Appendix for Improved Analysis of Penalty-Based Methods for Bilevel Optimization with Coupled Constraints

Liuyuan Jiang, Quan Xiao, Tianyi Chen
Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
{jiangl7,xiaoq5,chent18}@rpi.edu

Algorithm 3 ALT-PBGD

1: **inputs:** initial point x_0; stepsize η; counters T; inner Min Solver.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: update $y_t^g \approx \arg\min_{y \in \mathcal{Y}(x)} g(x_t, y)$ by Min Solver.
4: find $g_t = (\nabla_x f(x_t, y_t) + \gamma \nabla_x g(x_t, y_t) - \gamma \nabla_x g(x_t, y_t^g)$, update $x_{t+1} = \mathrm{Proj}_{\mathcal{X}}\big(x_t - \eta\big) g_t$, and $y_{t+1} = \mathrm{Proj}_{\mathcal{Y}(x)}\big(y_t - \eta(\nabla_x f(x_t, y_t) + \gamma \nabla_x g(x_t, y_t))\big)$.
5: **end for**
6: **outputs:** (x_T, y_T^g)

I. Experiments

In this section, we verify our main theoretical results using toy problems. Section I-A corroborates Proposition 2 and Section I-B illustrates the results in Proposition 4.

A. Comparison of ALT-PBGD and JNT-PBGD

We conduct experiments comparing the performance of ALT-PBGD and JNT-PBGD (as described in [6, Algorithm 2]) on solving Example 1. For clarity, we present JNT-PBGD in Algorithm 3.

In the experiments, we set $\gamma = 10$ with an initial point of $(x_0 = 0, y_0 = 0)$. Both ALT-PBGD and JNT-PBGD are applied with a safe step-size choice of $\eta = 0.01$, and we also test with a larger step-size, $\eta = 0.1$, for both methods. As shown in Figure 2, both algorithms exhibit similar convergence rates for the smaller, safe step-size of $\eta = 0.01$. However, when using $\eta = 0.01$, JNT-PBGD fails to converge, while ALT-PBGD performs well, achieving faster convergence within 5 steps. This observation aligns with the theoretical results in Lemma 3 and Proposition 2, which indicate that ALT-PBGD can tolerate a larger γ without requiring a smaller step-size due to the smoothness constant remaining unaffected by the scaling of γ.

B. Comparison of BLOCC with different γ

We illustrate Proposition 4 using Example 2 by applying BLOCC from Algorithm 2 [4] to solve $F_\gamma(x)$ with $\gamma = 10$ and $\gamma = 100$. The step-size choices include a safe option $\eta = \frac{1}{10\gamma}$ and a larger value $\eta = 0.05$.

Identify applicable funding agency here. If none, delete this.

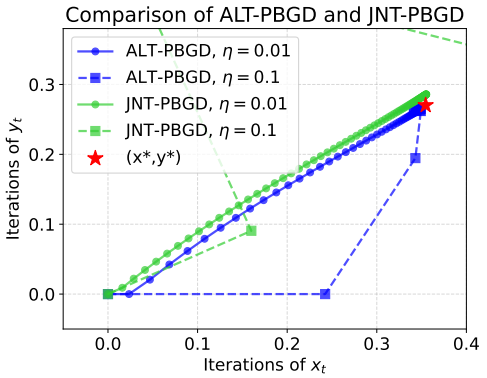

Fig. 2: Convergence performance for Example 1, comparing the iterations of (x_t, y_t) using ALT-PBGD from Algorithm 1 with joint PBGD applied to $\min_{x,y} H_\gamma(x, y)$, as JNT-PBGD in [6, Algorithm 2].

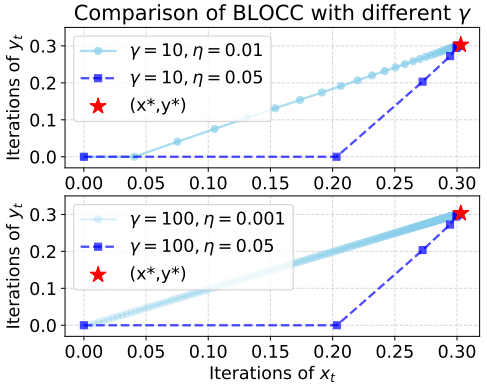

Fig. 3: Convergence performance for Example 2, comparing the iterations of (x_t, y_t) using BLOCC from Algorithm 2 [4] to solve $F_\gamma(x)$ with $\gamma = 10, 100$ and varying step-sizes.

Example 2. *With* $\mathcal{X} = [0, 3]$, $\mathcal{Y}(x) = \{y \in [0, 3] : y - x \leq 0\}$ *consider the coupled constrained BLO problem in* (1) *with the objectives as follows*

$$f(x, y) = \frac{\exp(-y + 2)}{(2 + \cos(4x)} + \frac{1}{2}\ln((4x - 2)^2 + 1) + x^2, \quad \text{(1a)}$$

$$g(x, y) = (y - 2x)^2. \quad \text{(1b)}$$

As shown in Figure 3, the algorithm converges for both $\gamma = 10$ and $\gamma = 100$ using either the large step-size $\eta = 0.05$ or the safe step-size $\eta = \frac{1}{10\gamma}$. Notably, using the larger step-size $\eta = 0.05$ results in faster convergence to the optimal solution, requiring fewer iterations compared to the literature. This supports the conclusion that $F_\gamma(x)$ exhibits constant-level smoothness, leading to improved complexity as stated in Proposition 4.

II. Preliminary Knowledge

Definition 1 (Lipschitz Continuity and Smoothness)**.** *Let $h : \mathcal{Q} \to \mathbb{R}$. We say $h(q)$ is $l_{h,0}$-Lipschitz on $\mathcal{S} \subseteq \mathcal{Q}$ if*

$$\|h(q_1) - h(q_2)\| \leq l_{h,1}\|q_1 - q_2\|, \quad \forall q_1, q_2 \in \mathcal{S}. \tag{2}$$

If $h(q)$ is differentiable on $\mathcal{S}$, we say $h(q)$ is $l_{h,1}$-smooth on $\mathcal{S}$ if $\nabla h(q)$ is $l_{h,1}$-Lipschitz on $\mathcal{S}$.

Definition 2 (Proximal Gradient)**.** *Let $h : \mathcal{Q} \to \mathbb{R}$ be differentiable on $\mathcal{S} \subseteq \mathcal{Q}$, and $\eta > 0$ be some small scalar. We say the proximal gradient of $h(q)$ on $\mathcal{S}$ is*

$$G_{h,\mathcal{S}}(q) = \frac{1}{\eta}\Big(q - \mathrm{Proj}_{\mathcal{S}}(q - \eta\nabla h(q))\Big), \quad \forall q \in \mathcal{S}. \tag{3}$$

Definition 3 (Proximal PL)**.** *Let $h : \mathcal{Q} \to \mathbb{R}$ be differentiable on $\mathcal{S} \subseteq \mathcal{Q}$. We say $h(q)$ satisfies proximal μ_h-Polyak-Łojasiewicz (PL) condition on $\mathcal{S}$ if*

$$h(q) - \min_{q\in\mathcal{S}} h(q) \leq \frac{1}{2\mu_h} G^2_{h,\mathcal{S}}(q), \quad \forall q \in \mathcal{S}. \tag{4}$$

Lemma 4 (Lipschitz-Continuity of $S^*_g(x)$ [6, Lemma 6])**.** *Suppose $\mathcal{Y}$ is closed and convex, $g(x, y)$ satisfies PL condition in $y \in \mathcal{Y}$, and $\nabla_y g(x, y)$ is Lipschitz in x. Then there exist a constant $L^g_y \geq 0$ such that for any $x_1, x_2 \in \mathcal{X}$ and $y_1 \in S^*_g(x_1) = \arg\min_{y\in\mathcal{Y}} g(x_1, y)$, there exists $y_2 \in S^*_g(x_1) = \arg\min_{y\in\mathcal{Y}} g(x_2, y)$ such that*

$$\|y_1 - y_2\| \leq L^g_y\|x_1 - x_2\|. \tag{5}$$

Lemma 5 (Lemma 3.1 in [2])**.** *Suppose $\mathcal{Q} \subseteq \mathbb{R}^{d_q}$ is convex, closed, and nonempty. For any $q_1 \in \mathbb{R}^{d_q}$ and any $q_2 \in \mathcal{Q}$, it follows that*

$$\langle \mathrm{Proj}_{\mathcal{Q}}(q_1) - q_2, \mathrm{Proj}_{\mathcal{Q}}(q_1) - q_1\rangle \leq 0. \tag{6}$$

In this way, take $q_1 = q_3 - \eta g$ for any $q_3 \in \mathcal{Q}$, and denote $q_3^{\eta,g} = \mathrm{Proj}_{\mathcal{Q}}(q_3 - \eta g)$ as a projected gradient update in direction g with stepsize η, we have,

$$\langle g, q_3^{\eta,g} - q_2\rangle \leq -\frac{1}{\eta}\langle q_3^{\eta,g} - q_2, q_3^{\eta,g} - q_3\rangle, \quad \forall q_2, q_3 \in \mathcal{Q}. \tag{7}$$

III. Improved Analysis under Smaller Smoothness Constant

A. Proof of Lemma 3

Before proceed to the proof of Lemma 3, we present the following useful lemma.

