A Barrier Function Approach for Bilevel Optimization with Coupled Lower-Level Constraints: Formulation, Approximation and Algorithms

Xiaotian Jiang, Jiaxiang Li, Mingyi Hong, Shuzhong Zhang*

January 20, 2025

Abstract

In this paper, we consider bilevel optimization problem where the lower-level has coupled constraints, i.e. the constraints depend both on the upper- and lower-level variables. In particular, we consider two settings for the lower-level problem. The first is when the objective is strongly convex and the constraints are convex with respect to the lower-level variable; The second is when the lower-level is a linear program. We propose to utilize a barrier function reformulation to translate the problem into an unconstrained problem. By developing a series of new techniques, we proved that both the hyperfunction value and hypergradient of the barrier reformulated problem (uniformly) converge to those of the original problem under minimal assumptions. Further, to overcome the non-Lipschitz smoothness of hyperfunction and lower-level problem for barrier reformulated problems, we design an adaptive algorithm that ensures a non-asymptotic convergence guarantee. We also design an algorithm that converges to the stationary point of the original problem asymptotically under certain assumptions. The proposed algorithms require minimal assumptions, and to our knowledge, they are the first with convergence guarantees when the lower-level problem is a linear program. Numerical experiments are conducted to show the effectiveness of the proposed method.

1 Introduction

Bilevel optimization (BLO) is drawing wide attention in the communities including machine learning, operations research and signal processing. It observes wide applications in various machine learning problems, such as hyperparameter optimization (Maclaurin et al., 2015b; Franceschi et al., 2018), meta learning (Finn et al., 2017; Franceschi et al., 2018; Ji et al., 2021) and reinforcement learning (Stadie et al., 2020; Zeng et al., 2024; Li et al., 2024a,b). In the field of operations research, BLO finds its applications in pricing, transportation design, game theory, among others (Labbé and Violin, 2016; Silvério et al., 2022). Standard BLO takes the form:

$$\min_{x, y^*(x)} \ f(x, y^*(x)) \quad \text{s.t.} \quad x \in \mathcal{X}, \ y^*(x) \in \operatorname*{argmin}_{y \in \mathcal{Y}(x)} g(x, y), \tag{1.1}$$

where f is called the upper-level objective and g is called the lower-level objective; $\mathcal{X} \subseteq \mathbb{R}^n$, and $\mathcal{Y}(x) \subseteq \mathbb{R}^m$ are feasible sets for the upper- and lower-problems, respectively. In game theory, the

---

*X. Jiang and S. Zhang are with Department of Industrial and System Engineering, University of Minnesota, Minneapolis, MN, USA. E-mails: `jian0851@umn.edu`, `zhangs@umn.edu`. J. Li and M. Hong are with Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. E-mails: `li003755@umn.edu`, `mhong@umn.edu`.

bilevel problem can be thought of as a two-player (Stackelberg) game, the lower-level and upper-level problems are also called the follower and the leader, respectively; see e.g. Kohli (2012); Liu et al. (2018).

In BLO with *unconstrained* lower-level problems where $\mathcal{Y} \equiv \mathbb{R}^m$, a popular class of algorithms are referred to as the *implicit gradient descent* method (Ghadimi and Wang, 2018; Yang et al., 2021; Hong et al., 2023; Pedregosa, 2016). Let us denote the *hyperfunction* ϕ as

$$\phi(x) := \min_{y^*(x)} f(x, y^*(x)), \text{ where } y^*(x) \in \operatorname*{argmin}_{y \in \mathbb{R}^m} g(x, y). \tag{1.2}$$

The *implicit gradient descent* method operates by using the implicit function theorem to compute (some approximated version of) the gradient of $\phi(x)$, which we refer to as the hypergradient $\nabla_x \phi(x)$[1], thus allowing us to apply gradient descent-type of methods to $\phi(x)$. Indeed, as we will discuss shortly, there has been tremendous recent success in developing efficient implicit gradient decent-type methods and analytical approaches for *unconstrained* problems where $\mathcal{Y}(x) \equiv \mathbb{R}^m$. However, it is still unclear how to effectively leverage these developments for problems with lower-level constraints. In particular, the constrained lower-level solution $y^*(x) \in \operatorname{argmin}_{y \in \mathcal{Y}(x)} g(x, y)$ could be nonsmooth, discontinuous or even set-valued based on the specific form of $\mathcal{Y}(x)$ and $g(x, y)$ (see Khanduri et al. (2023, Page 3) and our Example 3.1).

In this work, we consider the following BLO problem with lower-level constraints:

$$\min \ f(x, y^*(x)) \quad \text{s.t.} \quad x \in \mathcal{X}, \ y^*(x) \in \operatorname*{argmin}_{\{y: h_i(x, y) \leq 0, \ i=1, \dots, k\}} g(x, y), \tag{1.3}$$

where $f(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is (possibly) a non-convex function, also with the lower-level objective function $g(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and the constraints $h_i(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $\mathcal{X}$ is a compact set. We refer to lower-level constraints as *coupled*, meaning that (at least one) of the constraint functions $h_i(x, y)$, $i \in [k]$, depends on both x and y. Otherwise, they are referred to as *uncoupled*.

If the constraints of the lower-level problem are uncoupled, many algorithms have been studied to deal with such problems. However, these algorithms fail to solve coupled constrained problems. While there are studies addressing coupled constrained BLO, they typically rely on strong assumptions and often only guarantee asymptotic convergence. For more details about these existing works, see Section 1.1. To avoid assumptions that are difficult to verify and to achieve non-asymptotic convergence results, in this work we propose a barrier function-based method for BLO. More specifically, we consider a smooth approximation of the original problem (1.3) as follows

$$\min \ f(x, y) \quad \text{s.t.} \quad x \in \mathcal{X}, \ y = \operatorname*{argmin}_{y \in \mathbb{R}^m} \widetilde{g}_t(x, y) := g(x, y) - t \sum_{i=1}^{k} \log(-h_i(x, y)), \tag{1.4}$$

where t is a sufficiently small fixed constant. That is, we transform the lower-level constraints into a log-barrier added to the lower-level objective function, thereby converting the lower-level problem into an unconstrained one. Theoretically, this transformation makes the gradient of the hyperfunction computable, allowing us to use the standard implicit function approach to solve the resulting BLO problem efficiently and with theoretical guarantees. Practically, log-barrier reformulation models decision-making processes more realistically: instead of abruptly rejecting a decision at the boundary, it gradually decreases acceptance as the decision variable approaches the boundary, simulating a continuous and natural decline in feasibility.

Next, we will review related works for solving various forms of BLO problems discussed so far.

---

[1]Note, such a function exists under certain conditions, which can be broader than the one stated above; We will discuss those conditions in detail later.

Algorithm	$g(x,y)\&h(x,y)$	Reformulation	Complexity	Nonstandard Assumptions
IGBA†	S-C&C	BR	non-asymptotic	needed
GAM	S-C&C	OP	asymptotic	not needed
BSG†	S-C&C	KR	asymptotic	needed
iP-DwCA†	J-W-C&J-C	MVFR	asymptotic	needed
APPM†	C&C	VFR	non-asymptotic	needed
BiC-GAFFA†	C&C	LPR	non-asymptotic	needed
IGOGIC	S-C&L	LPR	non-asymptotic	needed
BLOCC†	S-C&C	LPR	non-asymptotic	needed
BFBM (proposed)	S-C&C or L&L	BR	non-asymptotic	not needed

Table 1: Comparison of BFBM algorithm (our proposed Algorithm 1) with different algorithms for addressing coupled constrained BLO problem: IGBA (Tsaknakis et al., 2023b), GAM (Algorithm 1 in Xu and Zhu (2023)), BSG (Giovannelli et al., 2021), iP-DwCA (Gao et al., 2023), APPM (Algorithm 4 in Lu and Mei (2024)), BiC-GAFFA (Yao et al., 2024), IGOGIC (Algorithm 5 in Kornowski et al. (2024)), BLOCC (Algorithm 1 and 2 in Jiang et al. (2024)). BR means barrier reformulation; KR means KKT-based reformulation; MVFR means relaxed Moreau envelope-based Value function reformulation; VFR means Value function reformulation; LPR means Lagrangian-based penalized reformulation; OP means original problem. C means convex in y; S-C means strongly convex in y; N-C means non-convex in y; J-C means jointly convex in (x,y); J-W-C means jointly weakly convex in (x,y); L means linear in y. $\dagger$ means that authors did not consider the relation of the stationary point between reformulation and original problem, or the relation of hypergradient between reformulation and original problem. Nonstandard assumptions refer to assumptions beyond LICQ, Lipschitz smoothness, compactness, and Slater condition.

1.1 Related Works

Bilevel Optimization without lower-level constraints. BLO without lower-level constraints has seen remarkable advancements through gradient-based approaches such as approximate implicit differentiation (AID) and iterative differentiation (ITD). AID, utilizing the implicit function theorem, approximates the hypergradient and has demonstrated effective finite-time, and finite sample (when the problem is stochastic) convergence for unconstrained, strongly convex lower-level problems (Ghadimi and Wang, 2018; Yang et al., 2021; Hong et al., 2023; Pedregosa, 2016). ITD further enhances convergence by differentiating the entire iterative algorithm used at the lower-level, supported by substantial subsequent works (Maclaurin et al., 2015a; Franceschi et al., 2017; Nichol et al., 2018; Ji et al., 2021). Utilizing AID and momentum-based algorithms, Khanduri et al. (2021); Yang et al. (2023) achieves $\mathcal{O}(\epsilon^{-3})$ rate of convergence to achieve an ϵ-stationary point for stochastic BLO. Additionally, penalty-based methods have become popular, simplifying computational complexities by transforming BLO into single-level problems using various penalty terms (Mehra and Hamm, 2019; Shen et al., 2023; Kwon et al., 2024). Both AID and ITD utilize the Hessian of the lower-level objective for estimating the hypergradients, and a recent line of works (Kwon et al., 2023; Yang et al., 2023) develops fully first order algorithms by leveraging finite-difference to estimate the Hessian-vector product needed to compute the hypergradients.

Bilevel Optimization with lower-level constraints. BLO with lower-level constraints is significantly more complex as compared with their unconstrained counterparts. Previous studies have typically addressed upper-level constraints in various works (Chen et al., 2022a,b). Approaches for uncoupled lower-level constraints include the SIGD method (Khanduri et al., 2023), which targets the constraint $Ay \leq b$ and demonstrates asymptotic convergence. Shi et al. (2024) proposes a double-momentum based algorithm for lower-level constrained BLO and provides convergence analysis toward the so-called (δ, ϵ)-stationary point, where they use a finite-difference technique to avoid the computation of Hessian-vector product and yield a dimension-dependent rate of convergence. Works such as Shen et al. (2023); Kwon et al. (2024) utilize penalty reformulations to handle both upper

and lower uncoupled constraints. In particular, if the lower-level constraint is a smooth manifold (for example, a sphere), Li and Ma (2024); Han et al. (2024); Dutta et al. (2024) consider manifold optimization techniques for efficiently solving the lower-level constrained BLO problems. However, none of these works are able to handle lower-level coupled constraints, i.e. the situation where the lower-level constraints vary according to x. For example, in Shen et al. (2023), the authors introduce a penalty-based bilevel gradient descent (PBGD) algorithm that reformulates the bilevel problem into a single-loop penalized problem solved using the projected gradient descent method. However, despite the convexity of the feasible set $\mathcal{Y}(x) := \{y : h_i(x, y) \leq 0, \, i = 1, ..., k\}$ for each x, the product set $\mathcal{X} \times \mathcal{Y}(x)$ may not retain convexity. Therefore, in the case of coupled constraints, the PBGD algorithm that transforms the original problem into a single loop algorithm will no longer be applicable, as the domain may now be non-convex. Moreover, the method of computing the gradient of the value function, which is necessary for the PBGD algorithm, is also not applicable in the case of coupled constraints. In Khanduri et al. (2023), the authors employed an implicit gradient descent-based method, but the calculation of the hypergradient is no longer applicable in the coupled case.

When we have lower-level coupled equality constraints which are affine in y (e.g. $\{y : h(x) + Ay = c\}$), Xiao et al. (2023) considers a direct extension of AID-based method for hypergradient estimation. The key in their analysis is that equality constraints will not affect the differentiability of the lower-level solver $y^*(x)$ (see Lemma 2 in Xiao et al. (2023)). Another work that deals with lower-level equality constraints is Kornowski et al. (2024), where the authors propose a finite-difference approximation method for the hypergradient, provided that the lower-level equality constraint is linear for both x and y.

However, lower-level (coupled) inequality constraints are much harder to deal with since inequality constraints will result in a nonsmooth lower-level solver $y^*(x)$. In Giovannelli et al. (2021), the authors consider stochastic BLO with lower-level coupled inequality constraints and provide an asymptotic convergence analysis to the stationary point under the assumption that for each x, there exists a lower-level solver $y^*(x)$ satisfying the lower-level KKT conditions such that LICQ, strict complementarity slackness condition (SCSC) and second-order sufficient condition (SOSC) are also satisfied. Xu and Zhu (2023) proposed a gradient approximation scheme for the hypergradient also under LICQ assumptions on the lower-level constraints and proved its asymptotic convergence to the stationary points. Gao et al. (2023) consider a Moreau Envelope-based algorithm and obtain an asymptotic convergence result. Recently, Jiang et al. (2024) considered a penalty-based primal-dual reformulation to translate the BLO with lower-level coupled constraints problem into a single-level minimax optimization. The authors proved a non-asymptotic rate of convergence for the reformulated problem toward ϵ-stationary points under the assumptions that LICQ is satisfied at each point and an upper bound for the dual variable (see Lemma 2 in Jiang et al. (2024)). Beyond LICQ, they also need a curvature condition assumption on the multipliers to obtain this convergence result (see Assumption 5 in Jiang et al. (2024)), and such an assumption is often not easy to verify. Another recent work Yao et al. (2024) also considers a penalty-type reformulation and provides a non-asymptotic rate of convergence toward the ϵ-stationary point of the reformulated problem. However, this algorithm is largely an infeasible algorithm and needs an explicit assumption on the upper bound of the function value at each iteration in order to achieve feasibility. Lu and Mei (2024) also considers a penalty reformulated problem of the lower-level constrained BLO and shows the asymptotic convergence toward stationary points under appropriate assumptions. When each of the coupled constraints is linear to both x and y, Kornowski et al. (2024) proposes a penalty and finite-difference based algorithm which converges to (δ, ϵ)-stationary point with a non-asymptotic rate under the assumption of having access to the optimal dual variable, which is difficult in practice. We summarize these literature on BLO with lower-level coupled inequality constraints along with

our proposed algorithm in Table 1.

A recent work Tsaknakis et al. (2023b) proposed a certain barrier approximation approach for BLO problems. Despite the fact that the idea of using barrier approximation is close to the current work, the relationship between the reformulation and the original problem is not fully investigated. Further, an explicit and unconventional assumption on the lower-level constraint function values is imposed to achieve theoretical convergence. Moreover, a black-box lower-level solver is required to avoid the non-Lipschitz smooth nature of the barrier reformulated lower-level problem.

Additionally, it is worth mentioning that the coupled constrained problem has recently been investigated for a special subclass of the BLO problem: the minimax problem. Tsaknakis et al. (2023a) defined a type of local minimum point for minimax problems with coupled constraints and designed an algorithm that converges to such points.

1.2 Contributions

As mentioned before, in this work we propose a reformulation (1.4), and show that it approximates well to the original problem (1.3), in terms of both function values and gradients. Further, we propose an algorithm that optimizes the reformulated problem, and analyze its non-asymptotic convergence under mild conditions. Our main contributions are summarized below:

1. To our knowledge, this is the first work that systematically study the approximation error between the barrier reformulated problem (1.4) and the original problem (1.3) in terms of both hyperfunction values and hypergradients[2]. In particular, we show that approximation errors of hyperfunction values can be properly quantified not only when the lower-level objective g is strongly convex in y and the constraints $h_i(x,y)$'s are convex in y, but also when the lower-level objective g and constraints $h_i(x,y)$'s are all linear in y. To the best of our knowledge, we are the first to prove such a result for the latter setting. We also proved the asymptotic convergence of hypergradient at **SCSC** point (see Definition 3.1) as the barrier coefficient t goes to 0 within these two settings. To show these results, we develop several novel proof techniques. In particular, we first show $\lim_{t\to 0} y_t^*(x) = y^*(x)$ where $y^*(x)$ and $y_t^*(x)$ are the optimal lower-level solution for (1.3) and (1.4) respectively. We further show that differentiation and limits are interchangeable, i.e. $\lim_{t\to 0} \nabla_x y_t^*(x) = \nabla_x \lim_{t\to 0} y_t^*(x) = \nabla_x y^*(x)$, by establishing the uniform convergence of $\nabla_x y_t^*(x)$. This interchangeability is crucial for the convergence of the hypergradient.

