf{y}\in\mathcal{Y} \Leftrightarrow \langle \mathbf{y}^* - \mathbf{y}^* - \beta\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{y} - \mathbf{y}^*\rangle \geq 0, \forall \mathbf{y}\in\mathcal{Y}. \end{aligned}$$

From the optimality conditions of the projection operator (88), (89), equivalently we get

$$\begin{aligned}\mathbf{x}^* &= \text{proj}_{\mathcal{X}}\left(\mathbf{x}^* - \alpha\nabla_x f(\mathbf{x}^*,\mathbf{y}^*)\right)\\ \mathbf{y}^* &= \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^* + \beta\nabla_y f(\mathbf{x}^*,\mathbf{y}^*)\right).\end{aligned}$$

Finally, notice that $\widetilde{H}(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \text{proj}_{\mathcal{X}}\left(\mathbf{x} - \alpha\nabla_x f(\mathbf{x},\mathbf{y})\right), \mathbf{y} = \text{proj}_{\mathcal{Y}}\left(\mathbf{y} + \beta\nabla_y f(\mathbf{x},\mathbf{y})\right)$ and $\widetilde{H}(\mathbf{x},\mathbf{y}) \geq 0,\ \forall \mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}$. The proof is now complete. □