Lemma 6. *Denote $v(x) = \min_{y\in\mathcal{Y}} g(x, y)$, $S^*_g(x) = \arg\min_{y\in\mathcal{Y}} g(x, y)$. Suppose $\mathcal{Y}$ is closed and convex, $g(x, y)$ is $l_{g,1}$-smooth, and there exists constant L^g_y such that for arbitrary $x_1, x_2 \in \mathcal{X}$, $d_{S^*_g(x_2)}(y_1) \leq L^g_y\|x_1 - x_2\|$ for any $y_1 \in S^*_g(x_1)$. Then for $v(x) = \min_{y\in\mathcal{Y}} g(x, y)$, there is*

$$\nabla v(x) = \nabla_x g(x, y^*_g(x)), \quad \forall y^*_g(x) \in S^*_g(x). \tag{8}$$

Moreover, fix any $y^*_g(x) \in S^*_g(x)$, and for arbitrary $d \in \mathbb{R}^{d_x}$, denote $y_{g,r}$ as some element in $S^*_g(x + rd)$. There is

$$\langle \nabla_y g(x, y^*_g(x)), \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\rangle = 0. \tag{9}$$

Proof. The gradient of value function

$$\nabla v(x) = \nabla_x g(x, y^*_g(x)), \quad \forall y^*_g(x) \in S^*_g(x) \tag{10}$$

is an established outcome, see e.g. [6, Proposition 7]. In this way, denoting $y_{g,r} \in S^*_g(x + rd)$, the directional derivative $v(x)$ in unit direction $d \in \mathbb{R}^{d_x}$ is

$$\begin{aligned} D_d(v(x)) &\overset{(a)}{=} \langle \nabla v(x), d\rangle = \langle \nabla_x g(x, y_{g,0}), d\rangle && (11)\\ &\overset{(b)}{=} \lim_{r\downarrow 0} \frac{1}{r}(g(x + rd, y_{g,r}) - g(x, y_{g,0})) && (12)\\ &\overset{(c)}{=} \langle \nabla_x g(x, y_{g,0}), d\rangle && (13)\\ &\quad + \langle \nabla_y g(x, y_{g,0}), \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\rangle && (14) \end{aligned}$$

where (a) follows the $\nabla v(x) = \nabla_x g(x, y_{g,0})$; (b) is the definition of directional derivative; and (c) applies Taylor expansion. In this way, we can conclude

$$\langle \nabla_y g(x, y_{g,0}), \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\rangle = 0. \tag{15}$$

□

With Lemma 6 prepared, we are ready to explore some features of the directional derivative of $y^*_g(x)$ and show the first part of Lemma 3.

Lemma 7. *Suppose all conditions in Lemma 3 hold. For any unit direction $d \in \mathbb{R}^{d_x}$, there exists an index set $\mathcal{I} \subseteq [d_y]$ such that the directional derivative of $y^*_g(x)$, $D_d(y^*_g(x))$ satisfies*

$$\Big[\lim_{r\downarrow 0} \frac{y^*_g(x + rd) - y^*_g(x)}{r}\Big]_{[d_y]\setminus\mathcal{I}} = 0, \quad \textit{and} \tag{16}$$

$$\Big[\lim_{r\downarrow 0} \frac{y^*_g(x + rd) - y^*_g(x)}{r}\Big]_{\mathcal{I}} = C(x)^{-1}_{[\mathcal{I},\mathcal{I}]} B(x)_{[:,\mathcal{I}]}{}^\top d, \tag{17}$$

and the second-order directional derivative of $v(x)$ is

$$D^2_{dd}(v(x)) = \Big(A(x) - B(x)\begin{bmatrix} C(x)^{-1}_{[\mathcal{I},\mathcal{I}]} B(x)_{[:,\mathcal{I}]}{}^\top \\ 0 \end{bmatrix}\Big)d, \tag{18}$$

where

$$\begin{aligned} A(x) &= \nabla_{xx} g(x, y^*_g(x)),\\ B(x) &= \nabla_{xy} g(x, y^*_g(x)),\\ C(x) &= \nabla_{yy} g(x, y^*_g(x)). \end{aligned}$$

Remark 2. *This means the non-zero part of $D_d(y^*_g(x))$, indexed by $\mathcal{I}$, is determined by the submatrices of $\nabla_{yy} g(x, y^*_g(x))$ and $\nabla_{xy} g(x, y^*_g(x))$ also indexed by $\mathcal{I}$.*

Proof. According to Lemma 2,

$$\nabla v(x) = \nabla_x g(x, y_g^*(x)), \quad \forall y_g^*(x) \in S_g^*(x). \tag{19}$$

Fix any $x \in \mathcal{X}$. For arbitrary $d \in \mathbb{R}^{d_x}$, denote $y_{g,r} = \arg\min_{y \in \mathcal{Y}(x)} g(x+rd, y)$ for some scalar $r \geq 0$. Here, $y_{g,0} = y_g^*(x)$. According to Lemma 6, there is

$$\begin{aligned}
0 =& \lim_{r,r' \downarrow 0} \frac{1}{rr'} \langle \nabla_y g(x + rd, y_{g,r}) - \nabla_y g(x, y_{g,0}), \\
& \qquad y_{g,r'} - y_{g,0} \rangle \\
=& \lim_{r \downarrow 0} \Big\langle \langle \nabla_{xy} g(x, y_{g,0}), d \rangle + \langle \nabla_{yy} g(x, y_{g,0}), \frac{y_{g,r} - y_g}{r} \rangle \\
& + \mathcal{O}(\|rd\|^2), \lim_{r' \downarrow 0} \frac{y_{g,r'} - y_g}{r'} \Big\rangle \\
=& \Big\langle B(x)^\top d + C(x) \Big(\lim_{r \downarrow 0} \frac{y_{g,r} - y_g}{r} \Big), \lim_{r' \downarrow 0} \frac{y_{g,r'} - y_g}{r'} \Big\rangle.
\end{aligned} \tag{20}$$

where the second equality is by Taylor's expansion, and the fact that $\|y_{g,0} - y_{g,r}\| \leq L_y^g \|x - (x + rd)\|$ for some constant $L_y^g > 0$ according to [6, Lemma 5], and the third is from rearranging.

For simplicity, denote the directional derivative of $y_g^*(x)$ as

$$dy = D_d(y_g^*(x)) = \lim_{r \downarrow 0} \frac{y_{g,r} - y_g}{r}. \tag{21}$$

In this way, (20) can be written as

$$0 = dy^\top C(x) dy + dy^\top B(x)^\top d. \tag{22}$$

Denote $\mathcal{I} = \{i : dy_i \neq 0\} \subseteq [d_y]$ as the collection of indices of non-zero elements in dy. Without the loss of generality, take $dy = [dy'; \mathbf{0}]^\top$ where $dy' = dy_{\mathcal{I}}$ is the stack of non-zero elements in dy. In this way, writing A and B in block matrix form, (22) is equivalent to

$$\begin{aligned}
0 =& \begin{bmatrix} dy' & 0 \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} dy' \\ 0 \end{bmatrix} + \begin{bmatrix} dy' & 0 \end{bmatrix} \begin{bmatrix} B_1^\top d \\ B_2^\top d \end{bmatrix} \\
=& dy'^\top C_{11} dy' + dy'^\top B_1^\top d.
\end{aligned} \tag{23}$$

where $C_{11} = C(x)_{[\mathcal{I},\mathcal{I}]}$, $B_1 = B(x)_{[:,\mathcal{I}]}$.

As g is strongly convex in y, its hessian with respect to y, $C(x)$ is positive definite. Therefore, any of its principal sub-matrix C_{11} is positive definite and full rank. In this way, Solving (23) therefore gives

$$dy' = C_{11}^{-1} B_1^\top d. \tag{24}$$

This proves (62).

