Supplementary Material for "A Primal-Dual-Assisted Penalty Approach to Bilevel Optimization with Coupled Constraints"

Table of Contents

A Preliminary Knowledge 14

B Analysis of the Penalty-Based Lagrangian Reformulation 15
    B.1 Proof of Lemma 1 15
    B.2 Proof of Theorem 1 15

C Analysis of the Differentiability of Value Functions 16
    C.1 Proof of Lemma 2 17
    C.2 Proof of Lemma 3 17

D Convergence Analysis of the Main Result 18
    D.1 Proof of Theorem 12 18
    D.2 Proof of Theorem 3 19
    D.3 Proof of Theorem 4 22

E Applications to SVM model training 26
    E.1 Problem introduction 26
    E.2 Additional Experiments 26

F Applications to Transportation Network Planning 27
    F.1 Problem introduction 27
    F.2 Experiment roadmap 29
    F.3 A 3-node network experiment 30
    F.4 A 9-node network experiment 31
    F.5 Seville network experiment 31

A Preliminary Knowledge

Definition 5. *For a convex function $h : \mathbb{R}^{d_q} \to \mathbb{R}$ whose domain is $\mathcal{Q} \subseteq \mathbb{R}^{d_q}$, the Legendre conjugate of $h^* : \mathcal{Q}^* \to \mathbb{R}$ is defined as:*

$$h^*(y) := \sup_{q' \in \mathcal{Q}} \{\langle q', q \rangle - h(q')\} = -\inf_{q' \in \mathcal{Q}} \{-\langle q', q \rangle + h(q')\},$$

$$\forall q \in \mathcal{Q}^* := \{q \in \mathbb{R}^{d_q} : \sup_{q' \in \mathcal{Q}} \{\langle q', q \rangle - g(q')\} < \infty\}.$$

Remark 1. *When h is strongly convex in $\mathbb{R}^{d_q}$, it is lower bounded and therefore $\mathcal{Q}^* = \mathbb{R}^{d_y}$.*

Definition 6. *The function $h : \mathbb{R}^{d_q} \to \mathbb{R}$ is called closed if its epigraph on its domain $\mathcal{Q}$ is closed.*

Lemma 4. *Suppose $h : \mathbb{R}^{d_y} \to \mathbb{R}$ is $l_{h,1}$-smooth and α_h-strongly convex and its domain $\mathcal{Q} \subseteq \mathbb{R}^{d_q}$ is convex, closed and non-empty.*

    1. If additionally $\mathcal{Q} = \mathbb{R}^{d_q}$, the gradient mappings ∇h and ∇h^* are inverse of each other ([53]); and $h^* : \mathbb{R}^{d_q} \to \mathbb{R}$ is $\frac{1}{\alpha_h}$-smooth and $\frac{1}{l_{h,1}}$-strongly convex (Proposition 2.6 [3]).

    2. If $\mathcal{Q} \subset \mathbb{R}^{d_q}$, h^* is $\frac{1}{\alpha_h}$-smooth ([34]) and and convex (Theorem 4.43 [30]).

Lemma 5. *Suppose $h : \mathbb{R}^{d_q} \to \mathbb{R}$ is strongly convex on domain convex, closed and non-empty $\mathcal{Q}$, $h^c : \mathbb{R}^{d_q} \to \mathbb{R}^{d_c}$ is convex in q and d_c is finite, and $\{q \in \mathcal{Q} : h^c(q) \le 0\}$ is non-empty.*

    1. The problem $\min_{q \in \{q \in \mathcal{Q} : h^c(q) \le 0\}} h(q)$ has a unique feasible solution.

14

2. *When linear independence constraint qualification (LICQ) condition additionally holds for g^c, the corresponding Lagrange multiplier, i.e. solution to the problem $\max_{\mu \in \mathbb{R}_+^{d_c}} \min_{q \in \mathcal{Q}} h(y) + \langle \mu, h^c(q) \rangle$ is unique [65].*

Lemma 6 (Lemma 3.1 in [8]; Lemma 2.11 in [54]). *Suppose $\mathcal{Q} \subseteq \mathbb{R}^{d_q}$ is convex, closed, and nonempty. For any $q_1 \in \mathbb{R}^{d_q}$ and any $q_2 \in \mathcal{Q}$,*

$$\langle \operatorname{Proj}_{\mathcal{Q}}(q_1) - q_2, \operatorname{Proj}_{\mathcal{Q}}(q_1) - q_1 \rangle \leq 0. \tag{20}$$

In this way, take $q_1 = q_3 - \eta g$ for any $q_3 \in \mathcal{Q}$, and denote $q_3^+ = \operatorname{Proj}_{\mathcal{Q}}(q_3 - \eta g)$,

$$\langle g, q_3^+ - q_2 \rangle \leq -\frac{1}{\eta} \langle q_3^+ - q_2, q_3^+ - q_3 \rangle. \tag{21}$$

Lemma 7 (Theorem 3.10 [8]). *Suppose a differentiable function h is $l_{h,1}$-smooth and α_{h_2}-strongly convex. Consider the constrained problem $\min_{q \in \mathcal{Q}} h(q)$ where $\mathcal{Q}$ is non-empty, closed and convex. Projected Gradient Descent with $\eta \leq \frac{1}{l_{h,1}}$ converges linearly to the unique $q^* = \arg \min_{q \in \mathcal{Q}} h(q)$:*

$$\| \operatorname{Proj}_{\mathcal{Q}}(q^t - \eta \nabla h(q^t)) - q^* \| \leq (1 - \alpha \eta)^{1/2} \| q^t - q^* \| \leq (1 - \alpha \eta / 2) \| q^t - q^* \|. \tag{22}$$

B Analysis of the Penalty-Based Lagrangian Reformulation

B.1 Proof of Lemma 1

Proof. According to Lemma 5, for any fixed x, there exist a unique $\mu_g^*(x)$ such that the primal problem is $g(x, y) + \langle \mu_g^*(x), g^c(x, y) \rangle$. This problem is α_g-strongly convex with respect to y. This α_g is independent from x and therefore the quadratic growth in statement 1 can be concluded following Theorem 2 in [35].

As g is strongly convex and continuous, and $\mathcal{Y}(x)$ is a closed set, there exists a unique solution $y_g^*(x)$ such that $g(x, y_g^*(x)) = v(x)$. If $y \neq y_g^*(x)$ and $y \in \mathcal{Y}(x)$, $g(x, y) > v(x)$, which completes the proof of statement 2. $\qquad\square$

B.2 Proof of Theorem 1

Proof. We know from Lemma 1 that $g(x, y) - v(x) \geq \|y - y_g^*(x)\|^2$ and $g(x, y) = v(x)$ if and only if $y = y_g^*(x)$. This is a squared-distance bound following Definition 1 in [57]. Under Lipschitzness of $f(x, y)$ with respect to y, the ϵ-approximate problem is equivalent to its penalty reformulation

$$\min_{(x,y) \in \{\mathcal{X} \times \mathcal{Y} : g^c(x,y) \leq 0\}} f(x, y) + \gamma(g(x, y) - v(x)) \tag{23}$$

with $\gamma = o(\epsilon^{-0.5})$ following Theorems 1 and 2 in [57]. This is in equivalence to

$$\min_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}(x)} f(x, y) + \gamma(g(x, y) - v(x)). \tag{24}$$

Suppose $(x_0, y_0) \in \{\mathcal{X} \times \mathcal{Y} : g^c(x, y) \leq 0\}$ being a solution to (23). Suppose for any $x \in \mathcal{X}$, $y_F^*(x) \in \arg \min_{y \in \mathcal{Y}(x)} f(x, y) + \gamma(g(x, y) - v(x))$. We know that for any $x \in \mathcal{X}$, $y \in \mathcal{Y}(x)$,

$$f(x_0, y_0) + \gamma(g(x_0, y_0) - v(x_0)) \leq f(x, y_F^*(x)) + \gamma(g(x, y_F^*(x)) - v(x))$$
$$\leq f(x, y) + \gamma(g(x, y) - v(x)).$$

This means any solution to (23) is a solution to (24). On the other hand, suppose $x_0 \in \mathcal{X}$, $y_F^*(x_0) \in \mathcal{Y}(x_0)$ is a solution to (24). We know that for any $(x, y) \in \{\mathcal{X} \times \mathcal{Y} : g^c(x, y) \leq 0\}$,

$$f(x_0, y_F^*(x_0)) + \gamma(g(x_0, y_F^*(x_0)) - v(x_0)) \leq f(x, y_F^*(x)) + \gamma(g(x, y_F^*(x)) - v(x))$$
$$\leq f(x, y) + \gamma(g(x, y) - v(x)).$$

This means any solution to (24) is a solution to (23).

Besides, we know $f(x, y)$ is $l_{f,1}$-smooth, $g(x, y)$ is α_g-strongly convex in y, by the definitions, we know $f(x, y) + \gamma(g(x, y) - v(x))$ is $(\gamma \alpha_g - l_{f,1})$-strongly convex in y as

$$f(x, y_1) + \gamma(g(x, y_1) - v(x)) - f(x, y_2) + \gamma(g(x, y_2) - v(x))$$

15

$$=f(x,y_1) - f(x,y_2) + \gamma(g(x,y_1) - g(x,y_2))$$

$$\geq \langle \nabla_y f(x,y_2), y_1 - y_2 \rangle - \frac{l_{f,1}}{2}\|y_1 - y_2\|^2 + \gamma\langle \nabla_y g(x,y_2), y_1 - y_2 \rangle + \gamma\frac{\alpha_g}{2}\|y_1 - y_2\|^2$$

$$=\langle \nabla_y f(x,y_2) + \gamma\nabla_y g(x,y_2), y_1 - y_2 \rangle + \frac{\gamma\alpha_g - l_{f,1}}{2}\|y_1 - y_2\|^2. \tag{25}$$

Moreover, according to Assumption 2 the constraint $g^c(x,y)$ is convex in y, and $\min_{y \in \mathcal{Y}(x)} f(x,y) + \gamma(g(x,y) - v(x))$ is equivalent to its equivalent *Lagrangian Dual Form*

$$\max_{\mu \in \mathbb{R}_+^{d_c}} \min_{y \in \mathcal{Y}} f(x,y) + \gamma(g(x,y) - v(x)) + \langle \mu, g^c(x,y) \rangle \tag{26}$$

according to the Lagrangian Duality theory, as in Chapter 4 in [54]. Therefore, (23) can be recovered to (2a) and this completes the proof. $\square$