2. Based on the previous contribution, we develop a new adaptive algorithm that theoretically achieves non-asymptotic $\widetilde{\mathcal{O}}(1/(\epsilon^2 t^{4.5}))$ convergence rate to the ϵ stationary point of the barrier reformulated problem (1.4) under mild assumptions. One challenge in our algorithm design and analysis is that $\widetilde{g}_t(x, y_t^*(x))$ explodes when $h_i(x, y_t^*(x))$ is close to zero, so the hyperfunction and lower-level problem of (1.4) may not enjoy global Lipschitz gradient property. To overcome this difficulty, we present a key result (in Theorem 4.1), which states that if $h_i(x, y_t^*(x))$ tends towards zero as t approaches zero, it does so at most linearly with respect to t, and the linear coefficient is *independent* of both t and x. This result allows us to transform the original non-Lipschitz smooth lower-level problem into a Lipschitz smooth one by shrinking the feasible set of the lower-level problem, and to design the step sizes for the outer loop adaptively. Additionally, under strongly convex setting, when assuming that the upper-bounds for some provably bounded terms are known, the convergence rate can be improved significantly to $\widetilde{\mathcal{O}}(1/(\epsilon^2 t^{1.5}))$.

---

[2]Note that the approximation error in terms of hyperfunction values when g is strongly convex is first studied in Tsaknakis et al. (2023b), whereas the hypergradient is not studied before.

3. We provide numerical experiments on a class of strongly convex lower-level problems and a class of linear lower-level problems, both with linear inequality constraints. We compare our method with a number of existing works and verify the effectiveness of the proposed method. Our algorithm is the only one that always guarantees that the obtained solution is feasible for the lower-level problem, and it is the most effective algorithm among those we tested when lower-level is a linear program.

1.3 Notations and Terminology

The basic notations used throughout this paper are introduced in this section. Additional notations can be found in Appendix B.1 for clarity and reference:

- $\mathbb{R}^n$ is the Euclidean space of dimension n, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are the canonical inner product and norm of Euclidean space;

- $B_x(r)$ is the ball centered at the point x with radius r, i.e. $B_x(r) = \{z \in \mathbb{R}^n : \|z - x\| \leq r\}$;

- Let $y^*(x)$ and $y_t^*(x)$ be the optimal solution set of the original problem (1.1) and the barrier reformulation problem (1.4) respectively;

- Let $\phi(x)$ and $\widetilde{\phi}_t(x)$ be hyperfunction (see (1.2) and (3.1) for definition) of the original problem (1.1) and the barrier reformulation problem (1.4) respectively, and $\widetilde{g}_t(x, y) := g(x, y) - t \sum_{i=1}^{k} \log(-h_i(x, y))$ be the reformulated lower-level objective function;

- Define $\mathcal{Y}(x) := \{y : h_i(x, y) \leq 0, \ i = 1, ..., k\}$ as the feasible set of lower-level problem, and k is the number of constraints for lower-level problem;

- $\widetilde{\mathcal{O}}(\cdot)$ represents $\mathcal{O}(\cdot)$ with logarithmic term omitted.

2 Motivating Application

Let us briefly discuss a few applications for the considered BLO problems.

Linear Setting: In the first setting, both $g(x, y)$ and $h_i(x, y)$ are linear functions of y for any i. One relevant application is the bilevel price-setting problem in the transportation context (see Labbé and Violin (2016)), which can be formulated as:

$$\max_{T} \quad T^\top x \tag{2.1}$$
$$\text{s.t.} \quad (\text{x}, \text{y}) \in \arg \min_{\text{x,y}} \quad (c_1 + T)^\top x + c_2^\top y$$
$$\text{s.t.} \quad A_1(T)x + A_2(T)y \geq b(T), \quad x, y \geq 0$$

where $T \in \mathbb{R}^{d_x}$ is the upper-level decision variable, representing the tax value, and $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ are the lower-level decision variables, corresponding to the allocation of transportation activities across taxable and untaxable routes. The matrices $A_1(T) \in \mathbb{R}^{k \times d_x}$ and $A_2(T) \in \mathbb{R}^{k \times d_y}$ represent the constraints that govern the feasibility of transportation along taxed and untaxed routes.

In this specific transportation setting, the leader is a regulatory authority or city planner, who decides the tax value T on transportation activities. The leader's goal is to maximize the revenue from the tax, represented by $T^\top x$. The follower is a freight transporter who needs to minimize their total transportation cost, which includes both the tax on taxable routes and the cost of using taxed and untaxed routes. The cost function is $(c_1 + T)^\top x + c_2^\top y$, where c_1 and c_2 are the cost

coefficients for the taxable and nontaxable routes, respectively, and T adjusts the cost of the taxable routes. The matrices $A_1(T)$ and $A_2(T)$ represent coefficients that influence the capacity or feasibility of using taxed and untaxed routes for transportation, and $b(T)$ means demands from different customers. These matrices depend on T because the tax may impact the availability or cost of using taxed routes. The inequality $A_1(T)^\top x + A_2(T)^\top y \geq b(T)$ ensures that customer demands are met. The constraints $x, y \geq 0$ enforce that the quantities of goods transported along both types of routes cannot be negative, reflecting the real-world constraint that transportation volumes must be non-negative.

Strongly Convex Setting: In the second setting, $g(x, y)$ is strongly convex, and $h_i(x, y)$ are convex in y for any i. The primary applications in this setting arise from machine learning. For example, support vector machines (SVM) inherently possess a bilevel optimization structure, where hyper-parameter optimization can be framed as a constrained BLO problem; we refer the readers to recent works such as (Jiang et al., 2024; Xu and Zhu, 2023) for more detailed discussions.

3 Barrier Reformulation of Bilevel Problems

In this section, we will explore the hyperfunction of the barrier reformulation (1.4), defined below, and its relationship with the hyperfunction of the original problem (1.3):

$$\widetilde{\phi}_t(x) = f(x, y_t^*(x)), \text{ where } y_t^*(x) = \operatorname*{argmin}_{y \in \mathbb{R}^m} \widetilde{g}_t(x, y) := g(x, y) - t \sum_{i=1}^k \log(-h_i(x, y)). \tag{3.1}$$

Specifically, we first demonstrate that the barrier reformulation is differentiable under certain conditions. This is mainly because the lower-level problem is strictly convex (see Proposition 3.1). Subsequently, we investigate the convergence of the function values and gradients of the hyperfunction as t approaches 0, under various different assumptions of the lower-level problem.

3.1 Preliminaries

In this section, we study some basic properties of the barrier reformulation problem (1.4). We make the following assumption throughout this section.

Assumption 3.1. *The following holds for every $x \in \mathcal{X}, y \in \mathcal{Y}(x)$, and $i \in \{1, ..., k\}$*

1. *$f(x, y)$ is once while $g(x, y)$ and $h_i(x, y)$ are twice continuously differentiable;*

2. *$\mathcal{X}$ is convex and compact, and for any $x \in \mathcal{X}$ there exists $y \in \mathcal{Y}(x)$ such that $h_i(x, y) < 0$ for any $i \in \{1, 2, ..., k\}$;*

3. *Linear Independence Constraint Qualification (LICQ) is satisfied for any (x, y) where $x \in \mathcal{X}$ and $y \in y^*(x) = \arg\min_{y \in \mathcal{Y}(x)} g(x, y)$, i.e. the gradients $\{\nabla_y h_i(x, y) : i \text{ is active}\}$ are linearly independent for any $x \in \mathcal{X}$ and $y \in y^*(x)$.*

Assumption 3.1(1) is a basic assumption in the BLO literature (Ghadimi and Wang, 2018; Hong et al., 2023; Xu and Zhu, 2023). Assumption 3.1(2) says that Slater's condition is satisfied for any x which is a common assumption for studying BLO problems under constraints (Tsaknakis et al., 2023b; Schmidt and Beck, 2023; Beck et al., 2023). Assumption 3.1(3) is also a standard requirement, commonly used across various studies in BLO (Jiang et al., 2024; Kornowski et al., 2024; Khanduri et al., 2023).

Next, we have two different set of assumptions on $g(x, y)$ and $h_i(x, y)$'s.

Assumption 3.2. *$g(x, y)$ is μ_g-strongly convex in y for any $x \in \mathcal{X}$; $h_i(x, y)$ is convex in y for any $x \in \mathcal{X}$ and $i \in \{1, ..., k\}$.*

Assumption 3.3. *$g(x, y)$, $h_i(x, y)$'s are all linear in y for for any $x \in \mathcal{X}$ and $i \in \{1, ..., k\}$, i.e. the lower-level problem is a linear program, and $\mathcal{Y}(x)$ is compact for any $x \in \mathcal{X}$.*

Remark 3.1. *When the lower-level problem is a linear program, to make sure the feasible set $\mathcal{Y}(x)$ is compact, the number of constraints should be greater than the dimension of the lower-level problem, i.e. $k \geq m$.*

The strong convexity requirement in Assumption 3.2 has been the focus of most recent research on BLO, such as Xu and Zhu (2023); Khanduri et al. (2023); Hong et al. (2023); Yao et al. (2024). In contrast, Assumption 3.3, which requires that the lower-level problem is a linear program, has not yet been explored. The compactness of the feasible set in the lower-level problem is a technical assumption made to ensure that the hyperfunction in the barrier reformulation is differentiable, which is observed in the following proposition. See Appendix D.1 for proof.

Proposition 3.1. *Suppose that Assumption 3.1(1) and Assumption 3.1(2) hold, and either Assumption 3.2 or Assumption 3.3 holds. Then $\widetilde{g}_t(x, y)$ is strictly convex in y, and $\widetilde{\phi}_t(x)$ is differentiable for any $t > 0$ and $x \in \mathcal{X}$.*

We note that the above result holds regardless of the number of the lower-level optimal solutions for the original problem (1.3). The differentiability of $\widetilde{\phi}_t(x)$ allows us to design algorithms for the barrier reformulated problem. However, before we design algorithms, we need to understand the relationship between $\widetilde{\phi}_t(x)$ and the original hyperfunction $\phi(x)$.

3.2 Convergence of Hyperfunction Value for Barrier Reformulation

In this section, we study the relationship between the hyperfunction of problems (1.3) and its barrier reformulation (1.4). Such a study is critical as it helps reveal the utility of the proposed reformulation.

Under Assumption 3.2 where the lower-level problem is strongly convex in y, it can be shown that hyperfunction of the original problem and that of the barrier reformulated problem can be bounded uniformly (see Tsaknakis et al. (2023b)):

Theorem 3.1 (Lemma 1 in Tsaknakis et al. (2023b)). *Suppose that Assumption 3.1(1), 3.1(2), and Assumption 3.2 hold. If f is Lipschitz continuous with coefficient L_f, then the following holds:*

$$|\widetilde{\phi}_t(x) - \phi(x)| \leq L_f \left(\frac{2kt}{\mu_g} \right)^{1/2}.$$

Under Assumption 3.3, i.e. the linear setting, we aim to establish the error bound between the hyperfunction values as a function of t when $y^*(x)$ is unique. The reason that we do not attempt to show global uniform bounds is that in the case where the lower-level problem is a linear program, the hyperfunction of the original problem may be discontinuous, and even for the continuous points, the bound may not be uniform. Below we provide an example.

Example 3.1. *Consider a BLO with $f(x, y) = y_1$ with $x \in \mathbb{R}$, $y = (y_1, y_2)^\top \in \mathbb{R}^2$. Further, the lower-level problem takes the following form:*

$$\min_y g(x, y) := xy_1 + y_2$$

8

$$\text{s.t.} \quad y_2 \geq 0, \ y_1 \geq -1, \ y_1 \leq 1.$$

It is easy to see that the lower-level problem has the following solutions:

$$\arg\min_y g(x,y) = \begin{cases} (-1,0), & \text{if } x > 0 \\ [-1,1] \times \{0\}, & \text{if } x = 0 \\ (1,0), & \text{if } x < 0. \end{cases}$$

The barrier-penalized reformulation of $g(x,y)$ is:

$$\widetilde{g}_t(x,y) = xy_1 + y_2 - t(\log(y_2) + \log(1 - y_1) + \log(1 + y_1)).$$

Direct calculation shows that its stationary point is

$$\begin{cases} y_1 = \begin{cases} 0, & \text{if } x = 0 \\ \frac{t - \sqrt{t^2 + x^2}}{x}, & \text{if } x \neq 0 \end{cases} \\ y_2 = t. \end{cases}$$

The hyperfunction of the original problem is

$$\phi(x) = \begin{cases} -1, & \text{if } x \geq 0 \\ 1, & \text{if } x < 0 \end{cases}$$

and the hyperfunction corresponding to the barrier-penalized reformulation is

$$\widetilde{\phi}_t(x) = \begin{cases} 0, & \text{if } x = 0 \\ \frac{t - \sqrt{t^2 + x^2}}{x}, & \text{if } x \neq 0. \end{cases}$$

However, $\left|\widetilde{\phi}_t(0) - \phi(0)\right| = 1$ for any t. Moreover, when $x \neq 0$, we have $\left|\widetilde{\phi}_t(x) - \phi(x)\right| = |\frac{t - \sqrt{t^2 + x^2}}{x} + 1|$ when $x > 0$ or $\left|\widetilde{\phi}_t(x) - \phi(x)\right| = |\frac{t - \sqrt{t^2 + x^2}}{x} - 1|$ when $x < 0$, and both of these terms do not admit a uniform bound with respect to t.

Clearly, the discontinuity at $x = 0$ arises because $y^*(x)$ experiences a jump when transitioning from $x > 0$ to $x < 0$, or, in other words, $y^*(0)$ is not unique. Further, it is important to note that even if we exclude the point 0, there does not exist a uniform error bound as a function of t and is independent of x. Therefore, under the linear setting, we will only analyze those x that admits *unique* $y^*(x)$, and we seek to establish a *pointwise* convergence relationship, that is, $\lim_{t \to 0} \widetilde{\phi}_t(x) = \phi(x)$. In the case where the lower-level problem is simply a linear program the uniqueness of the solution then depends on its dual solution being non-degenerate. Since vector x plays the role of the right-hand side of the dual constraints in this case, if x is subject to small perturbations or imprecision then by the perturbation theory for linear programming, the dual optimal solution of the lower-level problem is likely to be non-degenerate after perturbation. Hence, the primal, which is the lower-level problem itself, will likely have a unique optimal solution in that case.

To proceed with our analysis, we first present the following known result (exercise 11.12(b) in Chapter 11 of Boyd and Vandenberghe (2004)).

Lemma 3.1 (Optimality gap). *Suppose that Assumption 3.1(1), 3.1(2) hold, and $g(x,y)$, $h_i(x,y)$'s are convex in y, then we have the optimality gap*

$$g(x, y_t^*(x)) - g(x, y^*(x)) \leq kt.$$

To use the above result under linear setting, note that if $y^*(x)$ is unique, then for any point y within the feasible region of the lower-level problem, the cosine of the angle between $y - y^*(x)$ and the $\nabla_y g(x, y^*(x))$ has a positive lower bound for any fixed x, which we denote as $\tau(x)$, i.e.

$$\tau(x) = \operatorname*{argmin}_{y \in \mathcal{Y}(x)} \frac{\langle y - y^*(x), \nabla_y g(x, y^*(x)) \rangle}{\|y - y^*(x)\| \|\nabla_y g(x, y^*(x))\|}.$$

It follows that:

$$g(x, y) - g(x, y^*(x)) = \langle y - y^*(x), \nabla_y g(x, y^*(x)) \rangle \geq \tau(x) \|\nabla_y g(x, y^*(x))\| \cdot \|y - y^*(x)\|.$$

By uniqueness of $y^*(x)$ and Slater condition, $\nabla_y g(x, y^*(x)) \neq 0$, then we obtain:

$$\|y - y^*(x)\| \leq \frac{1}{\tau(x) \|\nabla_y g(x, y^*(x))\|} \left(g(x, y) - g(x, y^*(x)) \right). \tag{3.2}$$

Combining Lemma 3.1, we obtain

$$\|y - y^*(x)\| \leq \frac{kt}{\tau(x) \|\nabla_y g(x, y^*(x))\|}.$$

If we further assume that $f(x, y)$ is L_f-Lipschitz continuous, then

$$|\widetilde{\phi}_t(x) - \phi(x)| = |f(x, y_t^*(x)) - f(x, y^*(x))| \leq L_f \frac{kt}{\tau(x) \|\nabla_y g(x, y^*(x))\|}.$$

To conclude, we establish the following theorem:

Theorem 3.2 (Hyperfunction value convergence under Linear setting). *Suppose that Assumption 3.1(1), 3.1(2), and Assumption 3.3 hold. If $f(x, y)$ is Lipschitz continuous with coefficient L_f and $y^*(x)$ is unique at point x, then*

$$|\widetilde{\phi}_t(x) - \phi(x)| \leq L_f \frac{kt}{\tau(x) \|\nabla_y g(x, y^*(x))\|},$$

where $\tau(x) := \operatorname{argmin}_{y \in \mathcal{Y}(x)} \cos \langle \nabla_y g(x, y^*(x)), y - y^*(x) \rangle > 0$.