As $v(x)$ is differentiable, the first order directional derivative for $v(x)$ is

$$D_d(v(x)) = \nabla v(x)^\top d = \nabla_x g(x, y_g^*(x))^\top d \tag{25}$$

In this way, the second order directional derivative for $v(x)$ is

$$\begin{aligned}
& D_{dd}^2(v(x)) = \lim_{r \downarrow 0} \frac{1}{r} \Big(\nabla D_d(v(x + rd)) - D_d(v(x)) \Big) \\
=& \lim_{r \downarrow 0} \frac{1}{r} d^\top \Big(\nabla_x g\Big(x + rd, y_{g,r}\Big) - \nabla_x g\Big(x, y_{g,0}\Big) \Big) \\
=& d^\top \nabla_{xx} g(x, y_{g,0}) d + d^\top \nabla_{xy} g(x, y_{g,0}) \lim_{r \downarrow 0} \frac{1}{r} (y_{g,r} - y_{g,0}) \\
=& d^\top A(x) d - d^\top B(x) \begin{bmatrix} dy' \\ 0 \end{bmatrix} \\
=& d^\top A(x) d - d^\top B(x) \begin{bmatrix} C(x)_{11}^{-1} B(x)_1{}^\top d \\ 0 \end{bmatrix} \\
=& d^\top \Big(A(x) - B(x) \begin{bmatrix} C(x)_{[\mathcal{I},\mathcal{I}]}^{-1} B(x)_{[:,\mathcal{I}]}{}^\top \\ 0 \end{bmatrix} \Big) d
\end{aligned} \tag{26}$$

where the first equality is the definition of second-order directional derivative; the second is by plugging the first-order directional derivative; the third follows Taylor's expansion; and the following is by plugging $\lim_{r \downarrow 0} \frac{1}{r}(y_{g,r} - y_{g,0})$. This completes the proof. □

In this way, we finished proving the first part of Lemma 3. For the second part, we look into the relation between $\mathcal{I}$, the non-zero index of $D_d(y_g^*(x))$, and $\mathcal{I}_\gamma$, the non-zero index of the counterpart $D_d(y_\gamma^*(x))$.

Lemma 8. *Suppose $\lim_{\gamma \to \infty} y_\gamma^*(x) = y_g^*(x)$ for all $x \in \mathcal{X}$. Fix any unit direction $d \in \mathbb{R}^{d_x}$, denote $\mathcal{I} \subseteq [d_y]$ as the index set for non-zero elements in $D_d(y_g^*(x)) = \lim_{r \downarrow 0} \frac{y_g^*(x+rd) - y_g^*(x)}{r}$, and $\mathcal{I}_\gamma \subseteq [d_y]$ as the index set for non-zero elements in $D_d(y_\gamma^*(x)) = \lim_{r \downarrow 0} \frac{y_\gamma^*(x+rd) - y_\gamma^*(x)}{r}$. Then, for any $\delta > 0$, there exists a finite γ^* such that for all $\gamma > 0$,*

1) *$\mathcal{I} \subseteq \mathcal{I}_\gamma$, and*
2) *$[D_d(y_\gamma^*(x))]_i < \delta$ for all $i \in \mathcal{I}_\gamma \backslash \mathcal{I}$.*

Proof. Similarly denote $y_{\gamma,r} = \arg\min_{y \in \mathcal{Y}(x)} \gamma^{-1} f(x, y) + g(x, y)$ for scalar $r \geq 0$, the directional derivative of $y_\gamma^*(x)$ as

$$dy_\gamma = D_d(y_\gamma^*(x)) = \lim_{r \downarrow 0} \frac{y_{\gamma,r} - y_{\gamma,0}}{r}, \tag{27}$$

and $dy'_\gamma = [dy_\gamma]_{\mathcal{I}_\gamma}$ where $\mathcal{I}_\gamma = \{i : dy_{\gamma_i} \neq 0\} \subseteq [d_y]$ is the collection of indices of non-zero elements in dy_γ.

For any $i \in \mathcal{I}$, $dy_i \neq 0$. i.e.

$$\lim_{r \downarrow 0} \Big| \frac{[y_{g,r}]_i - [y_{g,0}]_i}{r} \Big| = |dy_i| > 0 \tag{28}$$

Moreover, according to Lemma 1, $y_{\gamma,r} \to y_g$ for $\gamma \to \infty$. Thus, for any $r > 0$,

$$\lim_{\gamma \to \infty} \frac{y_{\gamma,r} - y_{\gamma,0}}{r} = \frac{y_{g,r} - y_{g,0}}{r}. \tag{29}$$

i.e. for any $\delta > 0$, there exist $\gamma^* > 0$ such that for all $\gamma > \gamma^*$,

$$\Big| \frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r} - \frac{[y_{g,r}]_i - [y_{g,0}]_i}{r} \Big| < \delta \tag{30}$$

Fix $\delta = \min_{i\in\mathcal{I}} |dy_i|/2$, following triangle inequality, there is

$$\begin{aligned}
&\lim_{r\downarrow 0}\Big|\frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r}\Big| \\
\geq &\lim_{r\downarrow 0}\Big|\frac{[y_{g,r}]_i - [y_{g,0}]_i}{r}\Big| \\
&- \lim_{r\downarrow 0}\Big|\frac{[y_{g,r}]_i - [y_{g,0}]_i}{r} - \frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r}\Big| \\
\geq &|dy_i| - \min_{i\in\mathcal{I}} |dy_i|/2 > 0. \qquad (31)
\end{aligned}$$

i.e. $\frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r} \neq 0$. So $\mathcal{I} \subseteq \mathcal{I}_\gamma$ for all $\gamma > \gamma_1^*$. This completes the first part of the lemma.

For the second part, $i \in [d_y]\backslash\mathcal{I}$, similarly use triangle inequality, there is

$$\begin{aligned}
&\lim_{r\downarrow 0}\Big|\frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r}\Big| \\
\leq &\lim_{r\downarrow 0}\Big|\frac{[y_{g,r}]_i - [y_{g,0}]_i}{r}\Big| \\
&+ \lim_{r\downarrow 0}\Big|\frac{[y_{g,r}]_i - [y_{g,0}]_i}{r} - \frac{[y_{\gamma,r}]_i - [y_{\gamma,0}]_i}{r}\Big| \\
< &0 + \delta = \delta. \qquad (32)
\end{aligned}$$

for all $\gamma > \gamma^*$. In this way, $dy_{\gamma_i} < \delta$ for all $i \in \mathcal{I}_\gamma\backslash\mathcal{I}$. □

Now we are ready to prove the second part of Lemma 3.

Lemma 9. *Suppose all conditions in Lemma 7 hold. Fix any unit direction $d \in \mathbb{R}^{d_x}$, denote $\mathcal{I} \subseteq [d_y]$ as the index set for non-zero elements in $D_d(y_g^*(x))$. For any $\delta > 0$, there exists a finite $\gamma^* > 0$ such that the second-order directional derivative of $v_\gamma(x)$ is*

$$\begin{aligned}
&D_{dd}^2(v_\gamma(x)) \\
&= \Big(A_\gamma(x) - B_\gamma(x)\begin{bmatrix} C_\gamma(x)_{[\mathcal{I},\mathcal{I}]}^{-1} B_\gamma(x)_{[:,\mathcal{I}]}^\top \\ 0 \end{bmatrix}\Big)d + \mathcal{O}(\delta) \\
&\textit{where} \qquad (33) \\
&A_\gamma(x) = \nabla_{xx}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)), \\
&B_\gamma(x) = \nabla_{xy}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)), \\
&C_\gamma(x) = \nabla_{yy}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)).
\end{aligned}$$

Proof. Denote $y_{\gamma,r} = \arg\min_{y\in\mathcal{Y}(x)} \gamma^{-1}f(x,y) + g(x,y)$ for scalar $r \geq 0$, the directional derivative of $y_\gamma^*(x)$ as

$$dy_\gamma = D_d(y_\gamma^*(x)) = \lim_{r\downarrow 0}\frac{y_{\gamma,r} - y_{\gamma,0}}{r}, \qquad (34)$$

and $dy'_\gamma = [dy_\gamma]_{\mathcal{I}_\gamma}$ where $\mathcal{I}_\gamma = \{i : dy_{\gamma_i} \neq 0\} \subseteq [d_y]$ is the collection of indices of non-zero elements in dy_γ.