C   Analysis of the Differentiability of Value Functions

Lemma 8 (Theorem 2.16 in [30]). *Suppose $h(x,y)$ is strongly convex in y and is Lipschitz with respect to x, $h^c(x,y)$ is convex in y and is Lipschitz with respect to x, and both $\mathcal{Y}$ and $\{y \in \mathcal{Y} : h^c(x,y) \leq 0\}$ are non-empty, closed, and convex. For the problem $\min_{y \in \{y \in \mathcal{Y} : h^c(x,y) \leq \mathbf{0}\}} h(x,y)$, the unique solution $y_h^*(x)$ and unique Lagrange multiplier $\mu_h^*(x)$, defined as*

$$(y_h^*(x), \mu_h^*(x)) := \arg\max_{\mu \in \mathbb{R}_+^{d_x}} \min_{y \in \mathcal{Y}} h(x,y) + \langle \mu, h^c(x,y) \rangle, \tag{27}$$

is Lipschitz in x. In other words, there exist $L_h \geq 0$ that, for all $x_1, x_2 \in \mathcal{X}$,

$$\|(y_h^*(x_1); \mu_h^*(x_1)) - (y_h^*(x_2); \mu_h^*(x_2))\| \leq L_h\|x_1 - x_2\|.$$

Remark 2. *This also implies the L_h-continuity of both $y^*(x)$ and $\mu^*(x)$.*

Remark 3. *When $h(x,y) = g(y)$, and $h^c(x,y) = A^\top y - x$, the Lipschitzness of both $y^*(x)$ and $\mu^*(x)$ in x holds automatically.*

Lemma 9 (Theorem 4.24 in [6]). *Consider the value function for the constrained problem*

$$v_h(x) = \min_{y \in \mathcal{Y}} h_0(x,y) \quad s.t. \quad h_i(x,y) \leq 0, \ i = 1, \ldots, I, \tag{P_x}$$

where $\mathcal{Y}$ is convex, closed, and non-empty. Denote $(S(x), \Lambda(x))$ as the solution sets for y and the Lagrange multipliers $(\mu_1, \ldots, \mu_I)$:

$$(S(x), \Lambda(x)) := \arg\min_y \max_{(\mu_1, \ldots, \mu_I) \geq 0} h_0(x,y) + \sum_{i=1}^{I} \mu_i h_i(x,y).$$

If the following conditions hold:

1. *$h_0(x, \cdot)$ is convex and the solution set $S(x)$ is non-empty.*

2. *The directional regularity condition in a direction d, holds for all $y \in S(x)$.*

3. *For a sequence $t_n \to 0$, define the sequence $x_n := x + t_n d + O(t_n)$. If (P_{x_n}) is attained by an $O(t_n)$-optimal solution sequence y_n with a limit point (in the strong topology) $y \in S(x)$.*

Then $v_h(x)$ is Hadamard directionally differentiable at x in the direction d, and the directional derivative can be written as

$$v_h'(x,d) = \inf_{y \in S(x)} \sup_{(\mu_1, \ldots, \mu_I) \in \Lambda(x)} \nabla_x \left(h_0(x,y) + \sum_{i=1}^{I} \mu_i h_i(x,y) \right).$$

Remark 4. *When $y_h^*, \mu_h^* = (\mu_{h,1}^*, \ldots, \mu_{h,I}^*)$ are unique in $S(x)$ and $\Lambda(x)$, we have:*

$$v_h'(x,d) = \nabla_x \left(h_0(x,y^*) + \sum_{i=1}^{I} \mu_i^* h_i(x,y^*) \right).$$

16

617 Before proving Lemma 2 and 3, we would like to introduce a more general form.

618 **Lemma 10.** *Suppose $\mathcal{Y}$ and $\{y \in \mathcal{Y} : h^c(x,y) \leq 0\}$ are both non-empty, closed and convex, $h(x,y)$*
619 *is jointly smooth and strongly convex in y, $h^c(x,y)$ is convex in y, and both $h(x,y)$ and $h^c(x,y)$ are*
620 *Lipschitz with respect to x.*

$$v_h(x) = \min_{y \in \mathcal{Y}} h(x,y) \quad s.t. \quad h^c(x,y) \leq 0$$

621 *is differentiable with*

$$\nabla v_h(x) = \nabla_x h(x, y_h^*(x)) + \langle \mu_h^*(x), h^c(x, y_h^*(x)) \rangle,$$

622 *where $(y_h^*(x), \mu_h^*(x))$ defined in (27) are unique.*

623 *Proof.* As $h(x,y)$ being strongly convex in y, condition 1 in Lemma 9 is satisfied and the solution
624 sets are of singleton value $(y_h^*(x), \mu_h^*(x))$ according to Lemma 5. Moreover, the smoothness of
625 $f(x,y)$ guarantees Robinson's constraint qualification [2], which implies the directional regularity
626 condition for any direction d (Theorem 4.9. (ii) in [6]). Additionally, under the Lipschitzness of
627 $h(x,y)$ and $h^c(x,y)$ with respect to x, $y_h^*(x), \mu_h^*(x)$ are Lipschitz according to Lemma 8. This
628 guarantees Condition 2 in Lemma 9 can be satisfied for all directions d. Condition 3 is a direct
629 outcome from Lemma 8. This completes the proof. $\square$

C.1 Proof of Lemma 2

631 *Proof.* The problem $\min_{y \in \mathcal{Y}} g(x,y)$ s.t. $g^c(x,y) \leq 0$ fits in the setting of Lemma 10 by taking
632 $h(x,y) = g(x,y)$ and $h^c(x,y) = g^c(x,y)$. Therefore the derivative (8) can be obtained accordingly.
633 Moreover, for any $x_1, x_2 \in \mathcal{X}$,

$$\begin{aligned}
&\|\nabla v(x_1) - \nabla v(x_2)\| \\
=& \|\nabla_x g(x_1, y_g^*(x_1)) + \langle \mu_g^*(x_1), \nabla_x g^c(x_1, y_g^*(x_1)) \rangle - \nabla_x g(x_2, y_g^*(x_2)) \\
&- \langle \mu_g^*(x_2), \nabla_x g^c(x_2, y_g^*(x_2)) \rangle\| \\
\overset{(a)}{\leq}& \|\nabla_x g(x_1, y_g^*(x_1)) - \nabla_x g(x_2, y_g^*(x_2))\| \\
&+ \|\langle \mu_g^*(x_1), \nabla_x g^c(x_1, y_g^*(x_1)) \rangle - \langle \mu_g^*(x_1), \nabla_x g^c(x_2, y_g^*(x_2)) \rangle\| \\
&+ \|\langle \mu_g^*(x_1), \nabla_x g^c(x_2, y_g^*(x_2)) \rangle - \langle \mu_g^*(x_2), \nabla_x g^c(x_2, y_g^*(x_2)) \rangle\| \\
\overset{(b)}{\leq}& (l_{g,1} + B_g l_{g^c,1})(\|x_1 - x_2\| + \|y_g^*(x_1) - y_g^*(x_2)\|) + l_{g^c,0}\|\mu_g^*(x_1) - \mu_g^*(x_2)\| \\
\overset{(c)}{\leq}& ((l_{g,1} + B_g l_{g^c,1})(1 + L_g) + l_{g^c,0} L_g)\|x_1 - x_2\|,
\end{aligned}$$

634 where (a) follows triangle inequality; (b) leverage on the Lipschitzness of ∇g, g^c and ∇g^c, and
635 the upper bound for $\|\mu_g^*(x)\|$; and (c) uses the Lipschitzness of $y_g^*(x)$ and $\mu_g^*(x)$. As the bound
636 is loose due to the use of triangle inequality, we can conclude that $v(x)$ is $l_{v,1}$-smooth where
637 $l_{v,1} \leq ((1 + B_g)(1 + L_g)l_{g^c,1} + l_{g^c,0} L_g)$. $\square$

C.2 Proof of Lemma 3

639 *Proof.* As $\gamma > \frac{l_{f,1}}{\alpha_g}$, we know $f(x,y) + \gamma(g(x,y) - v(x))$ is $(\gamma \alpha_g - l_{f,1})$-strongly convex by (25).
640 By strong duality,

$$F_\gamma(x) = \min_{y \in \mathcal{Y}} f(x,y) + \gamma(g(x,y) - v(x))$$
$$\text{s.t. } g^c(x,y) \leq 0.$$

641 Considering the smoothness of $v(x)$ as presented in Lemma 2, all assumptions in Lemma 10 are
642 satisfied. Therefore the derivative (9) can be obtained. For any $x_1, x_2 \in \mathcal{X}$,

$$\begin{aligned}
&\|\nabla F(x_1) - \nabla F(x_2)\| \\
=& \|\nabla_x f(x_1, y_F^*(x_1)) + \gamma\left(\nabla_x g(x_1, y_F^*(x_1)) - \nabla v(x_1)\right) + \langle \mu_F^*(x_1), \nabla_x g^c(x_1, y_F^*(x_1)) \rangle \\
&- \nabla_x f(x_2, y_F^*(x_2)) - \gamma\left(\nabla_x g(x_2, y_F^*(x_2)) - \nabla v(x_2)\right) - \langle \mu_F^*(x_2), \nabla_x g^c(x_2, y_F^*(x_2)) \rangle\|
\end{aligned}$$

17

$$
\overset{(a)}{\leq} \|\nabla_x f(x_1, y_F^*(x_1)) - \nabla_x f(x_2, y_F^*(x_2))\| + \gamma\|\nabla_x g(x_1, y_F^*(x_1)) - \nabla_x g(x_2, y_F^*(x_2))\|
$$
$$
+ \gamma\|\nabla v(x_1) - \nabla v(x_2)\| + \|\langle \mu_F^*(x_1), \nabla_x g^c(x_1, y_F^*(x_1))\rangle - \langle \mu_F^*(x_1), \nabla_x g^c(x_2, y_F^*(x_2))\rangle\|
$$
$$
+ \|\langle \mu_F^*(x_1), \nabla_x g^c(x_2, y_F^*(x_2))\rangle - \langle \mu_F^*(x_2), \nabla_x g^c(x_2, y_F^*(x_2))\rangle\|
$$
$$
\overset{(b)}{\leq} (l_{f,1} + \gamma l_{g,1} + B_F l_{g^c,1})(\|x_1 - x_2\| + \|y_F^*(x_1) - y_F^*(x_2)\|) + \gamma l_{v,1}\|x_1 - x_2\|
$$
$$
+ l_{g^c,0}\|\mu_F^*(x_1) - \mu_F^*(x_2)\|
$$
$$
\overset{(c)}{\leq} ((l_{f,1} + \gamma l_{g,1} + B_F l_{g^c,1})(1 + L_F) + \gamma l_{v,1} + l_{f^c,0} L_F)\|x_1 - x_2\|,
$$

where (a) follows triangle inequality; (b) leverage on the Lipschitzness of ∇f, ∇g, g^c and ∇g^c, and the upper bound for $\|\mu_F^*(x)\|$; and (c) uses the Lipschitzness of $y_F^*(x)$ and $\mu_F^*(x)$. As the bound is loose due to the use of triangle equality, we can conclude that $F(x)$ is $l_{F,1}$-smooth where $l_{F,1} \leq (l_{f,1} + \gamma l_{g,1} + B_F l_{g^c,1})(1 + L_F) + \gamma l_{v,1} + l_{f^c,0} L_F$. $\qquad\square$