3.3 Convergence of Hypergradient for Barrier Reformulation

In this subsection, we explore the convergence behavior of the hypergradient of the barrier reformulation $\nabla_x \widetilde{\phi}_t(x)$ as t goes to zero. Understanding this convergence is vital because optimization algorithms commonly converge to points where the gradient is zero. Therefore, it is crucial to ensure that when $\nabla_x \widetilde{\phi}_t(x)$ approaches zero, $\nabla_x \widetilde{\phi}_t(x)$ also becomes sufficiently small. Since the hyperfunction of the original problem may be discontinuous (under linear setting, see Example 3.1) or non-smooth (under strongly convex setting, see the example in Khanduri et al. (2023, Page 3)), we aim to show asymptotic convergence at the smooth points of $\phi(x)$.

We present the following definition as an essential characterization of the smoothness of the original problem.

Definition 3.1 (SCSC point). *Suppose Assumption 3.1 holds. We say x is an **SCSC (Strict Complementarity Slackness Condition) point** if for any $y \in y^*(x)$, we have that $h_i(x, y) = 0$ implies $\lambda_i(x, y) > 0$, where $y^*(x)$ is the set of optimal solutions for the lower-level problem, and $\lambda_i(x, y)$ is the optimal Lagrangian multiplier corresponding the the i-th constraint.*

We first provide a remark regarding the notation of the multipliers.

Remark 3.2. *Under the strong convex setting, $y^*(x)$ is unique, so $\lambda_i(x, y^*(x))$ only depends on x. Under the linear setting, although $y^*(x)$ is not unique and forms a set, due to the LICQ assumption, the value of the multiplier $\lambda_i(x, y)$ actually is independent of the choice of y in the optimal solution set $y^*(x)$. For the specific proof, see $(a) \Rightarrow (b)$ part of the proof of the following Proposition 3.2. Therefore, in both cases we denote $\lambda_i(x, y^*(x))$ simply as $\lambda_i(x)$ without loss of generality.*

Note that $\nabla_x \phi(x) = \nabla_x f(x, y) + (\nabla_x y^*(x))^\top \nabla_y f(x, y)$, the existence of $\nabla_x y^*(x)$ is a sufficient condition of the differentiability of the hyperfunction $\phi(x)$. The following proposition shows the relation between **SCSC** point and the existence of $\nabla_x y^*(x)$. Please see Appendix E.1 for the proof.

Proposition 3.2. *Suppose Assumption 3.1 holds. We have the following results:*

1. *If Assumption 3.2 holds, then x is an **SCSC** point implies $\nabla_x y^*(x)$ exists, but not vice versa;*

2. *If Assumption 3.3 holds, then following conditions are equivalent:*

   *(a) x is an **SCSC** point;*

   (b) $y^*(x)$ is unique;

   (c) $\nabla_x y^*(x)$ exists.

We point out that Khanduri et al. (2023, Page 3) provides an example where the hyperfunction is nonsmooth when **SCSC** is not satisfied, under the case that the lower-level objective function is strongly convex.

Next, we develop an analysis of the asymptotic behavior of the error based on the **SCSC** assumption, that is, we will show $\lim_{t \to 0} \nabla_x \widetilde{\phi}_t(x) = \nabla_x \phi(x)$. By direct computation, we obtain:

$$\lim_{t \to 0} \nabla_x \widetilde{\phi}_t(x) = \lim_{t \to 0} \left(\nabla_x f(x, y_t^*(x)) + (\nabla_x y_t^*(x))^\top \nabla_y f(x, y_t^*(x)) \right)$$

$$= \nabla_x f(x, y^*(x)) + \left(\lim_{t \to 0} \nabla_x y_t^*(x) \right)^\top \nabla_y f(x, y^*(x)),$$

$$\nabla_x \phi(x) = \nabla_x f(x, y^*(x)) + (\nabla_x y^*(x))^\top \nabla_y f(x, y^*(x)).$$

Thus, the core issue is to prove that the Jacobian matrix converges, i.e. $\lim_{t \to 0} \nabla_x y_t^*(x) = \nabla_x y^*(x)$. We will prove this conclusion separately for the linear and strongly convex settings.

First, under the linear setting, the proof is relatively simple. This is because $\lim_{t \to 0} \nabla_x y_t^*(x)$ has a very straightforward form, and $\nabla_x y^*(x)$ also has a simple local expression (see (E.16)), which exactly equals $\lim_{t \to 0} \nabla_x y_t^*(x)$. In particular, we have the following result, whose proof can be found in Appendix E.2.

Theorem 3.3 (Jacobian convergence under Linear setting). *Suppose Assumption 3.1, 3.3 hold. If x is an **SCSC** point of BLO problem (1.3), then we have $\lim_{t \to 0} \nabla_x y_t^*(x) = \nabla_x y^*(x)$.*

Under the strongly convex setting, the situation is more complicated. The reason is that $\lim_{t \to 0} \nabla_x y_t^*(x)$ becomes very difficult to determine. Even though $\nabla_x y^*(x)$ is computable (Giorgi and Zuccotti, 2018, Section 3), proving that it equals $\lim_{t \to 0} \nabla_x y_t^*(x)$ becomes extremely challenging. Therefore, we use a different method to prove the conclusion.

Note that under the strongly convex setting, $y_t^*(x)$ uniformly converges to $y^*(x)$, which is a direct consequence of the optimality gap in Lemma 3.1. To prove $\lim_{t \to 0} \nabla_x y_t^*(x) = \nabla_x \lim_{t \to 0} y_t^*(x) = \nabla_x y^*(x)$, we essentially need to show that differentiation and limits are interchangeable. A common

sufficient condition for this interchangeability is that the Jacobian matrix $\nabla_x y_t^*(x)$ converges uniformly in a neighborhood around x. By proving the uniform convergence of $y_t^*(x)$ to $y^*(x)$ (see Lemma B.1 in Appendix B.2) and local uniform convergence of $\nabla_x y_t^*(x)$ (see Lemma E.7 in Appendix E.3) of the Jacobian around the **SCSC** point, we are able to obtain the following conclusion.

Theorem 3.4 (Jacobian convergence under Strongly convex setting)**.** *Suppose Assumption 3.1, and Assumption 3.2 hold, i.e. when $g(x, y)$ is strongly convex in y. If x is an **SCSC** point of BLO problem (1.3), then we have $\lim_{t \to 0} \nabla_x y_t^*(x) = \nabla_x y^*(x)$.*

To conclude, we have the following main result for the hypergradient convergence in two cases.

Theorem 3.5 (Hypergradient convergence)**.** *Suppose Assumption 3.1 holds, and either Assumption 3.2 or Assumption 3.3 holds. If x is an **SCSC** point of BLO problem (1.3), then $\lim_{t \to 0} \nabla_x \widetilde{\phi}_t(x) = \nabla_x \phi(x)$, where $\nabla_x \widetilde{\phi}_t(x)$ is the hypergradient of the barrier reformulation (1.4).*

The reader may wonder if the above convergence can be made uniform. However the remark below indicates that such a form of convergence is impossible.

Remark 3.3. *Denote the set of **SCSC** points as $\widetilde{\mathcal{X}}$. Under Assumption 3.1, Assumption 3.2 or Assumption 3.3, it is impossible to get uniform error estimation between $\nabla_x \phi(x)$ and $\nabla_x \widetilde{\phi}_t(x)$ even on the set $\widetilde{\mathcal{X}}$. This is because if the convergence of hypergradient for the barrier reformulation would be uniform on $\widetilde{\mathcal{X}}$, then it is not hard to prove that $\nabla_x \phi(x)$ is well defined for any x on the boundary of $\widetilde{\mathcal{X}}$, denoted as $\partial \widetilde{\mathcal{X}}$. However, the hypergradient of the original problem may not be well defined at some point on $\partial \widetilde{\mathcal{X}}$ (see Example 3.1 and the example in Khanduri et al. (2023, Page 3)), which leads to a contradiction.*

4 Algorithm Design

So far we have studied the relations between the original problem (1.3) and the barrier reformulated problem (1.4), including the function and gradient approximation errors. Now we turn to focus on the algorithm design for solving the barrier reformulated problem (1.4) and provide both theoretical and numerical evidence to support its effectiveness. We aim at designing algorithm that works both for the linear and the strongly convex setting. To concisely state the assumptions required for this section, we present them as follows:

Assumption 4.1. $g(x, y)$ is μ_g-strongly convex in y for any $x \in \mathcal{X}$; $h_i(x, y)$ is convex in y for any $x \in \mathcal{X}$ and $i \in \{1, ..., k\}$.

Assumption 4.2. $g(x, y)$, $h_i(x, y)$ are all linear in y for any $x \in \mathcal{X}$ and $i \in \{1, ..., k\}$, i.e. the lower-level problem is a a linear program.

Note that we only require one of the above two assumptions. Designing an algorithm that works for both settings is highly non-trivial since we lose the global Lipschitz gradient for the lower-level problem for both settings due to the existence of the log-barrier in (1.4), and the Lipschitz smoothness of the hyperfunction $\widetilde{\phi}_t$ is largely unknown. In Tsaknakis et al. (2023b), to avoid this difficulty, the authors impose an assumption (Assumption 3 in Tsaknakis et al. (2023b)) over the function values of h_i at the lower-level optimal points that directly results in the Lipschitz smoothness of the lower-level problem and hypergradient. However, this assumption is not verifiable and the level of difficulty drastically increases without such assumption. We thus design our algorithm by

carefully controlling the local Lipschitz smoothness constant around the lower-level optimal solution at each step, and meticulously controlling the stepsize for each update point. To proceed, we first impose the following assumptions.

Assumption 4.3. *The following holds for every $x \in \mathcal{X}, y \in \mathcal{Y}(x)$, and $i \in \{1, ..., k\}$*

1. *$f(x, y)$ is one time and $g(x, y)$, $h_i(x, y)$ are two times continuously differentiable;*

2. *$\mathcal{X}$ is convex and compact, and for any $x \in \mathcal{X}$ there exists $y \in \mathcal{Y}(x)$ such that $h_i(x, y) < 0$ for any $i \in \{1, 2, ..., k\}$;*

3. *For any $x \in \mathcal{X}$, $\mathcal{Y}(x)$ is compact, and $\|y\| \leq R$ for any y in $\mathcal{Y}(x)$;*

4. *LICQ is satisfied for any (x, y) where $x \in \mathcal{X}$ and $y \in y^*(x) = \arg\min_{y \in \mathcal{Y}(x)} g(x, y)$, i.e. the gradients $\{\nabla_y h_i(x, y) : i \text{ is active}\}$ are linearly independent for any $x \in \mathcal{X}$ and $y \in y^*(x)$.*

Note that differently as Assumption 3.2, we have assumed the compactness of $\mathcal{Y}(x)$ for *both* the linear and the strongly convex setting. We also need the following smoothness assumptions.

Assumption 4.4. *The following holds for every $x, \overline{x} \in \mathcal{X}, y, \overline{y} \in \mathcal{Y}(x)$, and $i \in \{1, ..., k\}$*

1. *$\|\nabla_{(x,y)} f(x, y)\| \leq L_f$;*

2. *$\|\nabla_{(x,y)} f(x, y) - \nabla_{(x,y)} f(\overline{x}, \overline{y})\| \leq \overline{L}_f \|(x, y) - (\overline{x}, \overline{y})\|$;*

3. *$\|\nabla_{(x,y)} g(x, y)\| \leq L_g$;*

4. *$\|\nabla_{yy}^2 g(x, y)\| \leq \overline{L}_g$; $\|\nabla_{xy}^2 g(x, y)\| \leq \overline{L}_g$;*

5. *$\|\nabla_{xy}^2 g(x, y) - \nabla_{xy}^2 g(\overline{x}, \overline{y})\| \leq \overline{\overline{L}}_g \|(x, y) - (\overline{x}, \overline{y})\|$; $\|\nabla_{yy}^2 g(x, y) - \nabla_{yy}^2 g(\overline{x}, \overline{y})\| \leq \overline{\overline{L}}_g \|(x, y) - (\overline{x}, \overline{y})\|$;*

6. *$\|\nabla_{(x,y)} h(x, y)\| \leq L_h$;*

7. *$\|\nabla_{yy}^2 h(x, y)\| \leq \overline{L}_h$; $\|\nabla_{xy}^2 h(x, y)\| \leq \overline{L}_h$;*

8. *$\|\nabla_{xy}^2 h(x, y) - \nabla_{xy}^2 h(\overline{x}, \overline{y})\| \leq \overline{\overline{L}}_h \|(x, y) - (\overline{x}, \overline{y})\|$; $\|\nabla_{yy}^2 h(x, y) - \nabla_{yy}^2 h(\overline{x}, \overline{y})\| \leq \overline{\overline{L}}_h \|(x, y) - (\overline{x}, \overline{y})\|$.*

These assumptions, except Assumption 4.4(3), are standard in BLO literature such as Ghadimi and Wang (2018); Tsaknakis et al. (2023b); Shen et al. (2023). Assumption 4.4(3), which is an assumption of an upper bound for $\nabla g(x, y)$, can be implied by the compactness of $\mathcal{Y}(x)$ and $\mathcal{X}$.

4.1 The Proposed Algorithm

Now we are ready to establish the algorithm framework. We follow the standard bilevel implicit gradient method as in Ghadimi and Wang (2018); Ji et al. (2021); Hong et al. (2023), where we estimate the hypergradient $\nabla_x \phi(x)$ at each update point x and conduct a gradient descent update to get the next iterate.

Specifically, we first find the approximate optimal solution $\hat{y}$ of the lower-level solver $y^*(x)$ for a fixed x, and such $\hat{y}$ is obtained by some careful design (see Algorithm 2). In particular, since the lower-level objective $\widetilde{g}_t$ is not globally Lipschitz smooth in y, we need to carefully shrink the feasible set and estimate a local Lipschitz smoothness constant in the shrunk set, which is reflected

in **Step 3** of Algorithm 2. Then an accelerated gradient descent step is employed to approximate the lower-level optimal solution in **Step 4** of Algorithm 2.

Upon obtaining the approximate optimal lower-level solution, we further design the update for variable x in Algorithm 1. In particular, we use the implicit gradient methods to compute the approximate gradient of the hyperfunction, which is

$$\hat{\nabla}_x \widetilde{\phi}_t(x) = \nabla_x f(x, \hat{y}) - \nabla_{xy}^2 \widetilde{g}_t(x, \hat{y}) \left(\nabla_{yy}^2 \widetilde{g}_t(x, \hat{y}) \right)^{-1} \nabla_y f(x, \hat{y}),$$

where recall that the barrier reformulation for the lower-level problem is defined as

$$\widetilde{g}_t(x, y) = g(x, y) - t \sum_{i=1}^{k} \log(-h_i(x, y)). \tag{4.1}$$

The **main technical challenge** to analyze the proposed algorithm is that the hypergradient may suffer from non-Lipschitzness due to log barriers, similarly as the lower-level objective. In addition, the inexact lower-level solution $\hat{y}$ introduces approximation errors to the exact lower-level solver $y^*(x)$ which further increases the difficulties of the analysis. We overcome these difficulties and obtain an $\widetilde{\mathcal{O}}(1/\epsilon^2)$ rate of convergence (Theorem 4.3) by carefully designing the upper-level stepsize η_s, which contains the information of local Lipschitz constants for the hypergradient. See Section 4.2 for details of the convergence analysis.

Algorithm 1: Hypergradient Based Bilevel Barrier Method for (1.4)

Step 0: Initialization. Given an initial point x_0, the accuracy level ϵ, total iteration S.

for $s = 0, 1, ..., S$ **do**

  Step 1: Solve the inner loop. Solve the lower-level problem for $\hat{y}_s$ via Algorithm 2 such that $\|\hat{y}_s - y^*(x_s)\| \leq \epsilon_s$.

  Step 2: Update x_s.

  Compute the following approximate gradient of $\widetilde{\phi}_t(x)$ at x_s

$$\hat{\nabla}_x \widetilde{\phi}_t(x_s) = \nabla_x f(x_s, \hat{y}_s) - \nabla_{xy} \widetilde{g}_t(x_s, \hat{y}_s)(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1} \nabla_y f(x_s, \hat{y}_s)$$

  Update x_s by

$$x_{s+1} = \mathsf{proj}_{\mathcal{X}} \left(x_s - \eta_s \hat{\nabla}_x \widetilde{\phi}_t(x_s) \right)$$

  where the stepsize η_s can be calculated according to (4.7) in Theorem 4.3.

Output x_s with $s = \mathrm{argmin}_{s=0,...,S-1} \{ \frac{1}{\eta_s} \|x_s - x_{s+1}\| \}$

4.2 Convergence Analysis

Now we aim to develop an algorithm with non-asymptotic convergence rate. We first discuss the main difficulty.

Technical Difficulty: Algorithm 1 leverages the following approximation for the hypergradient:

$$\hat{\nabla}_x \widetilde{\phi}_t(x) = \nabla_x f(x, \hat{y}) - \nabla_{xy}^2 \widetilde{g}_t(x, \hat{y}) \left(\nabla_{yy}^2 \widetilde{g}_t(x, \hat{y}) \right)^{-1} \nabla_y f(x, \hat{y}),$$

Algorithm 2: lower-level solver for Algorithm 1

    Step 0: Initialization. Given reference point x_s, stepsize γ, and constant $d = 1$.