Without loss of generality, partition $dy'_\gamma = [[dy_\gamma]_{\mathcal{I}_1}, [dy_\gamma]_{\mathcal{I}_2}]$ where $\mathcal{I}_1 = \mathcal{I}$ and $\mathcal{I}_2 = \mathcal{I}_\gamma\backslash\mathcal{I}$ as $\mathcal{I} \subseteq \mathcal{I}_\gamma$ according to Lemma 8. Following Lemma 7, there is

$$\begin{aligned}
dy'_\gamma =&C_\gamma(x)_{[\mathcal{I}_\gamma,\mathcal{I}_\gamma]}^{-1} B_\gamma(x)_{[:,\mathcal{I}_\gamma]}^\top d \\
=&\begin{bmatrix} C_\gamma(x)_{[\mathcal{I}_1,\mathcal{I}_1]} & C_\gamma(x)_{[\mathcal{I}_1,\mathcal{I}_2]} \\ C_\gamma(x)_{[\mathcal{I}_2,\mathcal{I}_1]} & C_\gamma(x)_{[\mathcal{I}_2,\mathcal{I}_2]} \end{bmatrix}^{-1} \begin{bmatrix} B_\gamma(x)_{[:,\mathcal{I}_1]}^\top d \\ B_\gamma(x)_{[:,\mathcal{I}_2]}^\top d \end{bmatrix}.
\end{aligned}$$

Denote $C_{11} = C_\gamma(x)_{[\mathcal{I}_1,\mathcal{I}_1]}$, $C_{12} = C_\gamma(x)_{[\mathcal{I}_1,\mathcal{I}_2]}$, $C_{21}C_\gamma(x)_{[\mathcal{I}_2,\mathcal{I}_1]}$, and $C_{22}C_\gamma(x)_{[\mathcal{I}_2,\mathcal{I}_2]}$, and S is the Schur complement of C_{11} in C, given by $S = C_{22} - C_{21}C_{11}^{-1}C_{12}$. There is

$$\begin{aligned}
\begin{bmatrix} dy_{\mathcal{I}_1} \\ dy_{\mathcal{I}_2} \end{bmatrix} = &\begin{bmatrix} C_{11}^{-1} + C_{11}^{-1}C_{12}S^{-1}C_{21}C_{11}^{-1} & -C_{11}^{-1}C_{12}S^{-1} \\ -S^{-1}C_{21}C_{11}^{-1} & S^{-1} \end{bmatrix} \\
&\times \begin{bmatrix} B_\gamma(x)_{[:,\mathcal{I}_1]}^\top d \\ B_\gamma(x)_{[:,\mathcal{I}_2]}^\top d \end{bmatrix}.
\end{aligned}$$

In this way, there is

$$\begin{aligned}
&dy_{\mathcal{I}_2} = -S^{-1}C_{21}C_{11}^{-1}B_\gamma(x)_{[:,\mathcal{I}_1]}^\top d + S^{-1}B_\gamma(x)_{[:,\mathcal{I}_2]}^\top d, \\
&\text{and } dy_{\mathcal{I}_1} = C_{11}^{-1}B_\gamma(x)_{[:,\mathcal{I}_1]}^\top d - C_{11}^{-1}C_{12}dy_{\mathcal{I}_2} \qquad (35)
\end{aligned}$$

From Lemma 8, we know $v_{\mathcal{I}_2} = \mathcal{O}(\delta)$ for all $\gamma > \gamma^*$. Additionally, as $C_\gamma(x)$ is the hessian of a strongly convex function, its principle submatrix C_{11}, is also of full rank with its smallest eigenvalue $\lambda_-(C_{11}) \geq \mu_g$ [3, Lemma F.6] and thus $\|C_1 1^{-1}\| \leq \frac{1}{\mu}$. Moreover, as $C_\gamma(x)$ is also the hessian of a smooth function, so $\|C_\gamma(x)\|$ is bounded so as its submatrix $\|C_{12}\|$. In this way, we know $\|C_{11}^{-1}C_{12}dy_{\mathcal{I}_2}\| = \mathcal{O}(\delta)$ and we can conclude that

$$dy_{\mathcal{I}_1} = C_{11}^{-1}B_\gamma(x)_{[:,\mathcal{I}_1]}^\top d + \mathcal{O}(\delta). \qquad (36)$$

In this way, we can similarly complete the proof as in (26). □

In this way, Lemma 3 follows directly from Lemma 7 and Lemma 15.

B. Proof of Theorem 1

Fix arbitrary $x \in \mathcal{X}$. For notation simplicity, we follow a similar notation rule as in Lemma 3, where

$$\begin{aligned}
A =&\nabla_{xx}g(x, y_g^*(x)), \\
B =&\nabla_{xy}g(x, y_g^*(x)), \\
C =&\nabla_{yy}g(x, y_g^*(x)), \\
E_A =&A - \nabla_{xx}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)), \\
E_B =&B - \nabla_{xy}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)), \\
E_C =&C - \nabla_{yy}\gamma^{-1}f(x, y_\gamma(x)) + g(x, y_\gamma^*(x)).
\end{aligned}$$

Here, $\|A\|, \|B\|, \|C\| \leq l_{g,1}$ by the $l_{g,1}$-Lipschitzness of ∇g. Moreover,

$$\begin{aligned}
\|E_A\| =&\|\gamma^{-1}\nabla_{xx}f(x, y_\gamma^*(x)) \\
&+ \nabla_{xx}g(x, y_\gamma^*(x)) - \nabla_{xx}g(x, y_g^*(x))\| \\
\leq&\gamma^{-1}\|\nabla_{xx}f(x, y_\gamma^*(x))\| \\
&+ \|\nabla_{xx}g(x, y_\gamma^*(x)) - \nabla_{xx}g(x, y_g^*(x))\| \\
\leq&\gamma^{-1}l_{f,1} + l_{g,2}\|y_\gamma^*(x) - y_g^*(x)\| \\
\leq&\frac{l_{f,1}}{\gamma} + l_{g,2}\frac{l_{f,0}}{\mu_g\gamma}. \qquad (37)
\end{aligned}$$

The first inequality applies triangle inequality; the second uses the $l_{f,1}$-smoothness of f and the $l_{g,2}$-Lipschitz of $\nabla^2 g$; and the last follows Lemma 1. Following similar analysis,

$$\begin{aligned}\|E_B\| =&\|\gamma^{-1}\nabla_{xy}f(x,y_\gamma^*(x))+\nabla_{xy}g(x,y_\gamma^*(x))\\ &-\nabla_{xy}g(x,y_g^*(x))\|\\ \le&\frac{l_{f,1}}{\gamma}+l_{g,2}\frac{l_{f,0}}{\mu_g\gamma},\end{aligned}\tag{38}$$

$$\begin{aligned}\text{and}\quad \|E_C\| =&\|\gamma^{-1}\nabla_{yy}f(x,y_\gamma^*(x))+\nabla_{yy}g(x,y_\gamma^*(x))\\ &-\nabla_{yy}g(x,y_g^*(x))\|\\ \le&\frac{l_{f,1}}{\gamma}+l_{g,2}\frac{l_{f,0}}{\mu_g\gamma}.\end{aligned}\tag{39}$$

Additionally, the μ_g-Strongly convexity condition guarantees the smallest eigenvalue of C, $\lambda_{\min}^+(C)\ge\mu_g$ and similarly $\lambda_{\min}^+(C+E_C)\ge\mu_g$ for $\gamma\ge\frac{l_{g,1}}{l_{f,1}}$, according to [3, Lemma F.6]. The same holds for the principle submatrices $C_{[\mathcal{I},\mathcal{I}]}$ and $(C+E_C)_{[\mathcal{I},\mathcal{I}]}$ as well. In this way, following [8], we have

$$\|E_{C^{-1}}\|=\|C_{[\mathcal{I},\mathcal{I}]}^{-1}-(C+E_C)_{[\mathcal{I},\mathcal{I}]}^{-1}\|\le\frac{1+\sqrt{5}}{2\mu_g}\|E_C\|.$$

As the penalty function $F_\gamma(x)=\gamma(v^h(x,y)-v(x))$, for any unit directional d, its second order directional derivative is

$$D_{dd}^2(F_\gamma(x))=\gamma\Big(D_{dd}^2(v_\gamma(x))-D_{dd}^2(v(x))\Big)\tag{40}$$

where

$$\begin{aligned}D_{dd}^2(v(x)) =&d^\top\Big(A-B\begin{bmatrix}C_{[\mathcal{I},\mathcal{I}]}^{-1}B_{[:,\mathcal{I}]}^\top\\0\end{bmatrix}\Big)d,\quad\text{and}\\ D_{dd}^2(v_\gamma(x)) =&d^\top(A+E_A)d+\mathcal{O}(\delta)\\ -&d^\top(B+E_B)\begin{bmatrix}(C+E_C)_{[\mathcal{I},\mathcal{I}]}^{-1}(B+E_B)_{[:,\mathcal{I}]}^\top\\0\end{bmatrix}d.\end{aligned}\tag{41}$$