D Convergence Analysis of the Main Result

D.1 Proof of Theorem 12

Define the bias term $b(x_t)$ as

$$
b(x_t) := \nabla F(x_t) - g_t
$$
$$
= \left(\nabla_x f(x_t, y_F^*(x_t)) + \gamma\left(\nabla_x g(x, y_F^*(x_t)) + \nabla v(x_t)\right) + \langle \mu_F^*(x_t), \nabla_x g^c(x_t, y_F^*(x_t))\rangle\right)
$$
$$
- \left(\nabla_x f(x_t, y_{F,t}^{T_F}) + \gamma\left(\nabla_x g(x_t, y_{F,t}^{T_F}) - \nabla_x g(x_t, y_{g,t}^{T_g}) + \mu_{g,t}^{T_g}\right) + \langle \mu_F^{T_F}, \nabla_x g^c(x_t, y_{F,t}^{T_F})\rangle\right).
$$

In this way,

$$
\|b(x_t)\| \overset{(a)}{\leq} \|\nabla_x f(x_t, y_{F,t}^{T_F}) - \nabla_x f(x_t, y_F^*(x_t))\|
$$
$$
+ \gamma\left(\|\nabla_x g(x_t, y_{F,t}^{T_F}) - \nabla_x g(x_t, y_F^*(x_t))\| + \|\nabla_x g(x_t, y_{g,t}^{T_g}) - \nabla_x g(x_t, y_g^*(x_t))\| + \|\mu_{t,2}^{T_g} - \mu_g^*(x_t)\|\right)
$$
$$
+ \|\langle \mu_F^{T_F}, \nabla_x g^c(x_t, y_{F,t}^{T_F})\rangle - \langle \mu_F^*(x_t), \nabla_x g^c(x_t, y_{F,t}^{T_F})\rangle\|
$$
$$
+ \|\langle \mu_F^*(x_t), \nabla_x g^c(x_t, y_{F,t}^{T_F})\rangle - \langle \mu_F^*(x_t), \nabla_x g^c(x_t, y_F^*(x_t))\rangle\|
$$
$$
\overset{(b)}{\leq} l_{f,1}\|y_{F,t}^{T_F} - y_F^*(x_t)\| + \gamma(l_{g,1}\|y_{F,t}^{T_F} - y_F^*(x_t)\| + l_{g,1}\|y_{g,t}^{T_g} - y_g^*(x_t)\| + \|\mu_{g,t}^{T_g} - \mu_g^*(x_t)\|)
$$
$$
+ l_{g^c,0}\|\mu_{F,t}^{T_F} - \mu_F^*(x_t)\| + B_F l_{g^c,1}\|y_{F,t}^{T_F} - y_F^*(x_t)\|
$$
$$
\overset{(c)}{=} (l_{f,1} + \gamma l_{g,1} + B_F l_{g^c,0})\|y_{F,t}^{T_F} - y_F^*(x_t)\| + l_{g^c,0}\|\mu_{F,t}^{T_F} - \mu_F^*(x_t)\|
$$
$$
+ \gamma\left(l_{g,1}\|y_{g,t}^{T_g} - y_g^*(x_t)\| + \|\mu_{g,t}^{T_g} - \mu_g^*(x_t)\|\right),
$$

where (a) uses triangle inequality, (b) relies on Assumption 2 and Cauchy-Schwartz inequality, and (c) is by rearrangement. Furthermore, according to Young's inequality,

$$
\|b(x_t)\|^2 \leq 2\left((l_{f,1} + \gamma l_{g,1} + B_F l_{g^c,0})\|y_{F,t}^{T_F} - y_{F,t}^*\| + l_{g^c,0}\|\mu_{F,t}^{T_F} - \mu_{F,t}^*\|\right)^2
$$
$$
+ 2\gamma^2\left(l_{g,1}\|y_{g,t}^{T_g} - y_{g,t}^*\| + \|\mu_{g,t}^{T_g} - \mu_{g,t}^*\|\right)^2
$$
$$
= O(\gamma^2 \epsilon_F + \gamma^2 \epsilon_g).
$$

According to Lemma 3, $F_\gamma(x)$ is $l_{F,1}$-smooth in $\mathcal{X}$. The projection guarantees that x_{t+1} and x_t are in $\mathcal{X}$. In this way,

$$
F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{l_{F,1}}{2}\|x_{t+1} - x_t\|^2
$$
$$
\leq F(x_t) + \langle g_t, x_{t+1} - x_t\rangle + \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 + \langle b(x_t), x_{t+1} - x_t\rangle, \qquad (28)
$$

18

655 where the second inequality is by $\eta \leq \frac{1}{l_{F,1}}$.

656 Following lemma 6, we know that

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

657 Plugging this back to (28),

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

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

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\|G_\eta(x_t)\|^2 \leq & \frac{4}{\eta T}(F(x_0) - F(x_T)) + \frac{4}{T}\sum_{t=0}^{T-1}\|b(x_t)\|^2 \\
= & O(\eta^{-1}T^{-1}) + O(\gamma^2\epsilon_F + \gamma^2\epsilon_g) \\
= & O(\gamma T^{-1} + \gamma^2\epsilon_F + \gamma^2\epsilon_g)
\end{aligned}
$$

659 where last equality comes from $\eta = O(\gamma^{-1})$. This completes the proof.

D.2 Proof of Theorem 3

661 In Algorithm 2, with T_y being sufficiently large, we are implementing an accelerated projected
662 gradient descent on $-D(\mu)$. The following lemma presents the convergence analysis of such an
663 accelerated method on smooth and convex functions.

664 **Lemma 11.** *Suppose $h(\cdot)$ is $l_{h,1}$-smooth, and there exist a unique $q^* = \arg\min_{q \in \mathcal{Q}} h(q)$. Consider*
665 *the constrained problem $\min_{q \in \mathcal{Q}} h(q)$ where $\mathcal{Q}$ is non-empty, closed and convex. Accelerated*
666 *projected gradient descent algorithm as in (29) and (30) with step size $\eta \leq \frac{1}{l_{h,1}}$, initial value*
667 *$q_0 = q_{-1}$,*

$$
q_{t+\frac{1}{2}} = q_t + \frac{t-1}{t+2}(q_t - q_{t-1}) \tag{29}
$$
$$
q_{t+1} = \mathrm{Proj}_{\mathcal{Q}}(q_{t+\frac{1}{2}} - \eta\nabla h(q_{t+\frac{1}{2}})) \tag{30}
$$

668 *for $t = 0, \ldots, T-1$ will lead to*

$$
h(q_T) - h(q^*) < \frac{2}{\eta(T+1)^2}\|q_0 - q^*\|^2.
$$

669 *Proof.* Denote $\theta_t = \frac{2}{t+1}$, and

$$
u_t = q_{t-1} + \frac{1}{\theta_t}(q_t - q_{t-1}).
$$

670 (29) can be reformulated as

$$
q_{t+\frac{1}{2}} = (1 - \theta_{t+1})q_t + \theta_{t+1}u_t.
$$

671 In this way, we have

$$
\begin{aligned}
& h(q_{t+1}) - h(q^*) - (1 - \theta_{t+1})(h(q_t) - h(q^*)) \\
= & h(q_{t+1}) - (\theta_{t+1}h(q^*) + (1 - \theta_{t+1})h(q_t)) \\
\overset{(a)}{\leq} & h(q_{t+1}) - h(\theta_{t+1}q^* + (1 - \theta_{t+1})q_t) \\
= & h(q_{t+1}) - h(q_{t+\frac{1}{2}}) + h(q_{t+\frac{1}{2}}) - h(\theta_{t+1}q^* + (1 - \theta_{t+1})q_t)
\end{aligned}
$$

19

$$\overset{(b)}{\leq} \langle \nabla h(q_{t+\frac{1}{2}}), q_{t+1} - q_{t+\frac{1}{2}} \rangle + \frac{1}{2\eta} \|q_{t+1} - q_{t+\frac{1}{2}}\|^2 + \langle \nabla h(q_{t+\frac{1}{2}}), q_{t+\frac{1}{2}} - (\theta_{t+1} q^* + (1 - \theta_{t+1}) q_t) \rangle$$

$$= \langle \nabla h(q_{t+\frac{1}{2}}), q_{t+1} - (\theta_{t+1} q^* + (1 - \theta_{t+1}) q_t) \rangle + \frac{1}{2\eta} \|q_{t+1} - q_{t+\frac{1}{2}}\|^2$$

$$\overset{(c)}{\leq} -\frac{1}{\eta} \langle q_{t+1} - (\theta_{t+1} q^* + (1 - \theta_{t+1}) q_t, q_{t+1} - q_{t+\frac{1}{2}} \rangle + \frac{1}{2\eta} \|q_{t+1} - q_{t+\frac{1}{2}}\|^2$$

$$= -\frac{1}{\eta} \langle \theta_{t+1} u_{t+1} - \theta_{t+1} \mu^*, \theta_{t+1} u_{t+1} - \theta_{t+1} u_t \rangle + \frac{1}{2\eta} \|\theta_{t+1} u_{t+1} - \theta_{t+1} u_t\|^2$$

$$= \frac{\theta_{t+1}^2}{2\eta} \left(\|u_t - q^*\|^2 - \|u_{t+1} - q^*\|^2 \right),$$

where (a) follows the convexity of h; (b) is by the smoothness of h and $\eta \leq \frac{1}{l_{h,1}}$, and the convexity of h; and (c) follows from Lemma 6 as $\theta_{t+1} q^* + (1 - \theta_{t+1}) q_t$ is a linear combination of $q_t, q^* \in \mathcal{Q}$ and is in $\mathcal{Q}$.

Rearranging gives

$$\frac{\eta}{\theta_{t+1}^2} (h(q_{t+1}) - h(q^*)) + \frac{1}{2} \|u_{t+1} - q^*\|^2 \leq (1 - \theta_{t+1}) \frac{\eta}{\theta_{t+1}^2} (h(q_t) - h(q^*)) + \frac{1}{2} \|u_t - q^*\|^2$$

$$\overset{(d)}{\leq} \frac{\eta}{\theta_t^2} (h(q_t) - h(q^*)) + \frac{1}{2} \|u_t - q^*\|^2$$

$$\overset{(e)}{\leq} \frac{\eta}{\theta_1^2} (h(q_1) - h(q^*))^t + \frac{1}{2} \|u_1 - q^*\|^2$$

$$\overset{(f)}{\leq} \frac{(1 - \theta_1)\eta}{\theta_1^2} (h(q_0) - h(q^*)) + \frac{1}{2} \|u_0 - q^*\|^2 = \|u_0 - q^*\|^2$$

where (d) is from $\frac{1-\theta_{t+1}}{\theta_{t+1}^2} \leq \frac{1}{\theta_t}$, (e) is the outcome of iteration, and (f) again uses the first inequality. Additionally, as $u_0 = q_0$, rearranging gives

$$h(q_T) - h(q^*) < \frac{2}{\eta(T+1)^2} \|q_0 - q^*\|^2. \tag{31}$$

This completes the proof. □

In this way, we are ready to proceed to the **proof of Theorem 3**.