    Step 1: Find the initial d such that $\mathcal{Y}_d$ is nonempty.

    while $\mathcal{Y}_d(x_s) := \{y : h_i(x_s, y) \leq -d, \ i = 1, ..., k\}$ *is empty* **do**

        $d = d/2$

    Denote $d_s := d$ and initialize $y_0 \in \mathcal{Y}_{d_s}(x_s)$,

    Step 2: Estimate the local Lipschitz smoothness constant.

      1. Compute scalar m_s by replacing d in (4.2) with d_s from **Step 1**. Thus $y_t^*(x_s)$ is contained in $\mathcal{Y}_{m_s}(x_s) := \{y : h_i(x_s, y) \leq -m_s, \ i = 1, ..., k\}$;

      2. Compute the Lipschitz smoothness constant $\overline{L}_{\widetilde{g}_t, m_s}$ of $\widetilde{g}_t(x, y)$ in y on $\mathcal{Y}_{m_s}(x_s)$; $\overline{L}_{\widetilde{g}_t, m_s}$ can be computed by replacing m in (4.3) with m_s.

    Step 3: Update based on current estimate of Lipschitz smoothness constant. for $j = 0, 1, ..., J - 1 := \mathcal{O}(\sqrt{\overline{L}_{\widetilde{g}_t, m_s}/\mu_{\widetilde{g}_t}} \log(1/\epsilon_s))$ **do**

      Update

$$
\omega = \begin{cases} y_0 & \text{if } j = 0 \\ y_j + \dfrac{1 - \sqrt{\mu_{\widetilde{g}_t}/\overline{L}_{\widetilde{g}_t, m_s}}}{1 + \sqrt{\mu_{\widetilde{g}_t}/\overline{L}_{\widetilde{g}_t, m_s}}}(y_j - y_{j-1}) & \text{if } j \geq 1, \end{cases}
$$

$$
y_{j+1} = \mathsf{proj}_{\mathcal{Y}_{m_s}(x)}\left(y_j - \frac{1}{\overline{L}_{\widetilde{g}_t, m_s}} \nabla_y \widetilde{g}_t(x_s, \omega) \right),
$$

      where the strongly convexity constant $\mu_{\widetilde{g}_t}$ is from Proposition 4.1.

    Step 4: Output results. Output $\hat{y}_s := y_J$ and constant d_s, m_s.

---

and it is necessary to estimate the Lipschitz constant of $\hat{\nabla}_x \widetilde{\phi}_t(x)$, which further requires an upper bound estimate of $\nabla_{xy}^2 \widetilde{g}_t(x, \hat{y})$. By directly computing the mixed Hessian (or referred to as Jacobian)

$$
\nabla_{xy}^2 \widetilde{g}_t(x, \hat{y}) = \nabla_{xy}^2 g(x, \hat{y}) + t \sum_{i=1}^k \left(\frac{\nabla_{xy}^2 h_i(x, \hat{y})}{-h_i(x, \hat{y})} + \frac{\nabla_x h_i(x, \hat{y})\nabla_y h_i(x, \hat{y})^\top}{h_i^2(x, \hat{y})} \right),
$$

we need a negative upper bound estimate of $h_i(x, \hat{y})$, which can be achieved by upper bounding $h_i(x, y_t^*(x))$ if $h_i(x, \hat{y})$ and $h_i(x, y_t^*(x))$ are close to each other. However upper bounding $h_i(x, y_t^*(x))$ is hard: if t is sufficiently small, for those indices i that are active in the original problem, $h_i(x, y_t^*(x))$ will also approach zero. On the other hand, in order to make sure that $h_i(x, \hat{y})$ and $h_i(x, y_t^*(x))$ are as close as possible, we need to design an effective algorithm for the lower-level problem. It is challenging to design an effective algorithm: due to the log-barrier, the gradient of barrier reformulation $\widetilde{g}_t(x)$ tends to explode near the boundary, making it not globally Lipschitz smooth. Specifically, observe that the gradient of lower-level problem, given below

$$
\nabla_y \widetilde{g}_t(x, y) = \nabla_y g(x, y) + t \sum_{i=1}^k \frac{\nabla_y h_i(x, y)}{-h_i(x, y)},
$$

may goes to infinity when $-h_i(x, y)$ goes to zero. To overcome this, we consider the lower-level problem in a shrunk set $\mathcal{Y}_d(x) := \{y : h_i(x, y) \leq -d, \ i = 1, ..., k\}$. In order to determine the size of

d, such that $y_t^*(x)$ is still in this shrunk set $\mathcal{Y}_d(x)$, we need to identify a negative value that upper bounds $h_i(x, y_t^*(x))$. The next result establishes such an upper bound. Please see Appendix F.2 for proof.

Theorem 4.1. *Suppose $0 < t \leq T$, Assumption 4.3(1)(2)(3) and Assumption 4.4 all hold, and either Assumption 4.1 or Assumption 4.2 holds. For a fixed x, if $\min_{y \in \mathcal{Y}(x)} \{\max_{i \in \{1,...,k\}} h_i(x, y)\} \leq -d < 0$. Then for any $i \in \{1, ..., k\}$, we have $h_i(x, y_t^*(x)) \leq -m$, where*

$$m := \min \left\{ t \frac{d^2}{4dRL_g + 4RTkL_h}, \frac{d}{2} \right\}. \tag{4.2}$$

Remark 4.1. *Theorem 4.1 still holds when $g(x, y)$ is non-convex in y, and $h_i(x, y)$ is convex in y for any i. In this case, for m computed in (4.2), $h_i(x, y) \leq -m$ holds for any y in the set $y_t^*(x)$.*

Theorem 4.1 provides an interesting way to estimate the negative upper bound at the optimal point $h_i(x, y_t^*(x))$: as long as there exists a tuple (y, d) in such that $y \in \mathcal{Y}(x)$ and $h_i(x, y) < -d$ for any i, it is possible to compute the desired negative upper bound. This enable us to solve the aforementioned difficulty without explicitly assuming a negative upper bound at the optimal point.

The next result computes the strong convexity constant of lower-level penalized objective $\widetilde{g}_t(x, y)$, allowing us to use accelerated gradient methods to obtain the global optimal solutions. The proof is provided in Appendix F.1.

Proposition 4.1. *Suppose the assumptions in Theorem 4.1 are satisfied, then for any fix x and t, the barrier reformulation for the lower-level problem, expressed in (4.1), is $\mu_{\widetilde{g}_t}$-strongly convex in $y \in \{y : h_i(x, y) < 0, \ i = 1, ..., k\}$. Further*

1. *$\mu_{\widetilde{g}_t} = \mu_g$ when Assumption 4.1 holds;*

2. *$\mu_{\widetilde{g}_t} = t\frac{\sigma}{H^2}$ when Assumption 4.2 holds, for some constant H and σ be defined as follows:*

   - *Denote $\sigma(x)$ as the smallest eigenvalue of the following matrix (which is independent of y due to the linearity of the lower-level problem)*

   $$(\nabla_y h_1(x, y), \cdots, \nabla_y h_k(x, y))^\top (\nabla_y h_1(x, y), \cdots, \nabla_y h_k(x, y)).$$

   Then set $\sigma := \min_{x \in \mathcal{X}} \sigma(x)$. We also have $\sigma > 0$;

   - *$H := \sup_{i \in \{1,...,k\}, x \in \mathcal{X}, y \in \mathcal{Y}_x} -h_i(x, y)$. We also have $H < \infty$.*

Remark 4.2. *When Assumption 4.2 holds, i.e. the lower-level problem is a linear program, the constant of strong convexity depends on σ, which is hard to compute for many applications. In practice, it can be made more computable by adding a constraint $h_{k+1}(x, y) = \|y\|^2 \leq R^2$ where R is from Assumption 4.3. Adding this constraint will not affect the original problem because the feasible set of the original problem is already contained within that ball, and we can immediately obtain $\widetilde{g}_t(x, y)$ is $\frac{2t}{R}$-strongly convex.*

Before proceeding to resolve the technical difficulty we discussed before, we need to emphasize the following remark:

Remark 4.3. *The assumed compactness of $\mathcal{X}$ and the Slater condition implies that there exists a positive constant D such that $\{y : h_i(x, y) \leq -D, \ i = 1, ..., k\} \neq \emptyset$ for any x, because $\max_{y \in \mathcal{Y}(x)} h_i(x, y) < 0$ for any $x \in \mathcal{X}$. In our analysis, we need to use D when proving the convergence rate; however, we do not need to know the value of D a priori during the design of the algorithm.*

In view of Theorem 4.1 and Remark 4.3, we can immediately obtain the following global estimate on the upper bound for the constraints $h_i(x, y_t^*(x))$ and $h_i(x, \hat{y}_s)$, which will also be used for the further convergence proof.

Corollary 4.1. *By Remark 4.3 and* **Step 1** *of Algorithm 2, the feasibility check must terminate when d is between $D/2$ to D, thus d_s is greater than $D/2$ for any s. By replacing d with $D/2$ in Theorem 4.1, we have a positive lower bound estimate of m_s, denoted as M. Note that by* **Step 3** *of Algorithm 2, the projection guarantees that $\hat{y}_s \in \mathcal{Y}_{m_s}(x_s)$. Therefore we have $h_i(x_s, \hat{y}_s) \leq -M$ for any s and i. Since $y^*(x_s)$ is also in $\mathcal{Y}_{m_s}(x_s)$ by Theorem 4.1, we have $h_i(x_s, y^*(x_s)) \in \mathcal{Y}_M(x_s)$. This implies $h_i(x_s, y^*(x_s)) \leq -M$ for any s and i.*

In conclusion, $-M$ is the upper bound of both the exact lower-level solution $h_i(x_s, y^*(x_s))$, and the inexact lower-level solution $h_i(x_s, \hat{y}_s)$ at each update point .

We are now ready to analyze the lower-level convergence of Algorithm 2. We first have the following Lipschitzness property for the barrier reformulated lower-level problem.

Proposition 4.2. *Suppose the assumptions in Theorem 4.1 are satisfied, then the following barrier reformulated lower-level objective function*

$$\widetilde{g}_t(x, y) = g(x, y) - t \sum_{i=1}^{k} \log(-h_i(x, y))$$

is $\overline{L}_{\widetilde{g}_t, m}$-Lipschitz smooth with respect to y in the set $\mathcal{Y}_m(x) := \{y : h_i(x, y) \leq -m, \ i = 1, ..., k\}$ for any x, and

$$\overline{L}_{\widetilde{g}_t, m} = \overline{L}_g + \frac{tk\overline{L}_h}{m} + \frac{tkL_h^2}{m^2}. \tag{4.3}$$

The proof of Proposition 4.2 is available in Appendix F.3. Now we state the convergence of the lower-level Algorithm 2.

Theorem 4.2. *Suppose Assumption 4.3(1)(2)(3) and Assumption 4.4 all hold, and either Assumption 4.1 or Assumption 4.2 holds. For the output of Algorithm 2, within $\mathcal{O}(\sqrt{\overline{L}_{\widetilde{g}_t, m_s}/\mu_{\widetilde{g}_t}} \log(1/\epsilon_s))$ number of iterations, we have*

$$\|y_J - y_t^*(x)\| \leq \epsilon_s.$$

Denote $\kappa_s = \overline{L}_{\widetilde{g}_t, m_s}/\mu_{\widetilde{g}_t}$, then κ_s has a upper bound estimate $\kappa = \overline{L}_{\widetilde{g}_t, M}/\mu_{\widetilde{g}_t}$, where $\mu_{\widetilde{g}_t}$ is from Proposition 4.1, $\overline{L}_{\widetilde{g}_t, M}$ can be computed by replacing m in (4.3) with M from Corollary 4.1, and $\overline{L}_{\widetilde{g}_t, m_s}$ can be computed by replacing m in (4.3) with m_s from the **Step 2** *of Algorithm 2.*

The proof follows the standard proof of the convergence of the accelerated proximal/projected gradient descent (see Nesterov (2012)) since the lower objective $\widetilde{g}_t$ is both $\overline{L}_{\widetilde{g}_t, m_s}$ Lipschitz smooth and $\mu_{\widetilde{g}_t}$ strongly convex in the region $\mathcal{Y}_{m_s}(x)$. By Corollary 4.1, m_s has a lower bound estimate M, so $\overline{L}_{\widetilde{g}_t, m_s}$ has a upper bound estimate $\overline{L}_{\widetilde{g}_t, M}$, which means κ_s has a upper bound estimate $\overline{L}_{\widetilde{g}_t, M}/\mu_{\widetilde{g}_t}$, denoted as κ. The proof is complete.

Remark 4.4. *We remark that the feasibility check (Step 1 of Algorithm 2) is achievable by the following scheme: let us assume that a solver is available, which is capable of computing the projection operation of a point onto a non-empty convex set. To determine whether the set $\{y : h_i(x, y) \leq -d, \ i = 1, ..., k\}$ is non-empty, we can employ the solver to project any arbitrary point onto the set.*

17

If this solver can return a solution, then we confirm that the set $\{y : h_i(x,y) \leq -d, \ i = 1, ..., k\}$ is indeed non-empty. Conversely, if the solver can not return a solution, then the set is empty. Therefore we believe that the feasibility check step is no more complicated than a projection onto a convex set.

The upper-level problem is more challenging than the lower-level problem. This difficulty arises because solving the lower-level problem requires only the value d_s at a point x_s, where there exists a y such that $h_i(x_s, y) \leq -d_s$ for any i. In contrast, designing the stepsize for the upper-level problem necessitates the global value D, where for any x, there exists a y such that $h_i(x, y) \leq -D$ for any i. Specifically, for lower-level problem, we only need the value d_s at the point x_s and then estimate the value of m_s in **Step 2** of Algorithm 2, so we can obtain the optimal value within the shrunk feasible set $\{y : h_i(x, y) \leq -m_s, \ i = 1, ..., k\}$. However, when addressing the upper-level problem, we need to estimate the Lipschitz constant of hypergradient, which requires the global upper bound of $h_i(x, y_t^*(x))$, as stated in Corollary 4.1 as $-M$. To obtain the $-M$, the global value D is necessary in view of Corollary 4.1. Yet, we do not know the value of D a priori.

Our approach begins with determining the value of d_s at a specific reference point x_s. Leveraging the Lipschitz continuity assumption of h, it is clear that for any x in the ball $B_{x_s}(d_s/2L_h)$ there exists a point y such that $h_i(x, y) \leq -d_s/2$ for any i. Therefore we can estimate the local upper bound of $h_i(x, y_t^*(x))$ within the ball $B_{x_s}(d_s/2L_h)$. This enables us to design an appropriate stepsize for the upper-level problem. Furthermore, we constrain the stepsize to ensure that next reference point x_{s+1} remains within this ball. The process is shown in Figure 1 and 2.
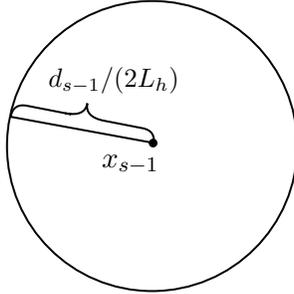

Figure 1: d_{s-1} is an output of Algorithm 2, which indicates that there exists one point y_0 such that $h_i(x_{s-1}, y_0) \leq -d_{s-1}$ for any i. By Lipschitz continuity of $h_i(x, y)$, when x is in the ball $B_{x_{s-1}}(d_{s-1}/(2L_h))$, we have $h_i(x, y_0) \leq -d_{s-1}/2$ for any i. In view of Theorem 4.1, we have a negative upper bound of $h_i(x, y_t^*(x))$ for any $i \in \{1, .., k\}$ and x in the ball $B_{x_{s-1}}(d_{s-1}/(2L_h))$.
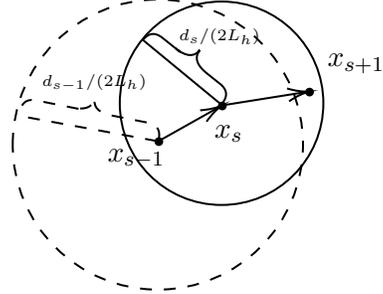
Figure 2: Utilizing the negative upper bound of $h_i(x, y_t^*(x))$ in the ball $B_{x_{s-1}}(d_{s-1}/(2L_h))$ obtained from left figure, we estimate the Lipschitz smoothness constant of $\widetilde{\phi}_t(x)$ in this ball, and design the stepsize η_{s-1}. The design of η_{s-1} ensures that x_s is still in $B_{x_{s-1}}(d_{s-1}/(2L_h))$. We update and get x_s by **Step 2** of Algorithm 1. We also get the ball $B_{x_s}(d_s/(2L_h))$ by repeating the argument in the left figure, then update to the next point x_{s+1} by repeating the above process.

We first establish the following lemma on the Lipschitz properties of the lower-level Hessian, whose proof can be found in Appendix F.4.