In this way,

$$\begin{aligned}&\|D_{dd}^2(F_\gamma(x))\|=\mathcal{O}(\delta)+\gamma\Big\|D_{dd}^2(v_\gamma(x))-D_{dd}^2(v(x))\Big\|\\ \le&\gamma\|E_A\|+\gamma\Big(l_{g,1}^2\|E_{C^{-1}}\|+2\frac{l_{g,1}}{\mu_g}\|E_B\|\\ &+(2l_{g,1}+\frac{1}{\mu_g})\|E_B\|\|E_{C^{-1}}\|+\|E_B\|^2\|E_{C^{-1}}\|\Big)\\ \le&\gamma\Big(1+l_{g,1}^2\frac{1+\sqrt{5}}{2\mu_g}+2\frac{l_{g,1}}{\mu_g}\Big)\Big(\gamma^{-1}l_{f,1}+l_{g,2}\|y_\gamma^*(x)-y_g^*(x)\|\Big)\\ &+\gamma\Big(2l_{g,1}+\mu_g^{-1}\Big)\frac{1+\sqrt{5}}{2\mu_g}\Big(\gamma^{-1}l_{f,1}+l_{g,2}\|y_\gamma^*(x)-y_g^*(x)\|\Big)^2\\ &+\gamma\frac{1+\sqrt{5}}{2\mu_g}\Big(\gamma^{-1}l_{f,1}+l_{g,2}\|y_\gamma^*(x)-y_g^*(x)\|\Big)^3+\mathcal{O}(\delta)\\ \le&C_1C_0+\frac{1}{\gamma}C_2C_0^2+\frac{1}{\gamma^2}C_3C_0^3+\mathcal{O}(\delta),\end{aligned}$$

$$\text{where}\quad\begin{cases}C_0=&l_{f,1}+l_{g,2}l_{f,0}\mu_g^{-1}\\ C_1=&1+l_{g,1}^2\frac{1+\sqrt{5}}{2\mu_g}+2\frac{l_{g,1}}{\mu_g}\\ C_2=&\Big(2l_{g,1}+\mu_g^{-1}\Big)\frac{1+\sqrt{5}}{2\mu_g}\\ C_3=&\frac{1+\sqrt{5}}{2\mu_g}\end{cases}$$

Here, the first inequality is from triangle inequality, Cauchy-Schwartz inequality and $\|B\|\le l_{g,1},\|C_\dagger\|\le\frac{1}{\mu_g}$; the second is are by the upper bounds of $\|E_A\|$, $\|E_B\|$, $\|E_C\|$, and $\|E_{C^{-1}}\|$, and the last is by the upper bound for $\|y_g^*(x)-y_\gamma^*(x)\|$ in Lemma 1.

C. Proof of Proposition 2

According to [5], to achieve

$$\|y_t-y_g^*(x_t)\|^2,\|y_t-y_g^*(x_t)\|^2<\epsilon^2,\tag{42}$$

we can apply PGD on

$$\min_{y\in\{y\in\mathcal{Y}:c(y)\le0\}}g(x,y),\quad\text{and}\tag{43}$$

$$\min_{y\in\{y\in\mathcal{Y}:c(y)\le0\}}\gamma^{-1}f(x,y)+g(x,y)\tag{44}$$

with algorithm complexity $\mathcal{O}(\ln(\epsilon^{-1}))$. In this way, the update bias

$$\begin{aligned}&\|b(x_t)\|^2=\|\nabla F_\gamma(x_t)-g_t\|^2\\ \le&\gamma\|\gamma^{-1}\nabla_x f(x_t,y_\gamma^*(x_t))+\nabla_x g(x_t,y_\gamma^*(x_t))-\nabla_x g(x_t,y_g^*(x_t))\\ &-\big(\gamma^{-1}\nabla_x f(x_t,y_t^\gamma)+\nabla_x g(x_t,y_t^\gamma)-\nabla_x g(x_t,y_t^g)\big)\|^2\\ \le&2\|\nabla_x f(x_t,y_\gamma^*(x_t))+\gamma\nabla_x g(x_t,y_\gamma^*(x_t))\\ &-\big(\nabla_x f(x_t,y_t^\gamma)+\gamma\nabla_x g(x_t,y_t^\gamma)\big)\|^2\\ &+2\|\gamma\nabla_x g(x_t,y_\gamma^*(x_t))-\gamma\nabla_x g(x_t,y_t^g)\|^2\\ \le&2(l_{f,1}+\gamma l_{g,1})^2\|y_\gamma^*(x_t)-y_t^\gamma\|^2+2\gamma^2l_{g,1}^2\|y_g^*(x)-t_t^g\|^2\\ \le&\mathcal{O}(\gamma^2\epsilon^2)=\mathcal{O}(\epsilon).\end{aligned}\tag{45}$$

where the second inequality is by Young's inequality; the third is from the smoothness of f and g; and the last is from (42) and $\gamma=\mathcal{O}(\epsilon^{-0.5})$.

According to smoothness of $F_\gamma(x)$, there is

$$\begin{aligned}F_\gamma(x_{t+1})\le&F_\gamma(x_t)+\langle\nabla F_\gamma(x_t),x_{t+1}-x_t\rangle+\frac{l_{F,1}}{2}\|x_{t+1}-x_t\|^2\\ \le&F_\gamma(x_t)+\langle g_t,x_{t+1}-x_t\rangle+\frac{1}{2\eta}\|x_{t+1}-x_t\|^2\\ &+\langle b(x_t),x_{t+1}-x_t\rangle,\end{aligned}\tag{46}$$

where the second inequality is by $\eta\le\frac{1}{l_{F,1}}$. The projection guarantees that x_{t+1} and x_t are in $\mathcal{X}$. In this way, Following lemma 5, we know that

$$\langle g_t,x_{t+1}-x_t\rangle\le-\frac{1}{\eta}\|x_{t+1}-x_t\|^2.$$

Plugging this back to (46),

$$\begin{aligned}F_\gamma(x_{t+1})\le&F_\gamma(x_t)-\frac{1}{2\eta}\|x_{t+1}-x_t\|^2+\langle b(x_t),x_{t+1}-x_t\rangle\\ \le&F_\gamma(x_t)-\frac{1}{2\eta}\|x_{t+1}-x_t\|^2+\eta\|b(x_t)\|^2\\ &+\frac{1}{4\eta}\|x_{t+1}-x_t\|^2\\ =&F_\gamma(x_t)-\frac{1}{4\eta}\|x_{t+1}-x_t\|^2+\eta\|b(x_t)\|^2,\end{aligned}$$

where the second inequality is from Young's inequality. Telescoping therefore gives

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\Big\|\frac{x_t-\mathrm{Proj}_{\mathcal{X}}(x_t-\eta g_t)}{\eta}\Big\|^2\\ \le&\frac{4}{\eta T}(F_\gamma(x_0)-F_\gamma(x_t))+\frac{4}{T}\sum_{t=0}^{T-1}\|b(x_t)\|^2. \end{aligned}\tag{47}$$

In this way

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\Big\|G_\eta(x_t)\Big\|^2\\ =&\frac{1}{T}\sum_{t=0}^{T-1}\Big\|\frac{x_t-\mathrm{Proj}_{\mathcal{X}}(x_t-\eta\nabla F_\gamma(x_t))}{\eta}\Big\|^2\\ =&\frac{1}{T}\sum_{t=0}^{T-1}\Big\|\frac{x_t-\mathrm{Proj}_{\mathcal{X}}(x_t-\eta g_t)}{\eta}\Big\|^2+\frac{1}{T}\sum_{t=0}^{T-1}\|b(x_t)\|^2\\ \le&\frac{4}{\eta T}(F_\gamma(x_0)-F_\gamma(x_t))+\frac{5}{T}\sum_{t=0}^{T-1}\|b(x_t)\|^2 \qquad (48)\\ =&\mathcal{O}(\eta^{-1}T^{-1}+\epsilon) \qquad (49)\end{aligned}$$

As $\eta=O(1)$, to achieve $\frac{1}{T}\sum_{t=0}^{T-1}\|G_\eta(x_t)\|^2\le\epsilon$, we require complexity

$$T=\mathcal{O}(\epsilon^{-1}).\tag{50}$$

In this way, counting in the inner $\mathcal{O}(\ln(\epsilon^{-1}))$ complexity, the total algorithm complexity is

$$\mathcal{O}(\epsilon^{-1}\ln(\epsilon^{-1}))=\tilde{\mathcal{O}}(\epsilon^{-1}).\tag{51}$$