Proof. To restate, for a fixed x, define

$$L_g(\mu, y) = g(x, y) + \langle \mu, g^c(x, y) \rangle,$$
$$L_F(\mu, y) = f(x, y) + \gamma(g(x, y) - v(x)) + \langle \mu, g^c(x, y) \rangle,$$

and

$$D_g(\mu) := \min_{y \in \mathcal{Y}} L_g(\mu, y),$$
$$D_F(\mu) := \min_{y \in \mathcal{Y}} L_F(\mu, y).$$

D_g and D_F are concave in μ according to Lemma 2.58 in [54]. Moreover, $L_g(\mu, y)$ is α_g-strongly convex and $(l_{g,1} + l_{g^c,1})$-smooth in y and $L_F(\mu, y)$ is $(\gamma \alpha_g - l_{f,1})$-strongly convex and $(l_{f,1} + \gamma l_{g,1} + l_{g^c,1})$-smooth in y. Therefore,

$$y_g^*(\mu; x) := \arg\min_{y \in \mathcal{Y}} L_g(\mu, y),$$
$$y_F^*(\mu; x) := \arg\min_{y \in \mathcal{Y}} L_F(\mu, y)$$

20

are respectively $\frac{1}{\alpha_g}$ and $\frac{1}{\gamma\alpha_g - l_{f,1}}$-Lipschitz to μ (Theorem F.10 in [16]; Theorem 4.47 in [30]). In this way, following Lemma 9, we have

$$\nabla D_g(\mu) = \nabla_\mu L_g(\mu, y_g^*(\mu; x)) = g^c(x, y_g^*(\mu; x)),$$
$$\nabla D_F(\mu) = \nabla_\mu L_F(\mu, y_F^*(\mu; x)) = g^c(x, y_F^*(\mu; x)).$$

Additionally, g^c is $l_{g^c,0}$-Lipschitz by Assumption 1, in this way, for any $\mu_1, \mu_2 \in \mathbb{R}_+^{d_c}$:

$$\|\nabla D_g(\mu_1) - \nabla D_g(\mu_2)\| = \|g^c(x, y_g^*(\mu_1; x)) - g^c(x, y_g^*(\mu_2; x))\|$$
$$\leq l_{g^c,0}\|y_g^*(\mu_1; x) - y_g^*(\mu_2; x)\| \leq \frac{l_{g^c,0}}{\alpha_g}\|\mu_1 - \mu_2\|. \tag{32}$$

and

$$\|\nabla D_F(\mu_1) - \nabla D_F(\mu_2)\| = \|g^c(x, y_F^*(\mu_1; x)) - g^c(x, y_F^*(\mu_2; x))\|$$
$$\leq l_{g^c,0}\|y_F^*(\mu_1; x) - y_F^*(\mu_2; x)\| \leq \frac{l_{g^c,0}}{\gamma\alpha_g - l_{f,1}}\|\mu_1 - \mu_2\|. \tag{33}$$

We can conclude that D_g and D_F are respectively $\frac{l_{g^c,0}}{\alpha_g}$ and $\frac{l_{g^c,0}}{\gamma\alpha_g - l_{f,1}}$-smooth.

Fixing $\mu_{t+\frac{1}{2}}$, steps 4-7 are T_y-step projected gradient descent on y with step size $\eta_{g,1} \leq \frac{1}{l_{g,1}+l_{g^c,1}}$ and $\eta_{F,1} \leq \frac{1}{l_{f,1}+\gamma l_{g,1}+l_{g^c,1}}$ respectively for the two problems to have linear convergence according to Lemma 7. For $T_y = O(\log(\epsilon_g^{-1}))$, we know $\|y_{t+1} - y_g^*(\mu_{t+\frac{1}{2}}; x)\| = O(\epsilon_g)$ to solve (7). For $T_y = O(\log(\epsilon_F^{-1}))$, we know $\|y_{t+1} - y_F^*(\mu_{t+\frac{1}{2}}; x)\| = O(\epsilon_F)$ to solve (5).

The algorithm is, therefore, an accelerated projected gradient descent method on $-D_g(\mu)$ and $-D_F(\mu)$, both of which are convex and smooth. By Lemma 11, we can conclude the complexity is $\tilde{O}(\epsilon_g^{-0.5})$ for conducting on (7) to achieve

$$D_g(\mu_g^*(x)) - D_g(\mu_{T_g}) < \epsilon_g, \tag{34}$$

and similarly, the complexity is $\tilde{O}(\epsilon_F^{-0.5})$ for conducting on (5) to achieve

$$D_F(\mu_F^*(x)) - D_F(\mu_{T_F}) < \epsilon_F. \tag{35}$$

Moreover, the problems

$$\max_{\mu\in\mathbb{R}_+^{d_c}} D_g(\mu) \quad \text{and} \quad \max_{\mu\in\mathbb{R}_+^{d_c}} D_F(\mu)$$

are respectively equivalent to the respective unconstrained problems with the Lagrange multipliers

$$\max_{\mu\in\mathbb{R}^{d_c}} \tilde{D}_g(\mu) := D_g(\mu) + \lambda_g^\top\mu \quad \text{and} \quad \max_{\mu\in\mathbb{R}^{d_c}} \tilde{D}_F(\mu) := D_F(\mu) + \lambda_F^\top\mu$$

for some λ_g, λ_F being non-negative and finite in all dimension, i.e. $0 \leq \lambda_g < \infty$, $0 \leq \lambda_F < \infty$, and

$$\lambda_g^\top\mu_g^*(x) = 0 \quad \text{and} \quad \lambda_F^\top\mu_F^*(x) = 0, \tag{36}$$

as $D_g(\mu)$ and $D_F(\mu)$ are both concave in μ and $\mu \in \mathbb{R}_+^{d_c}$ is equivalent to $\mu \geq 0$. These properties are well-known, see details in Chapter 4 in [54]. The first-order stationary condition requires $\nabla\tilde{D}_g(\mu_g^*(x)) = \nabla D_g(\mu_g^*(x)) + \lambda_g = 0$ and $\nabla\tilde{D}_F(\mu_F^*(x)) = \nabla D_F(\mu_F^*(x)) + \lambda_F = 0$ and therefore

$$\nabla D_g(\mu_g^*(x)) = -\lambda_g \quad \text{and} \quad \nabla D_F(\mu_F^*(x)) = -\lambda_F. \tag{37}$$

In this way, for all $\mu \in \mathcal{B}(\mu_g^*(x); \delta_g) \cap \mathbb{R}_+^{d_c}$.

$$D_g(\mu_g^*(x)) - D_g(\mu) = \int_{\tau=0}^1 \langle\nabla D_g(\mu + \tau(\mu_g^*(x) - \mu)), \mu_g^*(x) - \mu\rangle d\tau$$
$$= \int_{\tau=0}^1 \frac{1}{\tau}\langle\nabla D_g(\mu_g^*(x)) - D_g(\mu + \tau(\mu_g^*(x) - \mu)), \tau(\mu - \mu_g^*(x))\rangle d\tau$$

21

$$- \langle \nabla D_g(\mu_g^*(x)), \mu - \mu_g^*(x) \rangle$$
$$\overset{(a)}{\geq} \int_0^1 C_{\delta_g} \|\mu - \mu_g^*(x)\|^2 \tau d\tau - \langle \nabla D_g(\mu_g^*(x)), \mu - \mu_g^*(x) \rangle$$
$$\overset{(b)}{=} \frac{C_{\delta_g}}{2} \|\mu - \mu_g^*(x)\|^2 + \langle \lambda_g, \mu - \mu_g^*(x) \rangle$$
$$\overset{(c)}{\geq} \frac{C_{\delta_g}}{2} \|\mu - \mu_g^*(x)\|^2,$$

where (a) uses (15) and the fact that the $\mu, \mu_g^*(x) \in \mathcal{B}(\mu_g^*(x); \delta_g) \cap \mathbb{R}_+^{d_c}$ implies $\mu + \tau(\mu_g^*(x) - \mu) \in \mathcal{B}(\mu_g^*(x); \delta_g) \cap \mathbb{R}_+^{d_c}$; (b) solves the integral and $\lambda_g = -\nabla D_g(\mu_g^*(x))$; and (c) follows from the fact that $\langle \lambda, \mu_g^*(x) \rangle = 0$ by the nature of the Lagrangian reformulated objective (Chapter 4 in [54]) and $\mu, \lambda_g \geq 0$.

Analogously, for all $\mu \in \mathcal{B}(\mu_F^*(x); \delta_F) \cap \mathbb{R}_+^{d_c}$,

$$D_F(\mu_F^*(x)) - D_F(\mu) \geq \frac{C_{\delta_F}}{2} \|\mu - \mu_F^*(x)\|^2.$$

In this way, for all $\epsilon_g < \frac{C_{\delta_g}}{2} \delta_g$, when it achieves (34) with complexity $\tilde{O}(\epsilon_g^{-0.5})$ to solve (7),

$$\|\mu_{T_g} - \mu_g^*(x)\|^2 = O(\epsilon_g),$$
$$\text{and } \|y_{T_g} - y_g^*(x)\|^2 \leq \|y_{T_g} - y_g^*(\mu_{T_g}; x)\|^2 + \|\mu_{T_g} - \mu_g^*(x)\|^2$$
$$\leq (1/\alpha_g + 1) \|\mu_{T_g} - \mu_g^*(x)\|^2 = O(\epsilon_g).$$

Similarly, for (5), for all $\epsilon_g < \frac{C_{\delta_g}}{2} \delta_g$, the complexity to achieve (35) and

$$\|\mu_{T_F} - \mu_F^*(x)\|^2 = O(\epsilon_F),$$
$$\text{and } \|y_{T_F} - y_F^*(x)\|^2 \leq \|y_{T_F} - y_F^*(\mu_{T_F}; x)\|^2 + \|\mu_{T_F} - \mu_F^*(x)\|^2$$
$$\leq (1/\alpha_F + 1) \|\mu_{T_F} - \mu_F^*(x)\|^2 = O(\epsilon_F)$$

is $\tilde{O}(\epsilon_F^{-0.5})$. This completes the proof. $\square$

D.3 Proof of Theorem 4

In this section, we consider

$$g^c(x, y) = g_1^c(x)^\top y - g_2^c(x) \tag{38}$$

being affine in y, and $\mathcal{Y} = \mathbb{R}^{d_y}$.

Therefore, for fixed x, take (7) and (5) as $L(\mu, y)$ both fit into a special case of *strongly-convex-concave saddle point problems* in the following form:

$$\max_{\mu \in \mathbb{R}_+^{d_c}} \min_{y \in \mathbb{R}^{d_y}} -h_1(\mu) + y^\top A\mu + h_2(y). \tag{39}$$

For (7), $A = g_1^c(x)$, $h_1(\mu) = g_2^c(x)^\top \mu$ is convex (linear) in μ, and $h_2(y) = g(x, y)$ is α_g-strongly convex in y. For (5), $A = g_1^c(x)$, $h_1(\mu) = g_2^c(x)^\top \mu$ is convex (linear) in μ, $h_2(y) = g(x, y)$ is $\gamma \alpha_g - l_{f,1}$-strongly convex in y.

In this way, we would like to show the effectiveness of the single-loop algorithm, Algorithm 2 without acceleration and $T_y = 1$, on the problems in (39), which is a general form to 7 and (5). In other words, we are going to prove Theorem 7, which is a more general theorem to Theorem 4.