Lemma 4.1. *Suppose the assumptions in Theorem 4.1 are satisfied. If $h_i(x_1, y_1) \leq -m$, $h_i(x_2, y_2) \leq -m$ for any i, then $\|\nabla_{yy}^2 \widetilde{g}_t(x_1, y_1) - \nabla_{yy}^2 \widetilde{g}_t(x_2, y_2)\| \leq \overline{\overline{L}}_{\widetilde{g}_t, m}\|(x_1, y_1) - (x_2, y_2)\|$; $\|\nabla_{xy}^2 \widetilde{g}_t(x_1, y_1) -$*

$\nabla^2_{xy}\widetilde{g}_t(x_2, y_2)\| \leq \overline{\overline{L}}_{\widetilde{g}_t, m}\|(x_1, y_1) - (x_2, y_2)\|$, where

$$\overline{\overline{L}}_{\widetilde{g}_t, m} = \overline{\overline{L}}_g + tk\left(\frac{\overline{\overline{L}}_h}{m} + \frac{\overline{L}_h L_h}{m^2} + \frac{2L_h^3}{m^3} + \frac{2L_h \overline{L}_h}{m^2}\right). \tag{4.4}$$

The following lemma informs us that when we know, for a fixed x, the values of the constraints $h_i(x, y)$ for some interior point y, we can evaluate the Lipschitz smoothness constant of $\widetilde{\phi}_t$ within a spherical neighborhood $B_{x_s}(d_s/2L_h)$ of x which can help us to design the stepsize. The proof is available in Appendix F.5.

Lemma 4.2. *Suppose the assumptions in Theorem 4.1 are satisfied. For any fixed x, if there exists $y \in \mathcal{Y}(x)$ such that $h_i(x, y) \leq -d$ for any $i \in \{1, ..., k\}$, then*

1. *For any $x + \Delta x$, where $\|\Delta x\| \leq \frac{d}{2L_h}$, the following inequality holds: $h_i(x + \Delta x, y_t^*(x + \Delta x)) \leq -m^{loc}$, where m^{loc} is a positive constant which can be computed through replacing d by $d/2$ in (4.2);*

2. *The hyperfunction $\widetilde{\phi}_t(x)$ is local $\overline{L}^{loc}_{\widetilde{\phi}_t}$-Lipschitz smooth. That is, the following holds for any x_1, x_2 satisfies $\max\{\|x_1 - x\|, \|x_2 - x\|\} \leq \frac{d}{2L_h}$:*

$$\|\nabla_x \widetilde{\phi}_t(x_1) - \nabla_x \widetilde{\phi}_t(x_2)\| \leq \overline{L}^{loc}_{\widetilde{\phi}_t}\|x_1 - x_2\|.$$

$\overline{L}^{loc}_{\widetilde{\phi}_t}$ can be computed as follows:

$$\overline{L}^{loc}_{\widetilde{\phi}_t} = \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}}(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}) + L_f\frac{\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}}{(\mu_{\widetilde{g}_t})^2}(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}) + L_f\frac{\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}}{\mu_{\widetilde{g}_t}}\right)$$
$$\times \left(1 + \frac{1}{\mu_{\widetilde{g}_t}}(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2})\right),$$

where $\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}$ is from (4.4) by replacing m with m^{loc}, and $\mu_{\widetilde{g}_t}$ is from Proposition 4.1.

Replacing d in Lemma 4.2 with the current iteration d_s from the output of Algorithm 2, we have an estimate of local Lipschitz smoothness constant $\overline{L}_{\widetilde{\phi}_t, s} := \overline{L}^{loc}_{\widetilde{\phi}_t}$ in the ball $B_{x_s}(d_s/(2L_h))$. Now we are able to establish the following lemma on the upper bound of $\overline{L}_{\widetilde{\phi}_t, s}$, whose proof is provided in Appendix F.6:

Lemma 4.3. *Suppose the assumptions in Theorem 4.1 are satisfied. The estimate $\overline{L}_{\widetilde{\phi}_t, s}$ has a upper bound $\overline{L}_{\widetilde{\phi}_t} = \mathcal{O}(1/t^4)$.*

Remark 4.5. *Although $\overline{L}_{\widetilde{\phi}_t, s}$ is bounded by $\overline{L}_{\widetilde{\phi}_t}$, the constant $\overline{L}_{\widetilde{\phi}_t}$ is not computable if we do not know the constant D from Remark 4.3, thus this cannot help us design the stepsize. $\overline{L}_{\widetilde{\phi}_t, s}$ can make sure that the stepsize η_s has a lower bound, a fact which is essential for our main convergence result Theorem 4.3 to hold.*

Now that in practice we are not able to calculate the hypergradient $\nabla_x \widetilde{\phi}_t(x)$ exactly, and in Algorithm 1 we approximate the hypergradient by $\hat{\nabla}_x \widetilde{\phi}_t(x)$. We thus need the following lemma on the approximation error of the hypergradient, whose proof can be found in Appendix F.7.

Lemma 4.4. *Suppose the assumptions in Theorem 4.1 are satisfied. Let x_s be the reference point, and $\hat{y}_s$ is from the output of Algorithm 2. Define the following approximate gradient*

$$\hat{\nabla}_x\widetilde{\phi}_t(x_s) = \nabla_x f(x_s, \hat{y}_s) - \nabla^2_{xy}\widetilde{g}_t(x_s, \hat{y}_s)(\nabla^2_{yy}\widetilde{g}_t(x_s, \hat{y}_s))^{-1}\nabla_y f(x_s, \hat{y}_s). \tag{4.5}$$

Then the gradient approximation error can be upper bounded as $\|\nabla_x\widetilde{\phi}_t(x_s) - \hat{\nabla}_x\widetilde{\phi}_t(x_s)\| \leq \epsilon_s \overline{L}'_{\widetilde{\phi}_t, m_s}$, where m_s is from the output of Algorithm 2 and

$$\overline{L}'_{\widetilde{\phi}_t, m_s} = \overline{L}_f + \overline{L}_f\frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2}\right) + L_f\left(\frac{1}{\mu_{\widetilde{g}_t}}\right)^2\overline{\overline{L}}_{\widetilde{g}_t, m_s}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2}\right) + L_f\frac{1}{\mu_{\widetilde{g}_t}}\overline{\overline{L}}_{\widetilde{g}_t, m_s}, \tag{4.6}$$

where $\overline{\overline{L}}_{\widetilde{g}_t, m_s}$ is from Lemma 4.1 which can be computed by replacing m with m_s.

Finally, we state the convergence of the upper-level problem as follows.

Theorem 4.3. *Suppose Assumption 4.3(1)(2)(3) and Assumption 4.4 all hold, and either Assumption 4.1 or Assumption 4.2 holds. We recall and define the following terms:*

- *$\overline{L}_{\widetilde{\phi}_t, s}$ is computed by replacing d in Lemma 4.2 with d_s from the output of Algorithm 2;*

- *$\hat{\nabla}_x\widetilde{\phi}_t(x_s)$ is from (4.5);*

- *$\overline{L}'_{\widetilde{\phi}_t, m_s}$ is from (4.6);*

- *M is from Corollary 4.1;*

- *$\mu_{\widetilde{g}_t}$ is from Proposition 4.1;*

- *$\overline{L}_{\widetilde{\phi}_t}$ is from Lemma 4.3;*

- *D is from Remark 4.3;*

- *d_s is from the output of Algorithm 2;*

- *$\zeta = \min\left\{1, \frac{D}{4L_h} \times \frac{1}{L_f + \frac{L_f}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + tk\frac{\overline{L}_h}{M} + tk\frac{L_h^2}{M^2}\right)}, \frac{1}{\overline{L}_{\widetilde{\phi}_t}}\right\}.$*

We set the stepsize η_s as

$$\eta_s = \min\left\{1, \frac{d_s}{2L_h\|\hat{\nabla}_x\widetilde{\phi}_t(x_s)\|}, \frac{1}{\overline{L}_{\widetilde{\phi}_t, s}}\right\}, \tag{4.7}$$

and set the stop criterion $\epsilon_s = \epsilon/(4\overline{L}'_{\widetilde{\phi}_t, m_s})$ for the inner loop. Then the the sequence $\{x_s\}$ generated by Algorithm 1 satisfies

$$\min_{s=0,\ldots,S-1}\frac{1}{\eta_s}\|x_s - \mathsf{proj}_{\mathcal{X}}(x_s - \eta_s\nabla_x\widetilde{\phi}_t(x_s))\| \leq \epsilon$$

within $\widetilde{\mathcal{O}}(1/(\zeta\epsilon^2))$ number of iterations.

The proof of Theorem 4.3 is in Appendix F.8.

Remark 4.6. *In terms of the dependency on ϵ, Theorem 4.3 achieves the rate $\widetilde{\mathcal{O}}(\epsilon^{-2})$. However, if higher order (3rd order or even higher for the lower-level objective) smoothness is assumed for both lower- and upper-level objectives, the convergence could be improved to $\mathcal{O}(\epsilon^{-7/4})$ using the techniques proposed in Wang et al. (2024) for the special case of bilevel optimization, namely the minimax saddle point problems. We refer to Wang et al. (2024) and omit the details to keep the conciseness of this work.*

In terms of the dependency on t, from Lemma 4.3 and Theorem 4.3, we can see that $\zeta = \mathcal{O}(t^4)$ with respect to t. For the inner loop, we have that the number of iterations is $\widetilde{\mathcal{O}}(t^{-0.5})$ (more details see Appendix F.9). Therefore, the total number of iterations is $\widetilde{\mathcal{O}}(t^{-4.5})$. Moreover, under the strong convex setting, if we already know the upper bounds of some provably bounded terms a priori, then by adjusting the stepsize, the number of iterations with respect to t can be reduced to $\widetilde{\mathcal{O}}(t^{-1.5})$ (for details, see Appendix C).

In the next section, we discuss how to recover a stationary point for the original lower-level constrained bilevel problem (1.3).

4.3 Asymptotic Convergence to the Original Problem

The convergence result in Theorem 4.3 indicates a non-asymptotic rate of convergence is achievable for solving the barrier reformulated problem (1.4). Based on our previous discussion, achieving non-asymptotic convergence toward the original problem (1.3) is intractable since the hypergradient of the original problem (1.3) can be non-differentiable or even discontinuous (Example 3.1). Therefore we inspect the asymptotic convergence toward the original problem when the **SCSC** assumption is satisfied (which will make the original problem differentiable). In particular, we consider Algorithm 3 listed below, where we iteratively shrink the barrier parameter t in our formulation (1.4) and utilize Algorithm 1 for each t.

Algorithm 3: Hypergradient Based Bilevel Barrier Method for (1.3)

Step 0: Initialization. Given an initial point x_0, $t_0 = 1$, the accuracy level ϵ_0.
for $i = 0, 1, 2, ...$ **do**
    Step 1: Solve the problem with current t. Call Algorithm 1 with initial point x_i,
    accuracy $\epsilon = \epsilon_i$ and $t = t_i$ and other parameters specified as in Theorem 4.3, obtaining
    the output x_{i+1}
    Step 2: Decrease the parameter. Set $\epsilon_{i+1} = \epsilon_i/2$ and $t_{i+1} = t_i/2$

Algorithm 3 is a direct application of Algorithm 1 with shrinking t, which is analogous to the classical path-following scheme. Different from the path-following scheme, we do not have self-concordant barriers, thus we have to solve each of the sub-step to a certain precision for each fix t. We have the following result for the asymptotic convergence of Algorithm 3.

Theorem 4.4. *Assume the same assumptions as Theorem 4.3 and consider the output sequence $\{x_i\}$ of Algorithm 3. Suppose in addition Assumption 4.3(4) holds. If the limit point x^* of $\{x_i\} \subset \mathcal{X}$ is **SCSC** points of original BLO problem (1.3), then x^* is the stationary point of (1.3).*

The proof of Theorem 4.4 is available in Section F.10.

5 Numerical Experiments

In this section, we conduct numerical experiments to verify the effectiveness of the proposed Algorithm 1. Our experiments are conducted on a single PC with Intel Core 12400F CPU and 16GB of RAM. Our code is available at `https://github.com/jxxxxxxt/bilevel_ipm`.

5.1 Generated Strongly Convex Problem

In this experiment, we evaluate the performance of our proposed BFBM algorithm using a class of randomly generated problems. Both the upper-level and lower-level objective functions are designed to be strongly convex, and the constraints are all linear in y. The problem template is given as follows:

$$\min_{x,y} \ f(x,y) = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^\top A \begin{pmatrix} x \\ y \end{pmatrix} + b^\top \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\text{s.t.} \quad y \in \operatorname*{argmin}_{y} g(x,y) = \frac{1}{2} y^\top C y + d^\top y + x^\top D y$$

$$\text{s.t.} \quad x^\top E_i y \le e_i \quad \text{for } i = 1, \ldots, k, \quad -1 \le y_i \le 1 \quad \text{for } i = 1, \ldots, n,$$

where $x \in \mathbb{R}^n$ represents the upper-level decision variables, $y \in \mathbb{R}^n$ represents the lower-level decision variables, $A \in \mathbb{R}^{2n \times 2n}$ and $C \in \mathbb{R}^{n \times n}$ are symmetric positive definite matrices, $b \in \mathbb{R}^{2n}$ and $d \in \mathbb{R}^n$ are vectors, $D \in \mathbb{R}^{n \times n}$ and $E_i \in \mathbb{R}^{n \times n}$ are matrices, $e_i \in \mathbb{R}^+$ are positive numbers for $i = 1, .., k$. All these are randomly generated (see Appendix A).

Experimental Setup and Algorithm Comparison In this experiment, we set the problem dimension to $n = 60$ and the number of primary constraints to $k = 20$. We generated ten problem instances using seeds ranging from 0 to 9 to ensure variability and reproducibility. We allocated a two-minute time budget to each of the three algorithms—BFBM (Our proposed Algorithm), BSG (Giovannelli et al., 2021), and BLOCC (Jiang et al., 2024)—and evaluated them under two stepsize strategies: constant and diminishing. (For details about the parameters, refer to Appendix A.)

After the allotted time, the quality of the obtained solutions (x,y) was evaluated based on four metrics: the *upper-level function value* $f(x,y)$, the *hyperfunction value* $f(x, y_{\text{opt}})$ where y_{opt} is the optimal solution of the lower-level problem given x, the *lower-level optimality gap* $g(x,y) - g(x, y_{\text{opt}})$, and the *total constraint violation* defined as

$$\sum_{i=1}^{k} \max(0, x E_i^{(1)} y - e_i^{(1)}) + \sum_{i=1}^{n} \max(0, y_i - 1) + \sum_{i=1}^{n} \max(0, -y_i - 1),$$

which measures the overall lower-level feasibility of the solution. A higher total constraint violation indicates that the obtained solution deviates more from the feasible region of the lower-level problem. The total constraint violation equals zero if and only if the obtained solution is feasible in the lower-level problem. For each algorithm, the solution with the best hyperfunction value across both stepsize strategies was selected for comparison.

Table 2 compares the solutions obtained by the three algorithms across ten different seeds under a two-minute computational budget per run. Each solution is evaluated in terms of both its hyperfunction value $f(x, y_{\text{opt}})$ and its upper-level function value $f(x,y)$. As shown in the table, our proposed algorithm BFBM achieves the lowest hyperfunction value among the three algorithms in five out of the ten seeds.

Table 2: Evaluate solution qualities using upper-level metrics: Hyperfunction Value (Upper-Level Function Value)

Seed	BSG	BLOCC	BFBM
0	**1.1138 (1.0546)**	2.6983 (1.3556)	1.5372 (1.2014)
1	1.1550 (1.0850)	0.9524 (0.1207)	**0.1657 (-0.0186)**
2	1.9700 (1.8424)	3.1455 (2.1752)	**1.8765 (1.6664)**
3	**2.0662 (2.1330)**	2.7364 (2.5312)	2.1675 (1.9531)
4	**0.6963 (17.6067)**	1.4438 (1.4192)	0.8288 (0.6353)
5	0.7596 (0.7157)	0.8560 (1.4241)	**0.6675 (0.3143)**
6	**1.0723 (18.5577)**	1.2513 (0.2774)	2.2743 (1.8548)
7	0.6451 (1.0544)	**0.3920 (0.5784)**	2.8073 (2.2171)
8	17.2746 (146.8528)	2.8086 (1.1682)	**2.2652 (2.0724)**
9	8.4697 (15.3909)	5.8981 (5.9888)	**1.5066 (1.0964)**

Table 3: Evaluate solution qualities using lower-level metrics: Total Constraint Violation (Optimality Gap)

Seed	BSG	BLOCC	BFBM
0	0.0000(0.0289)	0.0000(0.1354)	0.0000(0.0515)
1	0.0084(0.0218)	0.0000(0.2183)	0.0000(0.0478)
2	0.0000(0.0425)	0.0000(0.1248)	0.0000(0.0643)
3	0.0100(-0.0204)	0.0000(0.0863)	0.0000(0.0443)
4	0.0122(-0.0392)	0.0066(0.0285)	0.0000(0.0527)
5	0.0000(0.0224)	0.0192(0.0099)	0.0000(0.0530)
6	0.0140(-0.0317)	0.0000(0.1670)	0.0000(0.0718)
7	0.0136(-0.0320)	0.0050(0.0171)	0.0000(0.0562)
8	1.2690(-1.0111)	0.0000(0.7070)	0.0000(0.0717)
9	0.0169(-0.0059)	0.0000(0.1192)	0.0000(0.0510)

As shown in Table 3, we compare the solutions obtained by the three algorithms across ten different random seeds in terms of their optimality gaps and total violations. It is crucial to consider only solutions with a total violation of zero as feasible; therefore, the comparison of optimality gaps is meaningful exclusively for these feasible solutions.