IV. Extension to Coupled Constraint setting

A. Extension to Coupled Constraint setting

Lemma 10 (Generalized Lemma 3)**.** *Consider $\mathcal{Y}(x)=\{y\in\mathcal{Y}:c(x,y)\le 0\}$. Suppose Assumption 1, 2, 3, 4, and 5 hold. For any unit direction $d\in\mathbb{R}^{d_x}$, there exists an index set $\mathcal{I}\subseteq[d_y]$ such that the second-order directional derivative of $v(x)$,*

$$\begin{aligned}&D^2_{dd}(v(x))\\ =&d^\top B'(x)_{[:,\mathcal{I}]}C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1\top}C(x)_{[\mathcal{I},\mathcal{I}]}C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^\top d\\ &+d^\top\nabla_{xx}A(x)d\\ &-d^\top B'(x)_{[:,\mathcal{I}]}C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B(x)_{[\mathcal{I},:]}d\\ &-d^\top B(x)_{[:,\mathcal{I}]}C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^\top d.\end{aligned}\tag{52}$$

where $A(x)$, $B(x)$, and $C(x)$ are defined in (12)*, and*

$$\begin{aligned}B'(x)&=\nabla_{xy}L_g(x,y_g^*(x),\lambda_g^*(x)),\\ C'(x)&=\nabla_{yy}L_g(x,y_g^*(x),\lambda_g^*(x)).\end{aligned}$$

Moreover, for any $\delta>0$, there exists a finite γ^* such that

$$\begin{aligned}&D^2_{dd}(v_\gamma(x))\\ =&d^\top B'_\gamma(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}C_\gamma(x)_{[\mathcal{I},\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'_\gamma(x)_{[:,\mathcal{I}]}{}^\top d\\ &+d^\top\nabla_{xx}A_\gamma(x)d\\ &-d^\top B'_\gamma(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B_\gamma(x)_{[\mathcal{I},:]}d\\ &-d^\top B(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'_\gamma(x)_{[:,\mathcal{I}]}{}^\top d.\end{aligned}\tag{53}$$

for all $\gamma>\gamma^*$, for the same $\mathcal{I}$, where

$$\begin{aligned}B'_\gamma(x)&=\nabla_{xy}L_\gamma(x,y_\gamma^*(x),\lambda_g^*(x)),\\ C'_\gamma(x)&=\nabla_{yy}L_\gamma(x,y_\gamma^*(x),\lambda_g^*(x)).\end{aligned}$$

Before we proceed to the proof of Lemma 10, we present the following as an extension to Lemma 6 for BLO with coupled constraint setting.

Lemma 11. *Suppose $\mathcal{Y}\subseteq\mathbb{R}^{d_x}$ is a closed and convex set with smooth boundary, $\nabla_y g(x,y)$ is continuous in y, and Lipschitz in x, and $g(x,y)$ satisfies μ_g-Proximal PL in y on $\mathcal{Y}$ and on a disturbed domain $\mathcal{Y}_\delta$ such that (i) $\mathcal{Y}\subset\mathcal{Y}_\delta$, and (ii) $d_{\mathcal{Y}}(y)=\delta$ for any $y\in bd(\mathcal{Y}_\delta)$. Fix any $y_g^*(x)\in S_g^*(x)$, and for arbitrary unit direction $d\in\mathbb{R}^{d_x}$, denote $y_{g,r}$ as some element in $S_g^*(x+rd)$. Then,*

$$\langle\nabla_y g(x,y_g^*(x)),\lim_{r\downarrow 0}\frac{y_{g,r}-y_g}{r}\rangle=0\tag{54}$$

Proof. When $\nabla_y g(x,y_g^*(x))=0$, the equality automatically hold. We consider in the following the case $\nabla_y g(x,y_g^*(x))\neq 0$.

When $\nabla_y g(x,y_g^*(x))\neq 0$, y_g is on the boundary of $\mathcal{Y}$, i.e. $y_g^*(x)\in\mathrm{bd}(\mathcal{Y})$, as the (local) optimum of a differentiable function happens either at stationary point or at extreme (i.e. boundary).

Additionally, $\{y^*\in\mathbb{R}^{d_y}:\nabla_y g(x,y^*)=0\}$ is a closed set, as $\{0\}$ is closed and the closeness is preserved by continuous image mapping, $\nabla_y g(x,\cdot)$. In this way, there exists $\delta_0>0$ such that $\|y_g^*(x)-y^*\|\ge\delta_0$ for any $y^*\in\{y^*\notin\mathcal{Y}:\nabla_y g(x,y^*)=0\}\subseteq\{y^*\in\mathbb{R}^{d_y}:\nabla_y g(x,y^*)=0\}$.

Denote some $y_g^\delta\in\min_{y\in\mathcal{Y}_\delta}g(x,y)$. According to [1, Theorem 2.87], $\|y_g^\delta-y_g^*(x)\|=\mathcal{O}(\delta)$. In this way, there exists some $\delta_1>0$ of the order δ_0 such that for all $0<\delta<\delta_1$, $\nabla_y g(x,y_g^\delta)\neq 0$, indicating that y_g^δ is on the boundary of $\mathcal{Y}_\delta$, i.e. $y_g^\delta\in\mathrm{bd}(\mathcal{Y}_\delta)$.

Moreover, under the condition that Proximal PL condition for $g(x,y)$ also holds for $y\in\mathcal{Y}_\delta$, there exist $y_{g,r}^\delta\in\arg\min_{y\in\mathcal{Y}_\delta}g(x+rd,y)$ such that $\|y_g^\delta-y_{g,r}^\delta\|\le L_y^g r$ according to Lemma 4. In this way, there exists some $r_0>0$ such that for all $0<r<r_0$, $\|y_g^\delta-y_{g,r}^\delta\|<\delta$, i.e. $y_{g,r}^\delta\in\mathcal{Y}_\delta\backslash\mathcal{Y}$ as y_g^δ is on the boundary of $\mathcal{Y}_\delta$, which is in δ distance to $\mathcal{Y}$. This directly implies $\nabla_y g(x,y_{g,r})\neq 0$ as otherwise $y_{g,r}^\delta=y_{g,r}\in\mathcal{Y}$. This gives $y_{g,r}\in\mathrm{bd}(\mathcal{Y})$, for all $0<r<r_0$.

In this way, as y_g, $y_{g,r}$ are both on $\mathrm{bd}(\mathcal{Y})$ and $y_{g,r}$ approaches to y_g, we know the limiting direction of y_g to $y_{g,r}$ is in the tangent cone of $\mathcal{Y}$ on boundary point y_g:

$$\lim_{\|y_{g,r}-y_g^*(x)\|\downarrow 0}\frac{y_{g,r}-y_g}{\|y_{g,r}-y_g^*(x)\|}\in T_{\mathcal{Y}}(y_g^*(x)).\tag{55}$$

Moreover, as $\|y_{g,r} - y_g^*(x)\| \le L_y^g r$, so $\lim_{r\downarrow 0} \frac{y_{g,r}-y_g}{r}$ either equals 0, which directly leads to the equality proved, or

$$\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r} \in T_{\mathcal{Y}}(y_g^*(x)). \tag{56}$$

In this way, as $\mathcal{Y}$ is smooth on the boundary, $\nabla_y g(x, y_g^*(x)) \perp T_{\mathcal{Y}}(y_g^*(x))$ and we can conclude the lemma. □

Lemma 12. *Suppose $\mathcal{Y}$ is smooth on its boundary, $\|\lambda_g^*(x)\| < B_\lambda$ for all $x \in \mathcal{X}$, and Assumption 2, 3 hold. Fix any $x \in \mathcal{X}$, denote $y_g^*(x) = \arg\min_{\mathcal{Y}} g^\lambda(x, y)$, and for arbitrary unit direction $d \in \mathbb{R}^{d_x}$, denote $y_{g,r} = \arg\min_{\mathcal{Y}} g^\lambda(x + rd, y)$. For $\lambda_g^*(x) = \arg\max_{\lambda \in \mathbb{R}_+^{d_x}} (\min_{y\in\mathcal{Y}} g(x, y) + \langle \lambda, c(x, y)\rangle$ being the unique Lagrangian multiplier [7], then,*

$$\Big\langle \nabla_y g(x, y_g^*(x)) + \langle \lambda_g, \nabla_x c(x, y_g^*(x))\rangle, \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r} \Big\rangle = 0 \tag{57}$$

Proof. The proof follows directly from Lemma 11. Denote $g^\lambda(x, y) = g(x, y) + \langle \lambda_g^*(x), c(x, y)\rangle$. Here we can easily conclude that $g^\lambda(x, y)$ is proximal-PL in $y \in \mathcal{Y}$ and any disturbed domain $\mathcal{Y}_\delta$ defined in Lemma 11, as $g(x, y)$ is strongly convex in y and $c(x, y)$ is convex in y, according to [5]. Moreover, as $\nabla_y g^\lambda(x, y) = \nabla_y g(x, y) + \langle \lambda_g^*(x), \nabla_y c(x, y)\rangle$, and as assuming $\|\lambda_g^*(x)\| < B_\lambda$ for all $x \in \mathcal{X}$, there is, for any $x_1, x_2 \in \mathcal{X}$,

$$\begin{aligned}&\|\nabla_y g^\lambda(x_1, y) - \nabla_y g^\lambda(x_2, y)\|\\ \le&(l_{g,1} + l_{c,1}\|\lambda_g^*(x)\|)\|x_1 - x_2\|\\ \le&(l_{g,1} + l_{c,1}B_\lambda)\|x_1 - x_2\|.\end{aligned}$$

i.e. $\nabla_y g^\lambda(x, y)$ is Lipschitz in x. So all conditions in Lemma 11 are checked and therefore proved.