Theorem 7. *Suppose $L(\mu, y)$ is in the form of (39) where A is full rank in column, h_1 is concave and $l_{h_1,1}$-smooth, h_2 is α_{h_2}-strongly convex and $l_{h_2,1}$-smooth satisfying $l_{h_1,1} = O(1)$, $l_{h_2,1}, l_{\alpha_2} \geq O(1)$, and $\frac{l_{h_2,1}}{\alpha_{h_2}} = O(1)$. Conduct Algorithm 2 without acceleration, $T_y = 1$, $\eta_1 = O(\frac{1}{l_{h_2,1}}) \leq \frac{1}{l_{h_2,1}}$, $\eta_2 = O(\epsilon) \leq \frac{1}{l_{h_1,1} + \sigma_{\max}^2(A)/\alpha_{h_2}}$ for arbitrary small positive $\epsilon \leq \left(\frac{4l_{h_2,1}\sigma_{\max}(A)}{\alpha_{h_2}\sigma_{\min}^2(A)} (l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}) \right)^{-1}$, yields output (μ_T, y_T) such that*

$$\|\mu_t - \mu^*\|^2 < \epsilon, \quad \text{and} \quad \|y_t - y^*\|^2 < \epsilon$$

with complexity $O(\log(\epsilon^{-1}))$. Here, $(\mu^*, y^*) = \arg\max_{\mu \in \mathbb{R}^{d_c}} \min_{y \in \mathbb{R}^{d_y}} L(\mu, y)$.

22

Remark 5. *It emphasizes on $l_{h_1,1} = O(1)$, $l_{h_2,1}, l_{\alpha_2} \geq O(1)$, and $\frac{l_{h_2,1}}{\alpha_{h_2}} = O(1)$ as $l_{h_2,1}, l_{\alpha_2} \propto \gamma$ when taking (5) as $L(\mu, y)$.*

Before proceeding, we would first look at $D(\mu)$ as (18) and conclude its smoothness and strong concavity as in the following Lemma.

Lemma 12. *Suppose h_1 is concave and $l_{h_1,1}$-smooth, h_2 is α_{h_2}-strongly convex and $l_{h_2,1}$-smooth, and A is full column rank. In this way, $D(\mu)$ defined in (18) equals*

$$D(\mu) = -h_1(\mu) - h_2^*(-A\mu),$$

and is $\frac{\sigma_{\min}^2(A)}{l_{h_2,1}}$-strongly concave and $(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})$-smooth with respect to μ.

Proof. Following Definition 5, we have

$$D(\mu) = -h_1(\mu) - h_2^*(-A\mu)$$

where h_2^* is $\frac{1}{l_{h_2,1}}$-strongly convex and $\frac{1}{\alpha_{h_2}}$-smooth according to Lemma 4.

For all μ_1, μ_2,

$$
\begin{aligned}
-D(\mu_1) - (-D(\mu_2)) =& h_2^*(-A\mu_1) - h_2^*(-A\mu_2) + h_1(\mu_1) - h_1(\mu_2) \\
\geq& \langle \frac{\partial h_2^*(-A\mu_2)}{\partial - A\mu_2}, -A\mu_1 + A\mu_2 \rangle + \frac{1/l_{h_2,1}}{2}\|A\mu_1 - A\mu_2\|^2 + \langle \nabla h_1(\mu_2), \mu_1 - \mu_2 \rangle \\
\geq& \langle \nabla D(\mu_2), \mu_1 - \mu_2 \rangle + \frac{\frac{\sigma_{\min}^2(A)}{l_{h_2,1}}}{2}\|\mu_1 - \mu_2\|^2.
\end{aligned}
$$

where the first inequality follows the strong convexity of h_2^* and the fact that $-h_1$ is convex as h_1 is concave. and the second inequality follows the chain rule to formulate $\nabla D(\mu_2)$. Therefore, $-D(\mu)$ is $\frac{\sigma_{\min}^2(A)}{l_{h_2,1}}$-strongly convex, and $D(\mu)$ is $\frac{\sigma_{\min}^2(A)}{l_{h_2,1}}$-strongly concave.

Moreover $D(\mu)$ is $(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})$-smooth as

$$
\begin{aligned}
D(\mu_1) - D(\mu_2) =& - h_2^*(-A\mu_1) - (-h_2^*(-A\mu_2)) - h_1(\mu_1) + h_1(\mu_2) \\
\leq& \langle \frac{\partial - h_2^*(-A\mu_2)}{\partial - A\mu_2}, -A\mu_1 - (-A\mu_2) \rangle + \frac{1/\alpha_{h_2}}{2}\| - A\mu_1 - (-A\mu_2)\|^2 \\
& + \langle -\nabla h_1(\mu_2), \mu_1 - \mu_2 \rangle + \frac{l_{h_1,1}}{2}\|\mu_1 - \mu_2\|^2 \\
\leq& \langle \nabla D(\mu_2), \mu_1 - \mu_2 \rangle + \frac{l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}}{2}\|\mu_1 - \mu_2\|^2.
\end{aligned}
$$

The first inequality holds as both h_2^* and h_1 are smooth. The second follows the chain rule.

Note $\sigma_{\max}(A) \geq \sigma_{\min}(A) > 0$ as A is full column rank. This completes the proof. $\qquad\square$

In this way, we are ready to proceed with the general convergence analysis to solve (39) as $L(\mu, y)$ using Algorithm 2 without acceleration and $T_y = 1$, which is a single-loop algorithm.

Proof of Theorem 7. We first look into the update of $\|y_t - \nabla h_2^*(-A\mu_t)\|$.

Fixing μ, define $y_\mu^* := \arg\min_y L(\mu, y)$. The first-order stationary optimality condition requires $\nabla_y L(\mu, y_\mu^*) = 0$, i.e. $\nabla h_2(y_\mu^*) = -A\mu_t$. This implies $y_\mu^* = \nabla h_2^*(-A\mu_t)$ because the mapping ∇h_2 and ∇h_2^* are the inverse of each other according to Lemma 4.

For a fixed μ_t, the update rule $y_{t+1} = y_t - \eta_1 \nabla_y L(\mu_t, y_t)$ is a gradient descent step for the objective function $L(\mu_t, y)$, which is also α_{h_2}-strongly convex and $l_{h_2,1}$-smooth to y. Following Lemma 7, take $\eta_1 \leq \frac{1}{l_{h_2,1}}$, we have

$$\|y_{t+1} - \nabla h_2^*(-A\mu_t)\| \leq (1 - \eta_1 \alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\| \tag{40}$$

23

Following triangle inequality, we also have

$$
\begin{aligned}
&\|y_{t+1} - \nabla h_2^*(-A\mu_{t+1})\| \\
\leq& \|y_{t+1} - \nabla h_2^*(-A\mu_t)\| + \|\nabla h_2^*(-A\mu_t) - \nabla h_2^*(-A\mu_{t+1})\| \\
\leq& (1 - \eta_1\alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\| + \frac{\sigma_{\max}(A)}{\alpha_{h_2}}\|\mu_{t+1} - \mu_t\|
\end{aligned}
\tag{41}
$$

where the second term in the last inequality comes from the smoothness of the conjugate function according to Lemma 4.

We then look into the update of $\|\mu_{t+1} - \mu_t\|$. According to Lemma 12,

$$
D(\mu) := \min_{y \in \mathbb{R}^{d_y}} -h_1(\mu) + y^\top A\mu + h_2(y) = -h_1(\mu) - h_2^*(-A\mu)
$$

is $\frac{\sigma_{\min}^2(A)}{l_{h_2,1}}$-strongly concave and $(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})$-smooth with respect to μ. Moreover, the problem

$$
\max_{\mu \in \mathbb{R}_+^{d_c}} D(\mu),
$$

is equivalent to the unconstrained problem with the Lagrange multiplier

$$
\max_{\mu \in \mathbb{R}^{d_c}} \tilde{D}(\mu) := D(\mu) + \lambda^\top \mu
$$

where unique λ is non-negative and finite in all dimension, i.e. $0 \leq \lambda < \infty$, as $D(\mu)$ is strongly convex and $\mu \in \mathbb{R}_+^{d_c}$ is equivalent to $\mu \geq 0$. We know that $\tilde{D}(\mu)$ is smooth and strongly concave with the same modulus as $D(\mu)$. The first-order stationary condition requires

$$
\nabla \tilde{D}(\mu^*) = \nabla D(\mu^*) + \lambda = 0.
\tag{42}
$$

In this way,

$$
\begin{aligned}
\frac{1}{\eta_2}\|\mu_{t+1} - \mu_t\| &= \frac{1}{\eta_2}\|[\mu_t + \eta_2(-\nabla h_1(\mu_t) + A^\top y_{t+1})]_{\mathbb{R}_+^{d_c}} - \mu_t\| \\
&\overset{(a)}{\leq} \|-\nabla h_1(\mu_t) + A^\top y_{t+1}\| \\
&= \|-\nabla h_1(\mu_t) + A^\top \nabla h_2^*(-A\mu_t) + \lambda + A^\top y_{t+1} - A^\top \nabla h_2^*(-A\mu_t) - \lambda\| \\
&\overset{(b)}{\leq} \|\nabla \tilde{D}(\mu_t)\| + \sigma_{\max}(A)\|y_{t+1} - \nabla h_2^*(-A\mu_t)\| + \|\lambda\| \\
&\overset{(c)}{\leq} \|\nabla \tilde{D}(\mu_t) - \nabla \tilde{D}(\mu^*)\| + \sigma_{\max}(A)(1 - \eta_1\alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\| + \|\lambda\| \\
&\overset{(d)}{\leq} (l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})\|\mu_t - \mu^*\| + \sigma_{\max}(A)(1 - \eta_1\alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\| + \|\lambda\|
\end{aligned}
\tag{43}
$$

Inequality (a) comes from the non-expansiveness (1-Lipschitzness) of the projection operation, (b) follows triangle inequality, (c) uses (42) and (40), and (d) comes from the smoothness of $\tilde{D}(\mu)$.

Now we are ready to find the bound of the update of $\|\mu_t - \mu^*\|$.