Our proposed algorithm, BFBM, consistently achieves a total violation of zero for all ten seeds, ensuring the feasibility of its solutions. Compared to BLOCC, BFBM achieves lower total constraint violation and optimality gap. Relative to BSG, although both exhibit small optimality gaps, BFBM maintains a lower total constraint violation. This stability underscores the reliability and high quality of the solutions generated by the BFBM algorithm.

5.2 Price Setting Problem

In this experiment, we assess the performance of our proposed BFBM algorithm in the context of a price setting problem (see Section 2). In this problem, both the upper-level and lower-level objective functions are linear, and the constraints are all linear in y. Here we consider the case that the lower-level problems are all uncoupled, i.e. they do not depend on T. All the matrices and vectors in (2.1) are randomly generated (see Appendix A).

Experimental Setup and Algorithm Comparison   In this experiment, we set the problem dimension to $n = 60$ and the number of primary constraints to $k = 20$. We generated ten problem instances using seeds ranging from 0 to 9 to ensure variability and reproducibility. We allocated a four-minute time budget to each of the three algorithms—BFBM (Our proposed Algorithm), BSG (Giovannelli et al., 2021), and BLOCC (Jiang et al., 2024)—and evaluated them under diminishing stepsize strategy. (For details about the parameters, refer to Appendix A.)

Table 4: Hyperfunction (Evaluate solution qualities using upper-level metrics: Hyperfunction Value (Upper-Level Function Value))

Seed	BSG	BLOCC	BFBM
0	-0.1279 (-0.0710)	-0.0046 (-0.0047)	**-0.2825 (-0.2750)**
1	-0.1337 (-0.0405)	-0.0034 (-0.0034)	**-0.2455 (-0.2382)**
2	-0.1291 (-0.1417)	-0.0000 (0.0000)	**-0.3629 (-0.3552)**
3	-0.0509 (-0.0112)	-0.0000 (0.0000)	**-0.2950 (-0.2868)**
4	-0.0559 (-0.1086)	-0.0059 (-0.0060)	**-0.3034 (-0.2888)**
5	-0.0499 (-0.0612)	-0.0000 (0.0000)	**-0.2992 (-0.2891)**
6	-0.0045 (-0.0445)	-0.0044 (-0.0043)	**-0.1590 (-0.1506)**
7	0.0000 (0.0000)	-0.0000 (0.0000)	**-0.3070 (-0.2945)**
8	-0.0314 (-0.1053)	-0.0000 (0.0000)	**-0.1790 (-0.1741)**
9	-0.0748 (-0.0603)	-0.0100 (-0.0095)	**-0.3348 (-0.3248)**

Table 5: Evaluate solution qualities using lower-level metrics: Total Constraint Violation (Optimality Gap)

Seed	BSG	BLOCC	BFBM
0	2.1335 (1.8119)	0.0676 (0.0088)	0.0000 (0.0120)
1	4.4754 (-1.0703)	0.0618 (0.0312)	0.0000 (0.0119)
2	3.0165 (0.0653)	0.1464 (0.0024)	0.0000 (0.0120)
3	4.6530 (-0.8170)	0.0875 (-0.0049)	0.0000 (0.0119)
4	3.4429 (-1.0803)	0.1834 (-0.0523)	0.0000 (0.0120)
5	3.7905 (-0.2300)	0.1748 (-0.0246)	0.0000 (0.0119)
6	3.7567 (-1.3199)	0.0917 (0.0293)	0.0000 (0.0120)
7	6.6517 (-2.5038)	0.1058 (0.0149)	0.0000 (0.0120)
8	2.5576 (0.0622)	0.2153 (-0.0415)	0.0000 (0.0120)
9	3.4771 (0.5839)	0.0814 (0.0316)	0.0000 (0.0120)

For the BSG and BLOCC algorithms, we conducted experiment on a regularized version of the price setting problem. In the regularized version, a small regularization term $0.001 \times (\|x\|^2 + \|y\|^2)$ was added to the lower-level objective function $g(T, x, y)$, thereby ensuring that it is strongly convex. From Table 4 and 5, it is evident that for the price-setting problem, the solution obtained by our BFBM algorithm is optimal in terms of hyperfunction quality. In addition, while maintaining strict feasibility (i.e., the total constraint violation is zero), the optimality gap remains consistently low, demonstrating the robustness and efficiency of our method when lower-level problem is a linear program.

6 Conclusion

In this paper, we proposed a barrier function reformulation approach for BLO optimization problems involving lower-level problems with coupled constraints, i.e. the constraints depend on both upper- and lower-level variables. We focused on two specific cases: one where the lower-level objective is strongly convex with convex constraints concerning the lower-level variable, and another where the lower-level problem is a linear program.

By developing a series of new techniques, we demonstrated that the barrier reformulated problem converges to the original bilevel problem in terms of both hyperfunction value and hypergradient. To solve the reformulated problem whose hyperfunction and lower-level problems are not Lipschitz smooth, we designed two effective algorithms: one that guarantees non-asymptotic convergence for the barrier reformulated problem for a fixed t, and the other that converges asymptotically to a stationary point of the original problem if this point is an **SCSC** point (see Definition 3.1). Notably, this is the first work providing convergence guarantees for BLO optimization problems where the lower-level problem is a linear program. For the general case that $g(x, y)$ and $h_i(x, y)$ are all convex in y, we have the convergence for hyperfunction value, and Algorithm 1 still works in this case (see Appendix D).

We conducted experiments on strongly convex and linear lower-level problems with linear inequality constraints. Our algorithm is the only one that ensure the solution is feasible for lower-level problem. For the price-setting problem, it was the most effective algorithm, demonstrating its robustness and effectiveness for bilevel optimization.

Our work establishes a unified framework that operates under minimal assumptions, significantly advancing the methodology in bilevel optimization. Future efforts will focus on extending our algorithms to scenarios with stochastic upper- and lower-level objective functions.

References

Y. Beck, I. Ljubić, and M. Schmidt. A survey on bilevel optimization under uncertainty. *European Journal of Operational Research*, 311(2):401–426, 2023. (Cited on page 7.)

D. P. Bertsekas. Nonlinear programming: 3rd edition. *Journal of the Operational Research Society*, 48(3):334–334, 1997. (Cited on page 76.)

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. (Cited on page 9.)

T. Chen, Y. Sun, Q. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2466–2488. PMLR, 28–30 Mar 2022a. (Cited on page 3.)

Z. Chen, B. Kailkhura, and Y. Zhou. A fast and convergent proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. *ArXiv*, abs/2203.16615, 2022b. URL https://api.semanticscholar.org/CorpusID:247839136. (Cited on page 3.)

D. Drusvyatskiy. Convex analysis and nonsmooth optimization. *University Lecture*, 2020. (Cited on page 40.)

S. Dutta, X. Cheng, and S. Sra. Riemannian bilevel optimization. *arXiv preprint arXiv:2405.15816*, 2024. (Cited on page 4.)

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. (Cited on page 1.)

L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:8026824. (Cited on page 3.)

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018. (Cited on page 1.)

L. L. Gao, J. J. Ye, H. Yin, S. Zeng, and J. Zhang. Moreau envelope based difference-of-weakly-convex reformulation and algorithm for bilevel programs. *ArXiv*, abs/2306.16761, 2023. URL https://api.semanticscholar.org/CorpusID:259286850. (Cited on pages 3 and 4.)

S. Ghadimi and M. Wang. Approximation methods for bilevel programming, 2018. URL https://arxiv.org/abs/1802.02246. (Cited on pages 2, 3, 7, 13, and 39.)

G. Giorgi and C. Zuccotti. A tutorial on sensitivity and stability in nonlinear programming and variational inequalities under differentiability assumptions. 2018. URL https://api.semanticscholar.org/CorpusID:127899202. (Cited on pages 11 and 39.)

T. Giovannelli, G. D. Kent, and L. N. Vicente. Inexact bilevel stochastic gradient methods for constrained and unconstrained lower-level problems. *arXiv preprint arXiv:2110.00604*, 2021. (Cited on pages 3, 4, 22, 24, 31, and 32.)

A. Han, B. Mishra, P. Jawanpuria, and A. Takeda. A framework for bilevel optimization on riemannian manifolds. *arXiv preprint arXiv:2402.03883*, 2024. (Cited on page 4.)

M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023. (Cited on pages 2, 3, 7, 8, and 13.)

K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021. (Cited on pages 1, 3, and 13.)

L. Jiang, Q. Xiao, V. M. Tenorio, F. Real-Rojas, A. Marques, and T. Chen. A primal-dual-assisted penalty approach to bilevel optimization with coupled constraints. *arXiv preprint arXiv:2406.10148*, 2024. (Cited on pages 3, 4, 7, 22, 24, 31, and 32.)

P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021. (Cited on page 3.)

P. Khanduri, I. C. Tsaknakis, Y. Zhang, J. Liu, S. Liu, J. Zhang, and M.-F. Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:260779157. (Cited on pages 2, 3, 4, 7, 8, 10, 11, and 12.)

B. Kohli. Optimality conditions for optimistic bilevel programming problem using convexifactors. *Journal of Optimization Theory and Applications*, 152:632–651, 2012. (Cited on page 2.)

G. Kornowski, S. Padmanabhan, K. Wang, Z. Zhang, and S. Sra. First-order methods for linearly constrained bilevel optimization, 2024. URL https://arxiv.org/abs/2406.12771. (Cited on pages 3, 4, and 7.)

J. Kwon, D. Kwon, S. J. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. *ArXiv*, abs/2301.10945, 2023. URL https://api.semanticscholar.org/CorpusID:256274518. (Cited on page 3.)

J. Kwon, D. Kwon, S. Wright, and R. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation, 2024. URL https://arxiv.org/abs/2309.01753. (Cited on page 3.)

M. Labbé and A. Violin. Bilevel programming and price setting problems. *Annals of operations research*, 240:141–169, 2016. (Cited on pages 1 and 6.)

C. Li, S. Zeng, Z. Liao, J. Li, D. Kang, A. Garcia, and M. Hong. Joint demonstration and preference learning improves policy alignment with human feedback. *arXiv preprint arXiv:2406.06874*, 2024a. (Cited on page 1.)

J. Li and S. Ma. Riemannian bilevel optimization. *arXiv preprint arXiv:2402.02019*, 2024. (Cited on page 4.)

J. Li, S. Zeng, H.-T. Wai, C. Li, A. Garcia, and M. Hong. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *arXiv preprint arXiv:2405.17888*, 2024b. (Cited on page 1.)

J. Liu, Y. Fan, Z. Chen, and Y. Zheng. Pessimistic bilevel optimization: a survey. *International Journal of Computational Intelligence Systems*, 11(1):725–736, 2018. (Cited on page 2.)

Z. Lu and S. Mei. First-order penalty methods for bilevel optimization, 2024. URL https://arxiv.org/abs/2301.01716. (Cited on pages 3 and 4.)

D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015a. (Cited on page 3.)

D. Maclaurin, D. K. Duvenaud, and R. P. Adams. Gradient-based hyperparameter optimization through reversible learning. *ArXiv*, abs/1502.03492, 2015b. URL https://api.semanticscholar.org/CorpusID:8540522. (Cited on page 1.)

A. Mehra and J. Hamm. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:207847954. (Cited on page 3.)

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125 – 161, 2012. URL https://api.semanticscholar.org/CorpusID:18206201. (Cited on page 17.)

A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999, 2018. URL https://api.semanticscholar.org/CorpusID:4587331. (Cited on page 3.)

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, 2016. URL https://api.semanticscholar.org/CorpusID:552516. (Cited on pages 2 and 3.)

M. Schmidt and Y. Beck. A gentle and incomplete introduction to bilevel optimization, 2023. (Cited on page 7.)

H. Shen, Q. Xiao, and T. Chen. On penalty-based bilevel gradient descent method, 2023. URL https://arxiv.org/abs/2302.05185. (Cited on pages 3, 4, and 13.)

W. Shi, Y. Chang, and B. Gu. Double momentum method for lower-level constrained bilevel optimization. *arXiv preprint arXiv:2406.17386*, 2024. (Cited on page 3.)

A. J. Silvério, R. D. S. Couto, M. E. M. Campista, and L. H. M. K. Costa. A bi-objective optimization model for segment routing traffic engineering. *Annals of Telecommunications*, 77:813 – 824, 2022. URL https://api.semanticscholar.org/CorpusID:247004593. (Cited on page 1.)

B. Stadie, L. Zhang, and J. Ba. Learning intrinsic rewards as a bi-level optimization problem. In *Conference on Uncertainty in Artificial Intelligence*, pages 111–120. PMLR, 2020. (Cited on page 1.)

I. Tsaknakis, M. Hong, and S. Zhang. Minimax problems with coupled linear constraints: Computational complexity and duality. *SIAM Journal on Optimization*, 33(4):2675–2702, 2023a. (Cited on page 5.)

I. Tsaknakis, P. Khanduri, and M. Hong. An implicit gradient method for constrained bilevel problems using barrier approximation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023b. (Cited on pages 3, 5, 7, 8, 12, and 13.)

N. Wang, J. Zhang, and S. Zhang. Efficient first order method for saddle point problems with higher order smoothness. *SIAM Journal on Optimization*, 34(2):1374–1401, 2024. (Cited on page 21.)

Q.-W. Xiao, H. Shen, W. Yin, and T. Chen. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259104711. (Cited on page 4.)

S. Xu and M. Zhu. Efficient gradient approximation method for constrained bilevel optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12509–12517, 2023. (Cited on pages 3, 4, 7, and 8.)

J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021. (Cited on pages 2 and 3.)

Y. Yang, P. Xiao, and K. Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ Complexity in Hessian/Jacobian-free Stochastic Bilevel Optimization. *arXiv preprint arXiv:2312.03807*, 2023. (Cited on page 3.)

W. Yao, H. Yin, S. Zeng, and J. Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions, 2024. URL https://arxiv.org/abs/2406.01992. (Cited on pages 3, 4, and 8.)

S. Zeng, C. Li, A. Garcia, and M. Hong. When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on page 1.)

Appendix

A   Details of Experiments

A.1   Generated Strongly Convex Problem

Data Generation Process   The problem data is generated using the following procedure:

1. **Random Seed Initialization:** A fixed random seed is set using `np.random.seed(seed)` to guarantee reproducibility of the results across different runs.

2. **Matrix Generation:**

   - **Matrix A:** Generated as a symmetric positive definite matrix:
   $$A = 0.5 \times (A_{\text{init}} \times A_{\text{init}}^{\top}) + I_{2n}.$$
   Here, each entry of the initial random matrix A_{init} is drawn from a standard normal distribution $\mathcal{N}(0, 1)$;

   - **Matrix C:** Generated as a symmetric positive definite matrix:
   $$C = 0.5 \times (C_{\text{init}} \times C_{\text{init}}^{\top}) + I_n.$$
   Each entry of the initial random matrix C_{init} is drawn from a standard normal distribution $\mathcal{N}(0, 1)$;

   - **Matrix D:** Generated as a random matrix:
   $$D = \text{randn}(n, n).$$
   Each entry of D is independently drawn from a standard normal distribution $\mathcal{N}(0, 1)$.

3. **Vector Generation:**

   - **Vector b:** Generated as a random vector:
   $$b = \text{randn}(2n).$$
   Each entry of b is drawn from a normal distribution $\mathcal{N}(0, 1)$ and then scaled by a factor of 10 to increase the magnitude of the coefficients;

   - **Vector d:** Generated as a random vector:
   $$d = \text{randn}(n).$$
   Each entry of d is drawn from a normal distribution $\mathcal{N}(0, 1)$ and scaled by a factor of 10.

4. **Constraints Generation:**

   - **Coupled Constraints:** A total of k coupled linear constraints are imposed on the lower-level variables:
   $$x^{\top} E_i y \leq e_i, \text{ for } i = 1, \ldots, k.$$
     - Each matrix $E_i \in \mathbb{R}^{n \times n}$ is generated with entries drawn independently from a standard normal distribution $\mathcal{N}(0, 1)$;
     - Each scalar $e_i \in \mathbb{R}$ is drawn uniformly from the interval $[0, 1]$, i.e., $e_i \sim \mathcal{U}(0, 1)$;
   - **Box Constraints:** Simple Box constraints are imposed to make sure the lower-level feasible set is compact.
   $$-\mathbf{1} \leq y \leq \mathbf{1}.$$

Implementation Details

1. **BSG Algorithm Giovannelli et al. (2021):**

   - For inner loop, we use a max-min solver (Algorithm 2 in Jiang et al. (2024)) to obtain the lower-level optimal solution y and Lagrange multiplier z_I;
   - We choose the initial point $x_{\text{init}} = 0$, $y_{\text{init}} = 0$, $z_{I,\text{init}} = 0$; For the main loop, the stepsize is set to $\alpha_k = 2 \times 10^{-6}$ or $5 \times 10^{-5}/\sqrt{k+1}$; For the max-min solver, the stepsizes for the first and second loops are set to $\eta_1 = 10^{-3}$ and $\eta_2 = 0.02$ respectively; The stopping criteria for the first and second loops in the max-min solver are $\epsilon_1 = 0.01$ and $\epsilon_2 = 0.01$.