□

Lemma 13. *Under all assumptions and notations in Lemma 12, and suppose $g(x, y)$, $c(x, y)$ is locally Lipschitz at $(x, y_g^*(x))$, there is*

$$\Big\langle \lim_{r\downarrow 0} \frac{\lambda_g^*(x + rd) - \lambda_g^*(x)}{r}, c(x, y_g^*(x)) \Big\rangle = 0. \tag{58}$$

Proof. The directional derivative of $v(x)$ in unit direction $d \in \mathbb{R}^{d_x}$ is

$$\begin{aligned}D_d(v) =& \lim_{r\downarrow 0} \frac{1}{r}(c(x + rd, y_{g,r}) + \langle \lambda_g^*(x + rd), c(x + rd, y_{g,r})\rangle\\ &- c(x + rd, y_{g,r}) + \langle \lambda_g^*(x + rd), c(x + rd, y_{g,r})\rangle)\\ =& \lim_{r\downarrow 0} \frac{1}{r}\big(c(x + rd, y_{g,r}) + \langle \lambda_g^*(x), c(x + rd, y_{g,r})\rangle\\ &+ \langle \lambda_g^*(x + rd) - \lambda_g^*(x), c(x + rd, y_{g,r})\rangle\\ &- c(x + rd, y_{g,r}) + \langle \lambda_g^*(x + rd), c(x + rd, y_{g,r})\rangle\big)\\ =&\langle \nabla_x g(x, y_g^*(x)) + \langle \lambda_g^*(x), \nabla_x c(x, y_g^*(x))\rangle, d\rangle\\ &+ \langle \nabla_y g(x, y_g^*(x)) + \langle \lambda_g^*(x), \nabla_y c(x, y_g^*(x))\rangle, \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\rangle\\ &+ \langle \lim_{r\downarrow 0} \frac{\lambda_g^*(x + rd) - \lambda_g^*(x)}{r}, c(x, y_g^*(x))\rangle\end{aligned}$$

where the last equality is by Taylor expansion and y_g being Lipschitz according to Lemma 4. By Lemma 12, there is $\langle \nabla_y g(x, y_g^*(x)) + \langle \lambda_g^*(x), \nabla_y c(x, y_g^*(x))\rangle, \lim_{r\downarrow 0} \frac{y_{g,r}-y_g}{r}\rangle = 0$. Moreover, according to [4, Lemma 2], when $g(x, y)$, $c(x, y)$ is locally Lipschitz at $(x, y_g^*(x))$, there is

$$\nabla v(x) = \nabla_x c(x, y) + \langle \lambda_g^*(x), c(x, y)\rangle. \tag{59}$$

Therefore, we know

$$\Big\langle \lim_{r\downarrow 0} \frac{\lambda_g^*(x + rd) - \lambda_g^*(x)}{r}, c(x, y_g^*(x)) \Big\rangle = 0. \tag{60}$$

□

In this way, we can similarly show a generalized version of Lemma 7.

Lemma 14. *Suppose all conditions in Lemma 10 hold. For any unit direction $d \in \mathbb{R}^{d_x}$, there exists an index set $\mathcal{I} \subseteq [d_y]$ such that the directional derivative of $y_g^*(x)$, $D_d(y_g^*(x))$ satisfies*

$$\Big[\lim_{r\downarrow 0} \frac{y_g^*(x + rd) - y_g^*(x)}{r}\Big]_{[d_y]\setminus\mathcal{I}} = 0, \quad \textit{and} \tag{61}$$

$$\Big[\lim_{r\downarrow 0} \frac{y_g^*(x + rd) - y_g^*(x)}{r}\Big]_{\mathcal{I}} = C'(x)_{[\mathcal{I},\mathcal{I}]}^{-1} B'(x)_{[:,\mathcal{I}]}{}^\top d, \tag{62}$$

and the 2nd-order directional derivative of $v(x)$ in direction d is

$$\begin{aligned}&D_{dd}^2(v(x))\\ =&d^\top B'(x)_{[:,\mathcal{I}]} C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1\top} C(x)_{[\mathcal{I},\mathcal{I}]} C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1} B'(x)_{[:,\mathcal{I}]}{}^\top d\\ &+ d^\top \nabla_{xx} A(x) d\\ &- d^\top B'(x)_{[:,\mathcal{I}]} C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1} B(x)_{[\mathcal{I},:]} d\\ &- d^\top B(x)_{[:,\mathcal{I}]} C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1} B'(x)_{[:,\mathcal{I}]}{}^\top d.\end{aligned} \tag{63}$$

where $A(x)$, $B(x)$, and $C(x)$ are defined in (12)*, and*

$$\begin{aligned}B'(x) &= \nabla_{xy} L_g(x, y_g^*(x), \lambda_g^*(x)),\\ C'(x) &= \nabla_{yy} L_g(x, y_g^*(x), \lambda_g^*(x)).\end{aligned}$$

Proof. Fix arbitrary unit direction d, denote $y_{g,r}$ as solution to $\min_{y\in\mathcal{Y}(x)} g(x + rd + \frac{1}{2}r'^2 d)$. According to [1], the second order directional derivative of $v(x)$ in direction unit direction

d is

$$
\begin{aligned}
&D^2_{dd}(v(x)) \\
&= \lim_{r,r'\downarrow 0} \frac{1}{\frac{1}{2}rr'}\Big(g(x+rd+\frac{1}{2}r^2 d, y_{g,r}) - g(x, y_g^*(x)) \\
&\qquad - \langle \nabla v(x), rd\rangle\Big) \\
&\overset{(a)}{=} \lim_{r,r'\downarrow 0} \frac{1}{\frac{1}{2}rr'}\Big(\langle \nabla_x g(x, y_g^*(x)), rd\rangle \\
&\qquad + \langle \nabla_y g(x, y_g^*(x)), y_{g,r} - y_g^*(x)\rangle \\
&\qquad + \frac{1}{2}(y_{g,r} - y_g^*(x))^\top \nabla_{yy} g(x, y_g^*(x))(y_{g,r} - y_g^*(x)) \\
&\qquad + \frac{rr'}{2} d^\top \nabla_{xx} g(x, y_g^*(x)) d \\
&\qquad + \frac{r'}{2} d^\top \nabla_{xy} g(x, y_g^*(x))(y_{g,r} - y_g^*(x)) \\
&\qquad + \frac{r}{2}(y_{g,r} - y_g^*(x))^\top \nabla_{yx} g(x, y_g^*(x)) d \\
&\qquad - \langle \nabla_x g(x, y_g^*(x)) + \lambda_g^\top \nabla_x c(x, y_g^*(x)), rd\rangle\Big) \\
&\overset{(b)}{=} \Big(\lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\Big)^\top \nabla_{yy} g(x, y_g^*(x)) \Big(\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\Big) \\
&\quad + d^\top \nabla_{xx} g(x, y_g^*(x)) d \\
&\quad + \Big(\lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r}\Big)^\top \nabla_{xy} g(x, y_g^*(x)) d \\
&\quad + d^\top \nabla_{xy} g(x, y_g^*(x)) \Big(\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\Big).
\end{aligned} \tag{64}
$$

where the (a) is by 2nd order of Taylor expansion; (b) is from $\langle \nabla_x g(x, y_g^*(x)), rd\rangle + \langle \nabla_y g(x, y_g^*(x)), y_{g,r} - y_g^*(x)\rangle = \langle \nabla_x g(x, y_g^*(x)) + \lambda_g^\top \nabla_x c(x, y_g^*(x)), rd\rangle$ and they are two equivalent expression of the directional $\nabla v(x)$. Notice that the directional hessian relies on $\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}$, we investigate it in the following.