Define an auxiliary update as

$$
\tilde{\mu}_{t+1} := [\mu_t + \eta_2 \nabla D(\mu_t)]_{\mathbb{R}_+^{d_c}} = [\mu_t + \eta_2(-\nabla h_1(\mu_t) + A^\top \nabla h_2^*(-A\mu_t))]_{\mathbb{R}_+^{d_c}}.
\tag{44}
$$

This is a projected gradient descent on strongly convex $-D(\mu)$. As $\mathbb{R}_+^{d_c}$ is closed and convex, following Lemma 7, for $\eta_2 \leq \frac{1}{(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})}$, we have

$$
\|\tilde{\mu}_{t+1} - \mu^*\| \leq \left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}}\right)\|\mu_t - \mu^*\|.
$$

As the real update is $\mu_{t+1} = [\mu_t + \eta_2(-\nabla h_1(\mu_t) + A^\top y_t)]_{\mathbb{R}_+^{d_c}}$, by the non-expansiveness (1-Lipschitzness) of projection operation, we have

$$
\|\tilde{\mu}_{t+1} - \mu_{t+1}\| \leq \|\eta_2 A^\top(y_{t+1} - \nabla h_2^*(-A\mu_t))\| \leq \eta_2\sigma_{\max}(A)\|y_{t+1} - \nabla h_2^*(-A\mu_t)\|
$$

24

774 By triangle inequality and (40), we have

$$\|\mu_{t+1} - \mu^*\| \le \left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}}\right) \|\mu_t - \mu^*\| + \eta_2 \sigma_{\max}(A)\|y_{t+1} - \nabla h_2^*(-A\mu_t)\|$$

$$\le \left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}}\right) \|\mu_t - \mu^*\| + \eta_2 \sigma_{\max}(A)(1 - \eta_1 \alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\|. \tag{45}$$

775 For some positive $\rho > 0$, denote

$$P_t := \rho\|\mu_t - \mu^*\| + \|y_t - \nabla h_2^*(-A\mu_t)\|. \tag{46}$$

776 We know from (41), (43), and (45) that

$$P_{t+1} = \rho\|\mu_{t+1} - \mu^*\| + \|y_{t+1} - \nabla h_2^*(-A\mu_{t+1})\|$$

$$\le \rho\left(\left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}}\right) \|\mu_t - \mu^*\| + \eta_2 \sigma_{\max}(A)(1 - \eta_1 \alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\|\right)$$

$$+ (1 - \eta_1 \alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\|$$

$$+ \frac{\sigma_{\max}(A)}{\alpha_{h_2}}\eta_2\left((l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})\|\mu_t - \mu^*\| + \sigma_{\max}(A)(1 - \eta_1\alpha_{h_2}/2)\|y_t - \nabla h_2^*(-A\mu_t)\| + \|\lambda\|\right)$$

$$\le \left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}} + \frac{1}{\rho}\frac{\sigma_{\max}(A)}{\alpha_{h_2}}\eta_2(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})\right)\rho\|\mu_t - \mu^*\|$$

$$+ (1 - \eta_1\alpha_{h_2}/2)\left(1 + \rho\eta_2\sigma_{\max}(A) + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}\eta_2\right)\|y_t - \nabla h_2^*(-A\mu_t)\| + \frac{\sigma_{\max}(A)}{\alpha_{h_2}}\eta_2\|\lambda\|.$$

777 To construct $P_{t+1} \le (1-c)P_t + \frac{\sigma_{\max}(A)}{\alpha_{h_2}}\eta_2\|\lambda\|$ for some $0 < c < 1$, it is sufficient to find
778 $\eta_1 \le \frac{1}{l_{h_2,1}}, \eta_2 \le \frac{1}{(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})}$, and $\rho > 0$ such that

$$\begin{cases} 0 < \left(1 - \eta_2 \frac{\sigma_{\min}^2(A)}{2l_{h_2,1}} + \frac{1}{\rho}\frac{\sigma_{\max}(A)}{\alpha_{h_2}}\eta_2(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})\right) \le 1 - \eta_2 \frac{\sigma_{\min}^2(A)}{4l_{h_2,1}} < 1 \\ 0 < (1 - \eta_1\alpha_{h_2}/2)\left(1 + \rho\eta_2\sigma_{\max}(A) + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}\eta_2\right) \le (1 - \eta_1\alpha_{h_2}/2)(1 + \eta_1\alpha_{h_2}/2) < 1 \end{cases}$$

779 This can be obtained when

$$\begin{cases} \rho \ge \frac{4l_{h_2,1}\sigma_{\max}(A)}{\alpha_{h_2}\sigma_{\min}^2(A)}(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}) \\ \eta_2 \le \frac{\eta_1\alpha_{h_2}}{2\left(\rho\sigma_{\max}(A) + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}\right)} \end{cases} \tag{47}$$

780 (47) can be obtained when $\epsilon > 0$ is sufficiently such that $\rho = \epsilon^{-1} \ge \frac{4l_{h_2,1}\sigma_{\max}(A)}{\alpha_{h_2}\sigma_{\min}^2(A)}(l_{h_1,1} + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}})$,
781 $\eta_1 = O(\frac{1}{l_{h_2,1}})$ and $\eta_2 = O(\frac{\alpha_{h_2}}{l_{h_2,1}}\rho^{-1}) = O(\epsilon^{-1})$. In this way,

$$P_{t+1} \le (1-c)P_t + O(\alpha_{h_2}^{-1}\epsilon)$$

782 where $c > 0$ is of the order $O(\epsilon)$. Iteration gives

$$P_t \le (1-c)^t P_0 + O(\alpha_{h_2}^{-1}). \tag{48}$$

783 Notice $P_0 = O(\epsilon^{-1})$ as $\rho = \epsilon^{-1}$. In this way, there exist $T_1 = O(\log(\epsilon^{-1}))$ such that for all $t > T$,
784 $(1-c)^t P_0 = O(1)$ and accordingly

$$P_t = O(1), \quad \forall t > T_1. \tag{49}$$

785 Moreover, as $P_t = \epsilon^{-1}\|\mu_t - \mu^*\| + \|y_t - \nabla h_2^*(-A\mu_t)\|$,

$$\|\mu_t - \mu^*\| \le \epsilon P_t = O(\epsilon), \quad \forall t > T_1. \tag{50}$$

786 Furthermore, choose $\eta_1 = O(\frac{1}{l_{h_2,1}})$ satisfying $\eta_1 \le \frac{1}{l_{h_2,1}}$, for $t > T_1$,

$$\|y_t - \nabla h_2^*(-A\mu_t)\| \le (1 - \eta_1\alpha_{h_2}/2)^{t-T_1}\|y_{T_1} - \nabla h_2^*(-A\mu_{T_1})\| + O(\epsilon). \tag{51}$$

25

Here $\eta_1 \alpha_{h_2}/2 = O(\frac{\alpha_{h_2}}{l_{h_2,1}}) = O(1)$. In this way, for another $T_2 = O(\log(\epsilon^{-1}))$ steps, we have:

$$\|y_t - \nabla h_2^*(-A\mu_t)\| = O(\epsilon), \tag{52}$$

$$\text{and} \quad \|y_t - y^*\| \leq \|y_t - \nabla h_2^*(-A\mu_t)\| + \|\nabla h_2^*(-A\mu_t) - \nabla h_2^*(-A\mu^*)\| \tag{53}$$

$$\leq \|y_t - \nabla h_2^*(-A\mu_t)\| + \frac{\sigma_{\max}^2(A)}{\alpha_{h_2}}\|\mu_t - \mu^*\| \tag{54}$$

$$= O\left(\epsilon + \alpha_{h_2}^{-1}\epsilon\right) = O\left(\epsilon\right) \tag{55}$$

we can see that the algorithm converges linearly with complexity $O(T_1 + T_2) = O(\log(\epsilon^{-1}))$. In this way, obtaining

$$\|y_T - y^*\|^2 < \epsilon \quad \text{and} \quad \|\mu_T - \mu^*\|^2 < \epsilon, \tag{56}$$

requires complexity $T = O(\log((\sqrt{\epsilon})^{-1})) = O(\log(\epsilon^{-1}))$. This completes the proof. $\qquad\square$

E   Applications to SVM model training

In this section, we provide further details about the SVM model training experiment for the linear SVM model, including the problem formulation, and detailed results analysis.

E.1   Problem introduction

The SVM is a supervised machine learning model used for classification and regression tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. For the hard-margin SVM, misclassification is not tolerated. For the soft-margin SVM, the violation of classification, ξ, is penalized to the training objective to consider misclassification. To train an efficient soft-margin linear SVM, we are interested in the following constraint BLO problem

$$\min_c \mathcal{L}_{\mathcal{D}_{val}}(w^*, b^*) = \sum_{(z_{val,i}, l_{val,i}) \in \mathcal{D}_{val}} \exp\left(1 - l_{val,i}\left(z_{val,i}^\top w^* + b^*\right)\right) \tag{57a}$$

$$\text{with } w^*, b^*, \xi^* = \arg\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{2}\|c\|^2 \tag{57b}$$

$$\text{s.t. } l_{tr,i}(z_{tr,i}^\top w + b) \geq 1 - \xi_i \quad \forall\, i \in \{1, \ldots, |\mathcal{D}_{tr}|\} \tag{57c}$$

$$\xi_i \leq c_i \quad\quad\quad\quad\quad \forall\, i \in \{1, \ldots, |\mathcal{D}_{tr}|\}. \tag{57d}$$

The upper-level objective is a validation loss, and the lower level is to train SVM on the training set $\mathcal{D}_{tr} := \{(z_{tr,i}, l_{tr,i})\}_{i=1}^{|\mathcal{D}_{tr}|}$ with the soft margin upper bounded by c. The lower-level objective function considers both maximizing the margin (minimizing $\|w\|^2$) and allowing violations to the separating hyperplane ξ, controlled by the hyperparameter (and upper-level variable) c. The idea behind the BLO formulation is to use the validation loss (upper-level objective) to tune the hyperparameter c, while the model parameters (lower-level variables) should be optimal in the training dataset.

E.2   Additional Experiments

In this section, we present the detailed experimental results for the SVM model training experiment using our BLOCC algorithm in comparison with two baselines, LV-HBA [69] and GAM [67], both are tailored for BLO problems with inequality coupled constraints.

We evaluate the proposed algorithms in two different datasets: diabetes [18] and fourclass [27]. The detailed results are illustrated in Figure 4, where we represent validation metrics in the left column and test metrics in the right column. The metrics include both loss and accuracy, for both the diabetes and fourclass datasets. Our algorithm is able to converge faster both in terms of accuracy and loss, and it achieves a lower loss value than the alternatives for both datasets in both validation and test.

VT: Simply to showcase different things. They can be unified if we consider so.

26

(a) Validation loss for the diabetes dataset.

(b) Test loss for the diabetes dataset.

(c) Validation accuracy for the diabetes dataset.

(d) Test accuracy for the diabetes dataset.

(e) Validation loss for the fourclass dataset.