2. **BLOCC Algorithm Jiang et al. (2024):**

   - We choose the initial point $x_{\text{init}} = 0$, $y_{g,\text{init}} = 0$, $y_{F,\text{init}} = 0$, $\mu_{g,\text{init}} = 0$, $\mu_{F,\text{init}} = 0$, and the penalty coefficient $\gamma = 10$; For the main loop, the stepsize is set to $\alpha_k = 2 \times 10^{-5}$ or $10^{-3}/\sqrt{k+1}$; For both max-min solvers, the stepsizes for the first and second loops are set to $\eta_1 = 10^{-3}$ and $\eta_2 = 0.5$ respectively; The stopping criteria for the first and second loops in both max-min solvers are $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.05$.

3. **BFBM Algorithm (Our proposed Algorithm):**

   - For inner loop we use Netwon method;
   - We set $t = 0.01$, and $M = 0.1$, which means we shrink the lower-level feasible set to $\{y | h_i(x, y) \leq -10^{-3} \text{ for } i = 1, ..., k\}$;
   - We choose the initial point $x_{\text{init}} = 0$, $y_{\text{init}} = 0$; For the main loop, the stepsize is set to $\alpha_k = 2 \times 10^{-5}$ or $5 \times 10^{-4}/\sqrt{k+1}$; For inner loop the stepsizes is set to $\eta = 0.1$; The stopping criteria for inner loop is $\epsilon_y = 0.01$.

A.2   Price Setting Problem

Data Generation Process   The problem data is generated using the following procedure:

1. **Random Seed Initialization:** A fixed random seed is set using `np.random.seed(seed)` to guarantee reproducibility of the results across different runs.

2. **Matrix Generation:**

   - **Matrix A_1 and A_2:** Generated as a matrix with entries drawn from a normal distribution $\mathcal{N}(0, 1)$ and then made non-negative by taking the absolute value. Additionally, a correction is applied: if all values in a row of A_1 or A_2 are less than 1, we set the maximum value in that row to be 1. This ensures that each city has at least one road leading to it. Finally, we multiply each element of A_2 by 0.2 to simulate the significantly lower carrying capacity of untaxed roads compared to taxed roads.

3. **Vector Generation:**

   - **Vector b, c_1 and c_2:** Generated as random vectors, with each entry drawn from $\mathcal{N}(0, 1)$, and then made non-negative by taking the absolute value.

Implementation Details

1. **BSG Algorithm Giovannelli et al. (2021):**

   - For inner loop, we use a max-min solver (Algorithm 2 in Jiang et al. (2024)) to obtain the lower-level optimal solution y and Lagrange multiplier z_I;

   - We choose the initial point $T_{\text{init}} = 0$, $x_{\text{init}} = 0$, $y_{\text{init}} = 0$, $z_{I,\text{init}} = 0$; For the main loop, the stepsize is set to $\alpha_k = 0.01/\sqrt{k+1}$; For the max-min solver, the stepsizes for the first and second loops are set to $\eta_1 = 2^{-5}$ and $\eta_2 = 30$ respectively; The stopping criteria for the first and second loops in the max-min solver are $\epsilon_1 = 5 \times 10^{-4}$ and $\epsilon_2 = 1$.

2. **BLOCC Algorithm Jiang et al. (2024):**

   - We choose the initial point $T_{\text{init}} = 0$, $x_{g,\text{init}} = 0$, $y_{g,\text{init}} = 0$, $x_{F,\text{init}} = 0$, $y_{F,\text{init}} = 0$, $\mu_{g,\text{init}} = 0$, $\mu_{F,\text{init}} = 0$, and the penalty coefficient $\gamma = 2$; For the main loop, the stepsize is set to $\alpha_k = 0.01/\sqrt{k+1}$; For both max-min solvers, the stepsizes for the first and second loops are set to $\eta_1 = 2 \times 10^{-5}$ and $\eta_2 = 30$ respectively; The stopping criteria for the first and second loops in both max-min solvers are $\epsilon_1 = 10^{-4}$ and $\epsilon_2 = 1$.

3. **BFBM Algorithm (Our proposed Algorithm):**

   - For inner loop we use Newton method;

   - We set $t = 10^{-4}$, and $M = 0.1$, which means we shrink the lower-level feasible set to $\{y|h_i(x,y) \leq -10^{-5} \text{ for } i = 1, ..., k\}$;

   - We choose the initial point $x_{\text{init}} = 0$, $y_{\text{init}} = 0$; For the main loop, the stepsize is set to $\alpha_k = 0.02/\sqrt{k+1}$; For inner loop the stepsizes is set to $\eta = 0.01$; The stopping criteria for inner loop is $\epsilon_{xy} = 5 \times 10^{-4}$.

B   Preliminaries

B.1   Notations

- Denote $[\lambda_1(x,y), ..., \lambda_k(x,y)] := \operatorname{argmax}_\lambda g(x,y) + \sum_{i=1}^k \lambda_i h_i(x,y)$ be the Lagrange multipliers at point (x,y), where $y \in y^*(x)$. If $y^*(x)$ is unique, or $\lambda_i(x,y)$ is independent of the choice of $y \in y^*(x)$, for simplicity, we write $\lambda_i(x,y) = \lambda_i(x)$, and, under this case, define $\mathcal{I}^*(x) := \{i : \lambda_i(x) > 0\}$ be the set of index that corresponding multiplier is positive;

- $\mathcal{Y}_m(x) := \{y : h_i(x,y) \leq -m, \ i = 1, ..., k\}$ is the shrinked feasible set;

- $\hat{y}_s$ is approximation of $y^*(x_s)$ from the output of Algorithm 2;

- D introduced in Remark 4.3 is a constant that for any x there exists y satisfies $h_i(x,y) \leq -D$ for any i;

- $\mu_{\widetilde{g}_t}$ introduced in Proposition 4.1 is the strongly convexity constant of $\widetilde{g}_t(x,y)$ in y;

- M defined in Corollary 4.1 is the positive lower bound of $-h_i(x,y^*(x))$ and $-h_i(x_s, \hat{y}_s)$;

- $\overline{L_{\widetilde{g}_t,m}}$ introduced in Proposition 4.2 is the Lipschitz smoothness constant of $\widetilde{g}_t(x,y)$ in y in the set $\mathcal{Y}_m(x)$;

- $\overline{\overline{L}}_{\widetilde{g}_t,m}$ introduced in Proposition 4.1 is the Lipschitz constant of Hessian for $\widetilde{g}_t(x,y)$ in the set $\mathcal{Y}_m(x)$;

- κ_s utilized in Theorem 4.2 is the condition number of the inner loop at the s-th step of the outer loop;

- $\overline{L}_{\widetilde{\phi}_t,s}$ studied in Lemma 4.3 is the estimate of local Lipschitz smoothness constant at the s-th step of the outer loop;

- $\overline{L}_{\widetilde{\phi}_t}$ introduced in Lemma 4.3 is the global upper bound of $\overline{L}_{\widetilde{\phi}_t,s}$;

- $\hat{\nabla}_x\widetilde{\phi}_t(x_s) := \nabla_x f(x,y) + \nabla^2_{xy}\widetilde{g}_t(x_s,\hat{y}_s)(\nabla^2_{yy}\widetilde{g}_t(x_s,\hat{y}_s))^{-1}\nabla_y f(x_s,\hat{y}_s)$ is the approximation of the hypergradient $\nabla_x\widetilde{\phi}_t(x_s)$ at the s-th step of the outer loop;

- $\overline{L}'_{\widetilde{\phi}_t,m_s}$ studied in Lemma 4.4 is the approximation error between $\hat{\nabla}_x\widetilde{\phi}_t(x_s)$ and $\nabla_x\widetilde{\phi}_t(x_s)$, i.e. $\|\hat{\nabla}_x\widetilde{\phi}_t(x_s) - \nabla_x\widetilde{\phi}_t(x_s)\| \le \overline{L}'_{\widetilde{\phi}_t,m_s}\|y^*(x_s) - \hat{y}_s\|$;

- $\overline{L}'_{\widetilde{\phi}_t}$ given in Lemma F.1 is the upper bound of $\overline{L}'_{\widetilde{\phi}_t,m_s}$.

B.2   Some basic lemmas

Lemma B.1. *Suppose Assumption 3.1(1), 3.1(2) and Assumption 3.2 hold, then*

$$\|y^*_t(x) - y^*(x)\| \le \sqrt{\frac{2}{\mu_g}kt}. \tag{B.1}$$

Proof. Since $g(x,y)$ is μ_g-strongly convex in y, we have

$$g(x,y) \ge g(x,y^*(x)) + \nabla_y g(x,y^*(x))(y - y^*(x)) + \frac{\mu_g}{2}\|y - y^*(x)\|^2. \tag{B.2}$$

By the optimal condition, $\nabla_y g(x,y^*(x))(y - y^*(x)) \ge 0$ for any feasible point y. Replacing y by $y^*_t(x)$ in Equation (B.2), we get

$$\frac{\mu_g}{2}\|y^*_t(x) - y^*(x)\|^2 \le g(x,y^*_t(x)) - g(x,y^*(x)) - \nabla_y g(x,y^*(x))(y^*_t(x) - y^*(x))$$
$$\le g(x,y^*_t(x)) - g(x,y^*(x)).$$

From lemma 3.1, we know $g(x,y^*_t(x)) - g(x,y^*(x)) \le kt$, which implies

$$\|y^*_t(x) - y^*(x)\| \le \sqrt{\frac{2}{\mu_g}kt}. \tag{B.3}$$

$\square$

Lemma B.2. *Suppose Assumption 3.1 holds, and either Assumption 3.3 or 3.2 holds, then $y^*_t(x)$ continuously depends on t for any fixed x.*

Proof. Denote $G_x(t, y) = g_t(x, y)$. We fix a x. Note that

$$\nabla_y(\nabla_y G_x(t, y)) = \nabla^2_{yy}\widetilde{g}_t(x, y) \succ 0$$

for any t and y, and $\nabla_y G_x(t, y_t^*(x)) = 0$. By implicit function theorem, we can find a neighborhood $U_t \times U_{y_t^*(x)}$ of $(t, y_t^*(x))$ such that there exists a continuous function $Y(t)$ satisfies $\nabla_y G_x(t, y) = 0$ if and only if $y = Y(t)$. Since G is strongly convex in y, $\nabla_y G_x(t, y) = 0$ means $y = y_t^*(x)$. Therefore, by the uniqueness of the optimal solution, locally we have $y_t^*(x) = Y(t)$ continuously depending on t. As we already know $y_t^*(x)$ is a function of t, we conclude that globally, i.e. for any t, $y_t^*(x)$ continuously depends on t. $\qquad\square$

Lemma B.3. *Suppose Assumption 3.1 holds, and either Assumption 3.2 or Assumption 3.3 holds, then the Lagrange multiplier $\lambda_i(x)$ is continuous for any i. Furthermor, since $\mathcal{X}$ is compact, $\lambda_i(x)$ has upper bound independent of i and x.*

Proof. When $g(x, y)$ is strongly convex in y and $h_i(x, y)$ is convex in y, then $y^*(x)$ is unique. If $\lambda_i(x)$ is not continuous for some point x_0 and some index i_0, then we can find a sequence $\{x_j\}_{j=1}^\infty$ such that $\lim_{j\to\infty} x_j = x_0$, and $\lim_{j\to\infty} \lambda_{i_0}(x_j) \neq \lambda_{i_0}(x_0)$, or does not exist.

Case 1: If $\lim_{j\to\infty} \lambda_{i_0}(x_j)$ does not exist. We consider separately the cases where $\lambda_{i_0}(x_j)$ is uniformly bounded and where there exists an unbounded subsequence.

If $\{\lambda_{i_0}(x_j)\}_{j=1}^\infty$ has an unbounded subsequence, still denoted as $\{\lambda_{i_0}(x_j)\}_{j=1}^\infty$, i.e. $\lim_{j\to\infty} |\lambda_{i_0}(x_j)| = \infty$, then set $\mu_j^* = \max_i |\lambda_i(x_j)|$. By the unboundedness assumption, $\lim_{j\to\infty} \mu_j^* = \infty$. We can choose a sub-sequence of $\{x_j\}_{j=1}^\infty$, still denoted as $\{x_j\}$, such that $i^* = \arg\max_i |\lambda_i(x_j)|$ for some fixed index i^*. Since $|\lambda_i(x_j)|/\mu_j^* \leq 1$ for any i and j, we can choose a subsequence, still denoted as $\{x_j\}$, such that $\lim_{j\to\infty} \lambda_i(x_j)/\mu_j^*$ exists for any i. Denote $\mu_i(x_0) := \lim_{j\to\infty} \lambda_i(x_j)/\mu_j^*$. It is clear that $\mu_{i^*}(x_0) = 1$. Replacing x with x_j in the KKT condition equation

$$\nabla_y g(x, y^*(x)) + \sum_{i=1}^k \lambda_i(x) \nabla_y h_i(x, y^*(x)) = 0, \tag{B.4}$$

dividing by μ_j^*, and letting j approach infinity, we get

$$\sum_{i=1}^k \mu_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) = 0. \tag{B.5}$$

We will show that $\mu_i(x_0) \neq 0$ means i is active at x_0. If $\mu_i(x_0) \neq 0$, then $\lambda_i(x_j) \neq 0$ except for at most finitely many j. So the index i is active at the point x_j, which implies $h_i(x_j, y^*(x_j)) = 0$. Thus, $h_i(x_0, y^*(x_0)) = \lim_{j\to\infty} h_i(x_j, y^*(x_j)) = 0$, implying that i is active at the point x_0. (B.5) can be written as

$$\sum_{i\in\mathcal{I}^*(x_0)} \mu_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) = 0,$$

where $\mathcal{I}^*(x_0)$ is the set of active index at x_0. By LICQ assumption, $\nabla_y h_i(x_0, y^*(x_0))$ are linear independent for $i \in \mathcal{I}^*(x_0)$. Note that $\mu_{i^*}(x_0) = 1 \neq 0$, this contradicts the LICQ assumption. Therefore, we know that $\{\lambda_{i_0}(x_j)\}_{j=1}^\infty$ is uniform bounded.

Case 2: If $\lim_{j\to\infty}\lambda_{i_0}(x_j)$ exists but not equals $\lambda_{i_0}(x_0)$, we will show that this also contradicts LICQ assumption. Case 1 demonstrates that $\lambda_i(x_j)$ has a uniform bound for any i and j. Thus, we can choose a subsequence of $\{x_j\}_{j=1}^\infty$, still denoted as $\{x_j\}_{j=1}^\infty$, such that $\lim_{j\to\infty}\lambda_i(x_j) =: \mu_i(x_0)$ exists for any i. Replacing x with x_j in (B.4), and letting j approach infinity, we get

$$\nabla_y g(x_0, y^*(x_0)) + \sum_{i=1}^k \mu_i(x_0)\nabla_y h_i(x_0, y^*(x_0)) = 0. \tag{B.6}$$

Replacing x with x_0 in the KKT condition (B.4), and subtracting (B.6), we obtain

$$\sum_{i=1}^k (\lambda_i(x_0) - \mu_i(x_0))\nabla_y h_i(x_0, y^*(x_0)) = 0.$$

If $\lambda_i(x_0) - \mu_i(x_0) \neq 0$, then either $\lambda_i(x_0) = 0$ or $\lambda_i(x_0) \neq 0$. $\lambda_i(x_0) = 0$ means $\mu_i(x_0) \neq 0$, which implies i is active at point x_0 similar as the argument in Case 1. $\lambda_i(x_0) \neq 0$ means that i is active at point x_0. Therefore, $\lambda_i(x_0) - \mu_i(x_0) \neq 0$ always means is active at point x_0. Note that $\lambda_{i_0}(x_0) - \mu_{i_0}(x_0) \neq 0$, which contradicts the LICQ assumption.

When $g(x, y)$ and $h_i(x, y)$ are linear in y for any i, then by Remark 3.2, we know that the multipliers $\lambda_i(x)$ are well-defined in the sense that is independent of $y \in y^*(x)$. We can also denote $\nabla_y g(x, y)$ as $w(x)$, and $\nabla_y h_i(x, y^*(x))$ as $v_i(x)$ since $\nabla_y g(x, y)$ and $\nabla_y h_i(x, y)$ is independent of y. Then the proof is the same as the strongly convex case. $\square$

C Improved Convergence Rate

In this section, we will demonstrate how to modify the choice of stepsize so that the convergence rate of Algorithm 1 with respect to t can be improved to $\widetilde{\mathcal{O}}(t^{-1.5})$ under strongly convex case. However, it should be noted that we need to utilize the upper bounds of certain terms. We will prove that these terms are indeed bounded, but these bounds are not computable in practice.

Specifically, in Lemma 4.2, when estimating the local Lipschitz constant, we frequently evaluate the term $\nabla_{xy}\widetilde{g}_t(x, y)(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}$. For the upper bound estimation of $(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}$, we directly employ the strongly convexity constant of $\nabla_{yy}\widetilde{g}_t(x, y)$ in y from Proposition 4.1. However, it is important to note that, under strongly convex case, when $\nabla_{xy}\widetilde{g}_t(x, y)$ explodes, which is because h approaches zero, $\nabla_{yy}\widetilde{g}_t(x, y)$ also explodes. Therefore, using the strong convexity of $\nabla_{yy}\widetilde{g}_t(x, y)$ for the upper bound estimation of $(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}$ results in significant losses. This leads to an underestimate of the dependence of $\nabla_{xy}\widetilde{g}_t(x, y)(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}$ on t.