Following Lemma 12 and 13, we have

$$
\begin{aligned}
&\langle \nabla_y g(x, y_g^*(x)) + \lambda_g^\top \nabla_y c(x, y_g^*(x)), \lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\rangle = 0, \\
&\text{and } \langle c(x, y_g^*(x)), \lim_{r\downarrow 0} \frac{\lambda_{g,r} - \lambda_g}{r}\rangle = 0,
\end{aligned} \tag{65}
$$

In this way, denote $y_{g,r}$ as the optimal solution for $\min_{y\in\mathcal{Y}(x)} g(x+rd, y)$ and $\lambda_{g,r}$ as the corresponding Lagrangian multiplier, we have

$$
\begin{aligned}
0 &= \lim_{r\downarrow 0} \frac{1}{r}\langle \nabla_y g(x+rd, y_{g,r}) + \lambda_{g,r}^\top \nabla_y c(x+rd, y_{g,r}) \\
&\quad - (\nabla_y g(x, y_g^*(x)) + \lambda_g^\top \nabla_y c(x, y_g^*(x))), \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\rangle \\
&= \lim_{r\downarrow 0} \frac{1}{r}\langle \nabla_y g(x+rd, y_{g,r}) + \lambda_g^\top \nabla_y c(x+rd, y_{g,r}) \\
&\quad - (\nabla_y g(x, y_g^*(x)) + \lambda_g^\top \nabla_y c(x, y_g^*(x))), \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\rangle \\
&\quad + \lim_{r\downarrow 0} \frac{1}{r}\langle (\lambda_{g,r} - \lambda_g)^\top \nabla_y c(x+rd, y_{g,r}), \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\rangle \\
&\overset{(a)}{=} \lim_{r\downarrow 0} \frac{1}{r}\langle \nabla_y g(x+rd, y_{g,r}) + \lambda_g^\top \nabla_y c(x+rd, y_{g,r}) \\
&\quad - (\nabla_y g(x, y_g^*(x)) + \lambda_g^\top \nabla_y c(x, y_g^*(x))), \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\rangle \\
&\quad + \lim_{r'\downarrow 0} \frac{1}{r'}(\lim_{r\downarrow 0} \frac{\lambda_{g,r} - \lambda_g}{r})^\top (c(x+rd, y_{g,r'}) \\
&\quad - c(x+rd, y_g)) \\
&\overset{(b)}{=} \lim_{r\downarrow 0}\langle \nabla_y g(x+rd, y_{g,r}) + \lambda_g^\top \nabla_y c(x+rd, y_{g,r}) \\
&\quad - (\nabla_y g(x, y_g^*(x)) + \lambda_g^\top \nabla_y c(x, y_g^*(x))), \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\rangle \\
&\overset{(c)}{=} \Big\langle (\nabla_{yx} g(x, y_g^*(x)) + \lambda_g^\top \nabla_{yx} c(x, y_g^*(x)))d \\
&\quad + (\nabla_{yy} g(x, y_g^*(x)) + \lambda_g^\top \nabla_{yy} c(x, y_g^*(x))) \Big(\lim_{r'\downarrow 0} \frac{y_{g,r} - y_g}{r}\Big), \\
&\quad \lim_{r'\downarrow 0} \frac{y_{g,r'} - y_g}{r'}\Big\rangle.
\end{aligned}
$$

where (a) uses Taylor expansion; (b) is by choosing $r' = r$ and calling (65); and (c) follows Taylor expansion as well.

In this way, denote $\mathcal{I}$ as the index set for non-zero elements in $\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}$, we have

$$
\begin{aligned}
0 =& \nabla_{xy} L_g(x, y_g^*(x), \lambda_g^*(x))_{[:,\mathcal{I}]}{}^\top d \\
&+ \nabla_{yy} L_g(x, y_g^*(x), \lambda_g^*(x))_{[\mathcal{I},\mathcal{I}]} \Big[\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\Big]_{\mathcal{I}}
\end{aligned} \tag{66}
$$

where $L_g(x, y, \lambda)$ is defined in (16). In this way, following a similar analysis as the one for Lemma 7, there is

$$
\begin{aligned}
&\Big[\lim_{r\downarrow 0} \frac{y_{g,r} - y_g}{r}\Big]_{\mathcal{I}} \\
&= -\nabla_{yy} L(x, y_g; \lambda_g)_{[\mathcal{I},\mathcal{I}]}{}^{-1} \nabla_{xy} L(x, y_g; \lambda_g)_{[:,\mathcal{I}]}{}^\top d \\
&= -C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1} B'(x)_{[:,\mathcal{I}]}{}^\top d
\end{aligned} \tag{67}
$$

This proves the first part of the Lemma.

In this way, plugging the result back in (64), we have

$$\begin{aligned}
&D^2_{dd}(v(x))\\
=&\Big(C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^{\top}d\Big)^{\top}\\
&C(x)_{[\mathcal{I},\mathcal{I}]}\Big(C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^{\top}d\Big)\\
&+d^{\top}\nabla_{xx}g(x,y_g^*(x))d\\
&+\Big(-C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^{\top}d\Big)^{\top}B(x)_{[\mathcal{I},:]}d\\
&+d^{\top}B(x)_{[:,\mathcal{I}]}\Big(-C'(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'(x)_{[:,\mathcal{I}]}{}^{\top}d\Big). \qquad (68)
\end{aligned}$$

Therefore, rearranging completes the proof. □

Moreover, Lemma 8 is also applicable to this coupled constraint setting $\mathcal{Y}(x) = \{y \in \mathcal{Y} : c(x,y) \leq 0\}$. In this way, the generalized version of Lemma 15 can directly be concluded with the same analysis as the one for Lemma 15. We omit the proof here as the analysis follows the same technique.

Lemma 15. *Suppose all conditions in Lemma 10 hold. Fix any unit direction $d \in \mathbb{R}^{d_x}$, denote $\mathcal{I} \subseteq [d_y]$ as the index set for non-zero elements in $D_d(y_g^*(x))$. For any $\delta > 0$, there exists a finite $\gamma^* > 0$ such that the second-order directional derivative of $v_\gamma(x)$ is*

$$\begin{aligned}
&D^2_{dd}(v_\gamma(x))\\
=&d^{\top}B'_\gamma(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}C_\gamma(x)_{[\mathcal{I},\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'_\gamma(x)_{[:,\mathcal{I}]}{}^{\top}d\\
&+d^{\top}\nabla_{xx}A_\gamma(x)d\\
&-d^{\top}B'_\gamma(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B_\gamma(x)_{[\mathcal{I},:]}d\\
&-d^{\top}B(x)_{[:,\mathcal{I}]}C'_\gamma(x)_{[\mathcal{I},\mathcal{I}]}{}^{-1}B'_\gamma(x)_{[:,\mathcal{I}]}{}^{\top}d. \qquad (69)
\end{aligned}$$

where $A(x)$, $B(x)$, and $C(x)$ are defined in (12)*, and*

$$\begin{aligned}
B'_\gamma(x) &= \nabla_{xy}L_\gamma(x, y_\gamma^*(x), \lambda_g^*(x)),\\
C'_\gamma(x) &= \nabla_{yy}L_\gamma(x, y_\gamma^*(x), \lambda_g^*(x)).
\end{aligned}$$

In this way, we can show Theorem 1 easily following a similar analysis as for Theorem 3 in Appendix III-B.

Proof of Theorem 3. As $\nabla^2 L$, $\nabla^2 L^\gamma$, $\nabla^2 g$, $\nabla^2 f$ are all Lipschitz in y, we have $\|\nabla^2 v(x) - \nabla^2 v^\gamma(x)\| = \mathcal{O}(\gamma^{-1})$ following similar analysis as in Appendix III-B. In this way,

$$\|D^2_{dd}(F_\gamma(x))\| = \gamma\|D^2_{dd}(v_\gamma(x)) - D^2_{dd}(v(x))\| = \mathcal{O}(1).$$

So we can conclude that $\nabla F_\gamma(x)$ is $\mathcal{O}(1)$-smooth. □

REFERENCES

[1] J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[2] Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity*. Foundations and Trends® in Machine Learning, 2015.

[3] Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.

[4] Liuyuan Jiang, Quan Xiao, Victor M Tenorio, Fernando Real-Rojas, Antonio Marques, and Tianyi Chen. A primal-dual-assisted penalty approach to bilevel optimization with coupled constraints. In *Advances in Neural Information Processing Systems*, 2024.

[5] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811, 2016.

[6] Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, Honolulu, HI, 2023.

[7] Gerd Wachsmuth. On licq and the uniqueness of lagrange multipliers. *Operations Research Letters*, 41(1):78–80, 2013.

[8] Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232, 1973.