(f) Test loss for the fourclass dataset.
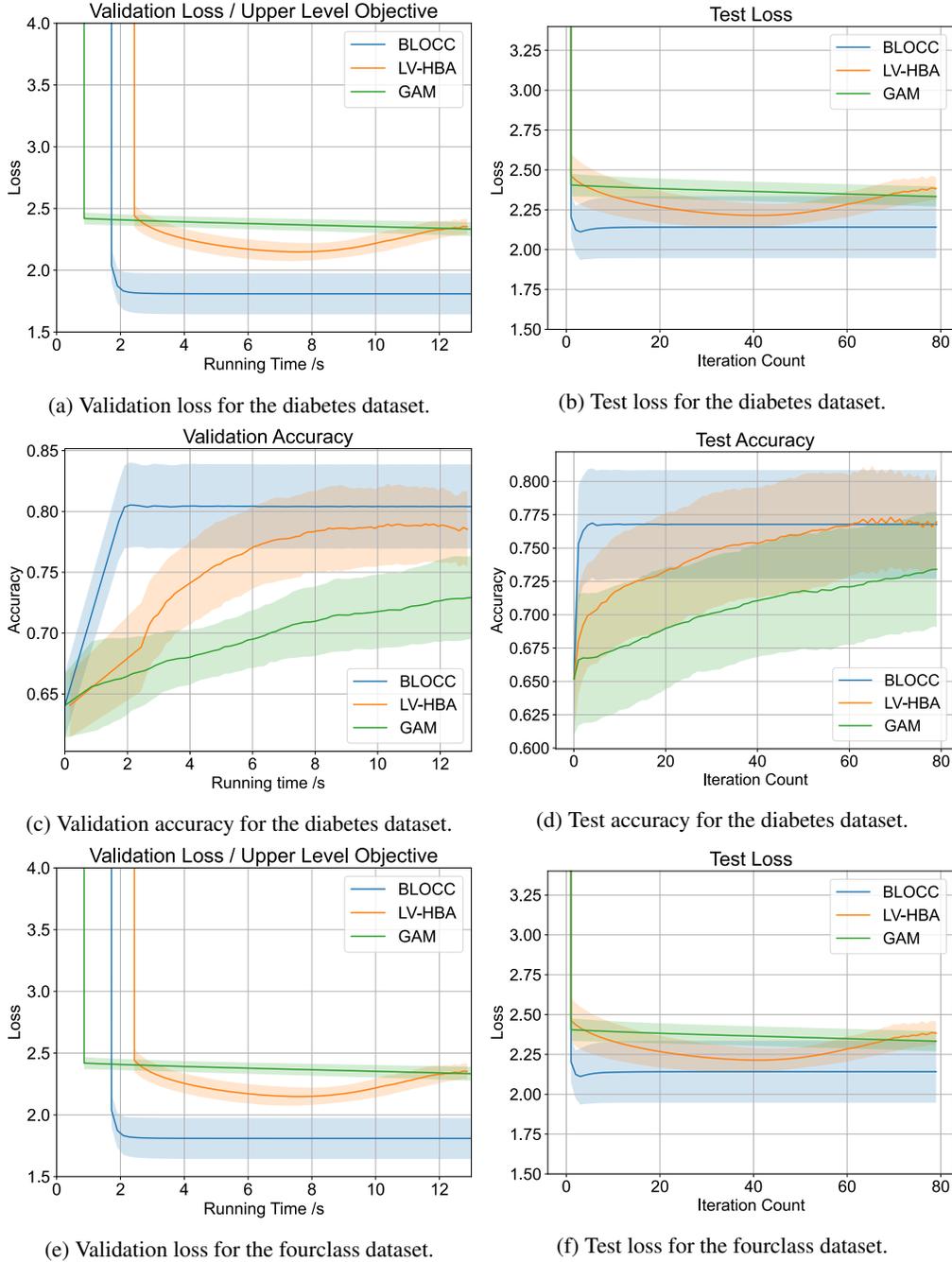
Figure 4: Results of the hyperparameter optimization experiment with an SVM model.

F   Applications to Transportation Network Planning

This section delves into applying the proposed algorithm to tackle a practical BLO problem in transportation science.

F.1   Problem introduction

In this transportation network planning problem, we are to design a capacitated transportation network connecting a set $\mathcal{S}$ of stations [9]. The network is designed to carry out passengers from a given

27

origin $o \in \mathcal{S}$ to a given destination $d \in \mathcal{S}$. For all the potential (o, d)-markets, we know the estimated demand of passengers, who can choose rationally among $\mathcal{N}$ networks including our designed network and the ones of our competitors', considering their trip (dis)utility function that depends on factors such as the price or the trip time [9].

Our goal is to maximize the operator's benefit modeled by a utility function (upper-level objective) knowing the passengers make the rational decisions on choosing the best route (lower-level objective) considering the available options (coupled constraints). Specifically, the network designer must determine the capacity for each link $(i, j) \in \mathcal{A}$, where $\mathcal{A}$ denotes the proposed network topology ($\mathcal{A} \subseteq \mathcal{S} \times \mathcal{S}$). The set $\mathcal{K}$ defines the markets between various origins o and destinations d, such that $(o, d) \in \mathcal{K}$ and $\mathcal{K} \subseteq \mathcal{S} \times \mathcal{S}$. The problem entails the joint optimization of construction and flow variables, which are described below.

- $x_{ij} \in \mathbb{R}^+$: capacity constructed for the link $(i, j) \in \mathcal{A}$.

- $y^{od} \in [0, 1]$: proportion of passengers from market $(o, d) \in \mathcal{K}$ choosing the new network for their travel. As we just consider 2 networks, the proportion of passengers choosing the incumbent network is $1 - y^{od}$.

- $y_{ij}^{od} \in [0, 1]$: proportion of passengers from market $(o, d) \in \mathcal{K}$ using link $(i, j) \in \mathcal{A}$ for their travel.

To be consistent with the rest of the manuscript, we use x to denote a tuple collecting all the construction variables so that x represents the set of variables in the upper-level (associated with the decisions of the operator). The feasible domain for x is $\mathcal{X} = \mathbb{R}^{+|\mathcal{A}|}$. Analogously, we use y to denote a tuple collecting all the flow variables. Tuple y represents the variables for the lower-level (associated with the passengers' decisions). The feasible domain for y is $\mathcal{Y} = [0, 1]^{|\mathcal{A}|} \times [0, 1]^{|\mathcal{A}||\mathcal{K}|}$.

In addition to the optimization variables, our objective and constraints involve other state variables and parameters:

- w^{od}: total estimated demand (number of passengers) between nodes $(o, d) \in \mathcal{K}$.

- m^{od}: revenue obtained by the operator from a passenger in the market $(o, d) \in \mathcal{K}$.

- c_{ij}: construction cost per passenger associated with link $(i, j) \in \mathcal{A}$.

- t_{ij}: travel time for the link $(i, j) \in \mathcal{A}$.

- t_{ext}^{od}: travel time on the alternative network for passengers in the market $(o, d) \in \mathcal{K}$.

- ω_t: coefficient associated with the travel time on passengers' utility function.

Now we are ready to introduce the objectives of our BLO problem. The network operator aims at maximizing profits and minimizing costs. As a result we have that the objective to minimize is

$$\min_{x, y} f(x, y) := -\left(\sum_{\forall (o, d) \in \mathcal{K}} m^{od} y^{od*}(x) - \sum_{\forall (i, j) \in \mathcal{A}} c_{ij} x_{ij} \right), \tag{58}$$

where $y^{od*}(x)$ are the passenger flows associated with the network design x. Regarding the lower level, for each transportation alternative and market, passengers aim to minimize the function

$$\min_{x, y} g(x, y) := w^{od} y^{od} (\log(y^{od}) - 1) + \sum_{(i, j) \in \mathcal{A}} w^{od} \omega_t t_{ij} y_{ij}^{od} \tag{59}$$

The second term represents the passenger's disutility. The role of the negative entropy in the first term is to ensure that decisions are made according to the so-called logit model [4]-[10]. This model states that the probability that a passenger selects network $n \in \mathcal{N}$ for market (o, d) is determined by the logit distribution:

$$P(n|(o, d)) = \frac{e^{-u_n^{od}}}{\sum_{n' \in \mathcal{N}} e^{-u_{n'}^{od}}}, \tag{60}$$

where u_n^{od} represents the disutility of the best available path within network $n \in \mathcal{N}$ for market $(o, d) \in \mathcal{K}$. For this study, we assume a scenario where: i) the disutility is given by the multiplication

28

of the sensitivity parameter ω_t and the travel time and ii) only one incumbent network exists, thus $\mathcal{N}$ consists of this incumbent network and the network under construction. Interestingly, it can be rigorously shown that the Karush-Kuhn-Tucker (KKT) conditions associated with of (59) lead to the expression in (60); see a formal proof of this result in [51].

With these considerations in mind, we are ready to formulate our constrained BLO problem

$$\min_{x \in \mathcal{X}} \quad - \sum_{\forall (o,d) \in \mathcal{K}} m^{od} y^{od*} + \sum_{\forall (i,j) \in \mathcal{A}} c_{ij} x_{ij} \tag{61a}$$

s.t.

$$(y^{od*}, y_{ij}^{od*}) = \arg\min_{y \in \mathcal{Y}} \sum_{(o,d) \in \mathcal{K}} w^{od} y^{od} (\log(y^{od}) - 1) \tag{61b}$$

$$+ \sum_{(o,d) \in \mathcal{K}} w^{od} (1 - y^{od}) (\log(1 - y^{od}) - 1)$$

$$+ \sum_{(o,d) \in \mathcal{K}} \sum_{(i,j) \in \mathcal{A}} w^{od} \omega_t t_{ij} y_{ij}^{od} + \sum_{(o,d) \in \mathcal{K}} w^{od} \omega_t t_{ext}^{od} (1 - y^{od})$$

s.t.

$$\sum_{\forall (o,d) \in \mathcal{K}} w^{od} y_{ij}^{od} \leq x_{ij} \qquad \forall (i,j) \in \mathcal{A} \tag{61c}$$

$$\sum_{\forall j | (i,j) \in \mathcal{A}} y_{ij}^{od} - \sum_{\forall j | (j,i) \in \mathcal{A}} y_{ji}^{od} = \begin{cases} y^{od} & \text{if } i = o \\ -y^{od} & \text{if } i = d \\ 0 & \text{otherwise} \end{cases} \qquad \forall i, (o,d) \in \mathcal{S} \times \mathcal{K}, \tag{61d}$$

where (61a) is the (operator's) upper-level objective and (61b) is the (passengers') lower-level objective. Note that for the lower-level objective, we aggregated the terms in (59) for all markets in $\mathcal{K}$ and the new and the alternative network, with the latter absorbing a fraction $(1 - y^{od})$ of the demand.

We shift now attention to the constraints. The capacity constraint in (61c) is critical for our approach since it relates to the upper and lower-level variables. Notice that we have one constraint per link and, in each of them $|\mathcal{K}|$ lower-level variables are involved. This implies that, even for medium-size networks (with tens or hundreds of nodes), thousands of coupled constraints, each with thousands of variables, will be present. In addition, (61d) represents flow conservation constraints: for every market $(o,d) \in \mathcal{K}$, these constraints ensure that the total flow departing from the origin o equals the total flow for that market. Similarly, the total flow entering destination d matches the flow leaving the origin. For nodes that are neither the origin nor the destination of the market, the flow conservation must be zero. The number of these constraints, which only involve lower-level variables, is $|\mathcal{S}||\mathcal{K}|$, scaling as a third-order polynomial with the number of nodes.

F.2 Experiment roadmap

In order to provide numerical results illustrating the behavior of our algorithm, we solve this optimization problem in three scenarios:

1. the design of a 3-node simple synthetic network;

2. the design of a 9-node synthetic network that has been previously analyzed in the transportation literature; and,

3. the design of a (real-world) subway network for the city of Seville, Spain, with 24 nodes.

In the case of the 3-node network, we will conduct a comparative analysis against other algorithms to evaluate the efficacy of our approach. Moving on to the 9-node and Seville networks, we will provide insights into the performance and behavior of our algorithm under varying parameters, shedding light on the versatility and adaptability of our approach to real-world transportation networks.

While one of the goals of these experiments was to compare our BLOCC algorithm against LV-HBA [69] and GAM [67], for the scenario at hand, the GAM algorithm cannot be implemented, since the

29
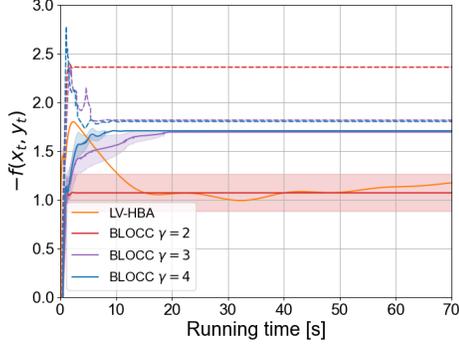
Figure 5: Negative upper-level objective $-f(x_t, y_t)$ evolution over time for a 3-node network design problem for 10 random initializations of the upper-level variables. The solid lines represent the mean value of $-f(x_t, y_{g,t}^{T_g})$ of the 10 realizations, and the shaded region is the $\pm$ standard deviation. The dashed lines represent the mean value of $-f(x_t, y_{F,t}^{T_F})$ of the 10 realizations, and the shaded region is the $\pm$ standard deviation. Three different γ values (red, purple, blue) are represented in our algorithm, and fixed stepsize $\eta = 1.6e-4$. The orange color represents the evolution of $-f(x_t, y_t)$ for the LV-HBA algorithm.
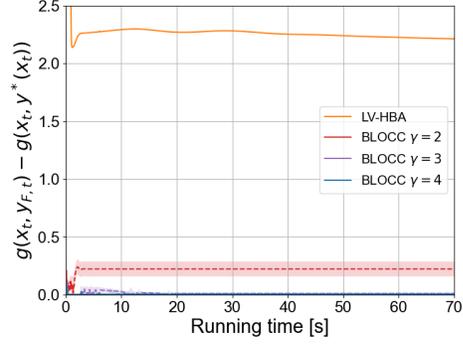
Figure 6: Loss of optimality on the lower-level optimization problem over time for a 3-node network design problem $g(x_t, y_t) - g(x_t, y_t^*)$, for 10 random initializations of the upper-level variables. The solid lines represent the mean value of the 10 realizations, the dashed lines represent performance of $g(x_t, y_{F,t}^{T_F}) - g(x_t, y_t^*)$, and the shaded region is the $\pm$ standard deviation. Three different γ values (red, purple, blue) are represented in our algorithm, and fixed stepsize $\eta = 1.6e-4$. The orange color represents the loss of optimality on the lower-level problem $g(x_t, y_t) - g(x_t, y_t^*)$ for the LV-HBA algorithm.

inverse of a matrix at each iteration for the problem in (61) is not tractable. In this way, we only conducted the experiments using our BLOCC and LV-HBA.

Moreover, besides the fact that Theorem 1 and Theorem 2 guarantees that $(x_T, y_{F,T}^{T_F})$ can be the solution to the ϵ-approximation problem, using $y_{g,T}^{T_g}$ as output can better attain the lower-level minimum $y_g^*(x_T)$ as it solves (7). In this way, we will presents both output using $y_{F,T}^{T_F}$ and $y_{g,T}^{T_g}$.

F.3 A 3-node network experiment

In this section, we solve the problem formulated in (61) using a network with 3 nodes, 6 potential links, and 6 markets. The state variable values for the simulation scenarios are available in the code repository. Figure 5 illustrates the performance of both algorithms, showing computation time on the horizontal axis and the upper-level objective value $-f(x_t, y_t)$ on the vertical axis for 10 random initializations. We analyze three instances of BLOCC with $\gamma \in 2, 3, 4$ and a stepsize of $\eta = 1.6e-4$. The upper-level objective values are computed using $-f(x_t, y_{g,t}^{T_g})$ and $-f(x_t, y_{F,t}^{T_F})$.

Our algorithm converges to the local optimum faster than LV-HBA, which fails to reach this optimum within the given time limit, resulting in a solution that does not satisfy the lower-level optimality constraint (61b).

To better understand the results, we compare the main differences between BLOCC and LV-HBA:

1. In BLOCC, either using output $y_{g,T}^{T_g}$ or $y_{F,T}^{T_F}$ has a guarantee of attaining optimality at the lower level, whereas in LV-HBA, the lower-level optimality can not be guaranteed, as shown in Figure 6

2. As already mentioned, the LV-HBA algorithm requires a joint projection into $\{\mathcal{X} \times \mathcal{Y} : g^c(x, y) \leq 0\}$ at each iteration, so when there are a large number of upper variables (99 in the presented scenario) and also a large number of constraints (24 in this simplified scenario), the computational time required for this projection increases considerably. In contrast, in BLOCC, it is only necessary to project onto $\mathcal{X}$ at each iteration, which simply represents box constraints and projection is straightforward.

30
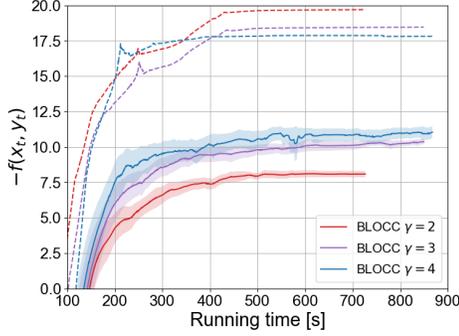
Figure 7: Negative upper-level objective $-f(x_t, y_t)$ evolution over time for a 9-node network design problem for 10 random initializations of the upper-level variables. The solid lines represent the mean value of $-f(x_t, y_{g,t}^{T_g})$ of the 10 realizations, and the shaded region is the $\pm$ standard deviation. The dashed lines represent the mean value of $-f(x_t, y_{F,t}^{T_F})$ of the 10 realizations, and the shaded region is the $\pm$ standard deviation. Three different γ values (red, purple, blue) are represented in our algorithm, and fixed stepsize $\eta = 1.6e - 4$.
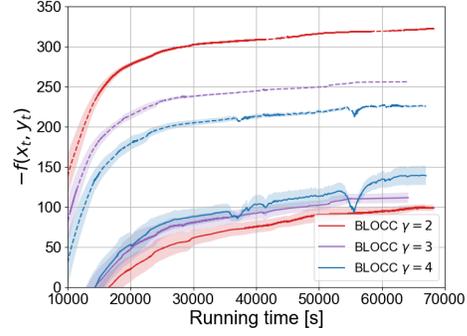
Figure 8: Negative upper-level objective $-f(x_t, y_t)$ evolution over time for a metro network design problem in the city of Seville, Spain, for 2 different random initializations of the upper-level variables. The solid lines represent the mean value of $-f(x_t, y_{g,t}^{T_g})$ of the 2 realizations, and the shaded region is the $\pm$ standard deviation. The dashed lines represent the mean value of $-f(x_t, y_{F,t}^{T_F})$ of the 2 realizations, and the shaded region is the $\pm$ standard deviation. Three different γ values (red, purple, blue) are represented in our algorithm, and fixed stepsize $\eta = 1.6e - 4$.

F.4 A 9-node network experiment

In this case, we consider a network with $|\mathcal{S}| = 9$ nodes and $|\mathcal{A}| = 72$ potential links, as well as $|\mathcal{K}| = 72$ markets. Figure 7 presents the obtained results for three different values of parameter $\gamma \in \{2, 3, 4\}$, and stepsize $\eta = 1.6e - 4$. It depicts computational time on the horizontal axis, while the evolution of $-f(x_t, y_{g,t}^{T_g})$ is provided on the vertical axis for 10 different random initializations of the upper-level variables. As in the previous network, BLOCC algorithm is able to converge, with different γ values leading to different optimums.

As mentioned in the main paper (reference main paper section), this parameter influences on the accuracy achieved regarding optimality at the lower-level for the variable y_F. For higher values of γ, the optimality condition at the lower-level when solving the problem associated with (reference algorithm 2) becomes more important in the objective function. Thus, the difference between $y_{g,t}^{T_g}$ and $y_{F,t}^{T_F}$ decreases for higher values of γ. Additionally, it can be seen how the best objective value $-f(x_t, y_{g,T}^{T_g}(x))$ is achieved for the highest value of γ, as well as the accuracy on the optimality of the lower-level problem for the solution of $y_{F,t}^{T_F}$ increases.

F.5 Seville network experiment

In this section, we aim to demonstrate the validity of BLOCC by applying it to a real transportation network design problem. Specifically, we address the design of a potential metro network in the city of Seville. This network consists of $|\mathcal{S}| = 24$ nodes and 552 possible links. However, we filter the set of possible links according to two criteria:

1. The link between nodes $(i, j) \in \mathcal{S} \times \mathcal{S}$ can only exist if node j is one of the 3 closest neighbors to i, or vice versa, in terms of travel time.

2. The link between nodes $(i, j) \in \mathcal{S} \times \mathcal{S}$ can only exist if the travel time t_{ij} is less than 7 minutes.

Thus, the set of possible links is reduced to $|\mathcal{A}| = 88$ possible links. The proposed topology for the network is shown in Figure 9. We consider all possible markets between nodes, so $|\mathcal{K}| = 552$.

31

Following the narrative in Sections F.3 and F.4, Figure 8 presents the evolution of the upper-level objective function with time for values of the parameter $\gamma \in \{2, 3, 4\}$ for 2 different realizations. As it can be observed, higher values of $-f(x_T, y_{T,g}^{T_g})$ are obtained for higher values of γ, as well as smaller gaps between $f(x_T, y_{T,g}^{T_g})$ and $-f(x_T, y_{T,F}^{T_F})$. In summary, it is demonstrated that the algorithm formulated in this document is able to solve problems with a large number of variables, which can have practical value in real-world applications, such as the one studied in this section.
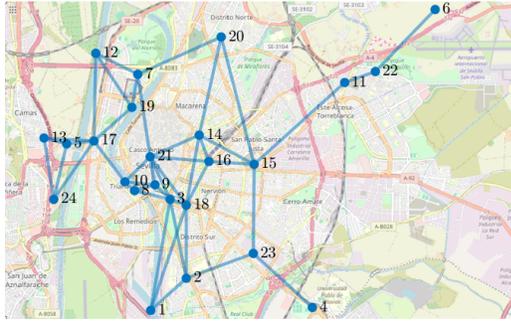
Figure 9: Topology of the Seville network.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: yes, we have discussed in the conclusions.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All have been clearly listed.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The supplementary materials contain the code used to run the simulations.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: the supplementary materials contain all the code used to run the simulation, as well as the data files considered.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: all the necessary details to run the experiments are available in the code.

32

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: the figures in the paper include confidence intervals representing the standard deviation for different realizations of the experiments. Also the tables include the standard deviation of the simulations.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: the paper is theory paper, and does not include much computation-heavy experiments

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: we follow the NeurIPS Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: the work is foundational research, and there is no societal impact of the work performed.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: the paper does not use existing assets.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

33

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

34