We assume that all values of t are less than or equal to 1. This assumption is justified, as in practical applications, we often choose a very small t. To improve the estimate, we establish the following result:

Lemma C.1. *Suppose Assumption 4.3, Assumption 4.4, and Assumption 4.1 hold, i.e. under the strongly convex setting. Then for any $0 < t \leq 1$ and $x \in \mathcal{X}$, the following holds*

$$\|\nabla_x y_t^*(x)\| = \|(\nabla_{yy}\widetilde{g}_t(x, y_t^*(x)))^{-1}\nabla_{yx}\widetilde{g}_t(x, y_t^*(x))\| \leq J_1$$

for some constant $J_1 > 0$.

The proof is provided in Section F.11. In addition, we also need the following technical results:

Lemma C.2. *Suppose Assumptions in Lemma C.1 are satisfied. Then for any $0 < t \le 1$, any $x \in \mathcal{X}$ and index $\hat{j}$ we have*

$$\left\| (\nabla_{yy}\widetilde{g}_t(x))^{-1} \frac{t\nabla_y h_{\hat{j}}(x, y_t^*(x))}{(-h_{\hat{j}}(x, y_t^*(x)))^2} \right\| \le J_2$$

for some constant $J_2 > 0$.

Lemma C.3. *Suppose Assumptions in Lemma C.1 are satisfied. There exists a constant β such that for any $0 < t \le 1$, x in $\mathcal{X}$ and $y \in int\mathcal{Y}(x)$, which is the interior of $\mathcal{Y}(x)$, such that $\|y - y^*(x)\| \le \beta$, we have $\|(\nabla_{yy}\widetilde{g}_t(x,y))^{-1}\nabla_{yx}\widetilde{g}_t(x,y)\| \le J_3$, for some constant $J_3 > 0$.*

Lemma C.2 is a byproduct of Lemma C.1. We provide the proof in Section F.12. The proof of Lemma C.3 utilizes some technical details from Lemma C.1, as discussed in Section F.13. Now we can estimate the improved local Lipschitz constant by the boundedness we proved before.

Lemma C.4. *Suppose the assumptions in Lemma C.1 are satisfied. For any $0 < t \le 1$, and any fixed x, if there exists $y \in \mathcal{Y}(x)$ such that $h_i(x,y) \le -d$ holds for any $i \in \{1,...,k\}$, then the hyperfunction $\widetilde{\phi}_t(x)$ is $\overline{L}_{\widetilde{\phi}_t}^{loc}$-Lipschitz smooth in the ball $B_x(\min\{\frac{d}{2L_h}, \frac{m}{2L_h(1+J_1)}\})$ which centered at x with radius $\min\{\frac{d}{2L_h}, \frac{m}{2L_h(1+J_1)}\}$, where J_1 is from Lemma C.1 and m is computed by (4.2). That is, the following holds for any x_1, x_2 satisfies $\max\{\|x_1 - x\|, \|x_2 - x\|\} \le \min\{\frac{d}{2L_h}, \frac{m}{2L_h(1+J_1)}\}$:*

$$\|\nabla_x\widetilde{\phi}_t(x_1) - \nabla_x\widetilde{\phi}_t(x_2)\| \le \overline{L}_{\widetilde{\phi}_t}^{loc}\|x_1 - x_2\|.$$

$\overline{L}_{\widetilde{\phi}_t}^{loc}$ *can be computed as follows*

$$\overline{L}_{\widetilde{\phi}_t}^{loc} = J_1\left(\overline{L}_f + \overline{L}_f J_1 + L_f J_1 \left(\frac{\overline{\overline{L}}_g}{\mu_g} + \frac{tk\overline{\overline{L}}_h}{\mu_g m^{loc}} + \frac{tk\overline{L}_h L_h}{\mu_g (m^{loc})^2} + 2\frac{tkL_h\overline{L}_h}{\mu_g (m^{loc})^2} + J_2\frac{kL_h^2}{m^{loc}} + J_2\frac{3kL_h^2}{m^{loc}}\right) + \overline{\overline{L}}_{\widetilde{g}, m^{loc}} J_1\right).$$

Here m^{loc} is a positive constant which can be computed through replacing d with $d/2$ in (4.2), and $\mu_{\widetilde{g}_t}$ is from Lemma 4.1, J_1 is from Lemma C.1, J_2 is from Lemma C.2. $\overline{L}_{\widetilde{\phi}_t}^{loc}$ has a upper bound $\overline{L}_{\widetilde{\phi}_t} = \mathcal{O}(\frac{1}{t})$.

The proof of Lemma C.4 can be found in Section F.14. Given our refined estimation of the Lipschitz constant, we are now able to optimize the stepsize selection. This adjustment significantly enhances the convergence rate with respect to t, leading to the following convergence results, whose proof is in (F.15):

Theorem C.1. *Suppose Assumption 4.3, Assumption 4.4, and Assumption 4.1 hold, i.e. under the strongly convex setting. We recall and define the following terms:*

- *β is from Lemma C.3;*

- *$\overline{L}_{\widetilde{\phi}_t,s}$ is computed by replacing d in Lemma C.4 with d_s from the output of Algorithm 2;*

- *$\overline{L}'_{\widetilde{\phi}_t,m_s}$ is from (4.6);*

- *$\hat{\nabla}_x\widetilde{\phi}_t(x_s)$ is from (4.5);*

- *J_1 is from Lemma C.1;*

- J_3 is from Lemma C.3;

- M is from Corollary 4.1;

- D is from Remark 4.3;

- m_s, d_s are from the output of Algorithm 2;

- $\mu_{\widetilde{g}_t}$ is from Proposition 4.1;

- $\overline{L}_{\widetilde{\phi}_t}$ is from Lemma C.4;

- $\zeta = \min\left\{1, \frac{D}{4L_h L_f(1+J_3)}, \frac{M}{4L_h L_f(1+J_1)(1+J_3)}, \frac{1}{\overline{L}_{\widetilde{\phi}_t}}\right\}$.

Assume $0 < t \leq \min\{\frac{\mu_g \beta^2}{2k}, 1\}$. We set the stepsize η_s as

$$\eta_s = \min\{1, \frac{d_s}{2L_h} \times \frac{1}{\|\hat{\nabla}_x \tilde{\phi}(x_s)\|}, \frac{m_s}{2L_h(1+J_1)} \times \frac{1}{\|\hat{\nabla}_x \tilde{\phi}(x_s)\|}, \frac{1}{\overline{L}_{\tilde{\phi}_t,s}}\}, \tag{C.1}$$

and set the stop criterion $\epsilon_s = \min\{\epsilon, \beta/2\}/\overline{L}'_{\tilde{\phi}_t, m_s}$ for the inner loop. Then the sequence $\{x_s\}$ generated by Algorithm 1 satisfies

$$\min_{s=0,\ldots,S-1} \frac{1}{\eta_s}\|x_s - \mathsf{proj}_{\mathcal{X}}(x_s - \eta_s \nabla_x \tilde{\phi}(x_s))\| \leq \epsilon$$

within $\widetilde{\mathcal{O}}(1/(\zeta\epsilon^2))$ number of iterations.

Remark C.1. *From Lemma C.4 and Theorem C.1, we can see that $\zeta = \mathcal{O}(t)$ with respect to t. Note that for inner loop we have $\kappa = \widetilde{\mathcal{O}}(t^{-0.5})$, the total number of iterations is $\widetilde{\mathcal{O}}(t^{-1.5})$.*

D Convex Case

When $g(x,y)$ is convex in y, some intrinsic difficulties arise in attempting to establish a connection between the barrier reformulation and the original problem. First, we will provide some counterexamples to illustrate these issues.

Before the discussion, we establish the following basic assumption for convex case:

Assumption D.1. *The following holds*

1. *$f(x,y)$ is one time and $g(x,y)$, $h_i(x,y)$ are two times continuously differentiable for every $x \in \mathcal{X}, y \in \mathcal{Y}(x)$, and $i \in \{1, ..., k\}$;*

2. *$\mathcal{X}$ is convex and compact, and for any $x \in \mathcal{X}$ there exists $y \in \mathcal{Y}(x)$ such that $h_i(x,y) < 0$ for some $i \in \{1, 2, ..., k\}$;*

3. *For any $x \in \mathcal{X}$, $\mathcal{Y}(x)$ is compact, and $\|y\| \leq R$ for any y in $\mathcal{Y}(x)$.*

When $g(x,y)$ and $h_i(x,y)$ are all convex in y, the lower-level objective function $\widetilde{g}_t(x)$ of the barrier reformulation 1.4 may not have the unique optimal solution. Consider the following counterexample:

Example D.1. *Define the lower-level objective function $g(x, y)$ as follows*

$$g(x, y) = \begin{cases} (y-1)^2 & \text{if } 1 \le y \le 2 \\ 0 & \text{if } -1 \le y < 1 \\ (y+1)^2 & \text{if } -2 \le y < -1, \end{cases}$$

and the lower-level constraint $h(x, y) := g(x, y) - 1$. Then $g(x, y)$ and $h(x, y)$ are both convex, and

$$\widetilde{g}_t(x, y) = g(x, y) - t \log(1 - g(x, y)) \ge 0 - t \log(1) = 0,$$

The optimal solution set of the barrier reformulation is $[-1, 1]$ for any t, which means $y_t^*(x)$ is not unique.

Because the lower-level problem does not have a unique solution, the hypergradient cannot be determined. Therefore, in this situation, we cannot use the implicit gradient methods for the barrier reformulation.

When $g(x, y)$ is convex in y, but there exist at least one strongly convex constraint $h_i(x, y)$ in y, it is not hard to see the lower-level problem of barrier reformulation is strongly convex. Therefore, the hyperfunction of barrier reformulation is well-defined. From this, we can explore the connection between the hyperfunctions of the original problem and the barrier reformulation. Like the counterexample 3.1 under the linear setting, we can still only attempt to prove the convergence of the hyperfunction when $y^*(x)$ is unique. Before proceeding, let us first present an assumption for the convex case.

Assumption D.2. *$g(x, y)$ is convex in y for any x, and $h_i(x, y)$ are convex in y for any i and x. There exists at least one strongly convex constraint $h_i(x, y)$ in y.*

We give the following convergence result for the convex case:

Theorem D.1. *Suppose Assumption D.1 and Assumption D.2 hold. If x satisfies that $y^*(x)$ is unique, then $\lim_{t \to 0} \phi_t(x) = \phi(x)$.*

The proof can be directly obtained by optimal condition

$$g(x, y_t^*(x)) - g(x, y^*(x)) \le kt,$$

which means $\lim_{t \to 0} g(x, y_t^*(x)) = g(x, y^*(x))$. Note that $y^*(x)$ is unique, it not hard to see $\lim_{t \to 0} y_t^*(x) = y^*(x)$, which implies $\lim_{t \to 0} \phi_t(x) = \phi(x)$.

However, establishing a relationship between the hypergradients may still be challenging, as the hypergradient of the original problem might not be determinable through the expression:

$$\nabla_x \phi(x) = \nabla_x f(x, y^*(x)) - \nabla_{xy}^2 g(x, y^*(x))(\nabla_{yy}^2 g(x, y^*(x)))^{-1} \nabla_y f(x, y^*(x))$$

for any x. Let's see the following example:

Example D.2. *Define the lower-level objective function $g(x, y) = y^4$, and the lower-level constraint $h(x, y) := y^2 - 1$. Then $g(x, y)$ is convex, and $h(x, y)$ is strongly convex. The optimal solution $y^*(x) = 0$ for any x. However, the hypergradient*

$$\nabla_x f(x, y^*(x)) - \nabla_{xy}^2 g(x, y^*(x))(\nabla_{yy}^2 g(x, y^*(x)))^{-1} \nabla_y f(x, y^*(x))$$

is not well-defined since $\nabla_{yy}^2 g(x, y) = 0$ is not invertible for any x and y. In this context, how to calculate the hypergradient remains unexplored.

Remark D.1. *For the strong convexity of the lower-level problem in the barrier reformulation, we need at least one constraint $h_i(x,y)$ to be strongly convex in y. However, by Assumption D.1(3), we can add a constraint $h_{k+1}(x,y) = \|y\|^2 \leq R^2$ which is strongly convex in y. After adding this constraint, it remains equivalent to the original problem. Therefore, we can say Assumption D.1(3) implies Assumption D.2.*

The most important result, Theorem 4.1, still holds in the general convex case. Therefore, Algorithm 1 and Algorithm 2 still work in the general convex case, and we can get the same convergence rate $\widetilde{\mathcal{O}}(1/(\epsilon^2 t^{4.5}))$ in Theorem 4.3. However, we do not have the convergence for hypergradient. The Algorithm 3 does not work for this case.

D.1   Proof of Proposition 3.1

Through direct computing the Hessian of $\widetilde{g}_t(x, y_t^*(x))$, we get

$$\nabla_{yy}^2 \widetilde{g}_t(x, y_t^*(x)) = \nabla_{yy}^2 g(x, y_t^*(x)) + t \sum_{i=1}^{k} \left(\frac{\nabla_{yy}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))} + \frac{\nabla_y h_i(x, y_t^*(x)) \nabla_y h_i(x, y_t^*(x))^\top}{h_i^2(x, y_t^*(x))} \right).$$

If Assumption 3.2 holds, then $\nabla_{yy}^2 \widetilde{g}_t(x, y) \succeq \mu_g I$, which means $y_t^*(x)$ is unique for any $t > 0$ and $x \in \mathcal{X}$. Thus the hypergradient can be computed directly by using implicit function Theorem (see equation (2.8) in Ghadimi and Wang (2018))

$$\nabla_x \widetilde{\phi}_t(x) = \nabla_x f(x, y_t^*(x)) - \nabla_{xy}^2 \widetilde{g}_t(x, y_t^*(x))(\nabla_{yy}^2 \widetilde{g}_t(x, y_t^*(x)))^{-1} \nabla_y f(x, y_t^*(x)). \tag{D.1}$$

This also illustrates that $\widetilde{\phi}_t(x)$ is differentiable.

If Assumption 3.3 holds (i.e., h_i's and g are linear), the Hessian can be bounded below by

$$\nabla_{yy}^2 \widetilde{g}_t(x, y_t^*(x)) = t \sum_{i=1}^{k} \frac{\nabla_y h_i(x, y_t^*(x)) \nabla_y h_i(x, y_t^*(x))^\top}{h_i^2(x, y_t^*(x))}$$

$$\geq \frac{t}{\min_i h_i^2(x, y_t^*(x))} \sum_{i=1}^{k} \nabla_y h_i(x, y_t^*(x)) \nabla_y h_i(x, y_t^*(x))^\top.$$

It is noteworthy that $h_i(x, y_t^*(x)) < 0$ for any i. This is because, if $h_i(x, y_t^*(x)) = 0$ for some i, then $\widetilde{g}_t(x, y^*(x)) = \infty$ due to the log-barrier, which contradicts that $y_t^*(x)$ minimize $\widetilde{g}_t(x, y)$. Since $h_i(x, y)$ is linear, we can set

$$h_i(x, y) = a_i(x)^\top y - b_i(x).$$

Define $A(x) = (a_1(x), \cdots, a_k(x))^\top$. Note that the feasible set $\mathcal{Y}(x)$ is compact, we get $A(x)^\top A(x)$ is positive definitive for any x. Otherwise, if there exists a vector v such that $v^\top A(x)^\top A(x)v = 0$, then $v^\top a_i(x) = 0$ for any i. However, in this case, if $h_i(x, y) \leq 0$, it implies that $h_i(x, y + \alpha v) \leq 0$ for any α, which contradicts the compactness of feasible set. Thus, we know $\nabla_{yy}^2 \widetilde{g}_t(x, y_t^*(x)) \succ 0$, which implies (D.1) still holds. We obtain that $\widetilde{\phi}_t(x)$ is differentiable.

E   Convergence of Hypergradient to the original problem

E.1   Proof of Proposition 3.2

(1) Under the strongly convex setting, we already have that $y^*(x)$ is unique. By Giorgi and Zuccotti (2018), it is known that $\nabla_x y^*(x)$ exists. As for the converse, it does not hold. Consider the following

counterexample: $g(x,y) = y^2$ and $h(x,y) = y$, where $y \in \mathbb{R}$. Then $y^*(x) = 0$ and $\nabla_x y^*(x) = 0$ for any x, and also $h(x,y)$ is always active. However, by KKT condition we have

$$0 = \nabla_y g(x, 0) + \lambda(x) \nabla_y h(x, 0) = \lambda(x) \nabla_y h(x, 0) = \lambda(x).$$

Thus the multiplier $\lambda(x)$ always equals 0.

(2) Under the linear setting, we fist prove **(a)** is equivalent to **(b)**, and then prove **(b)** is equivalent to **(c)**.

(a)$\Rightarrow$(b): Assuming x is an **SCSC** point, we want to show that $y^*(x)$ is unique. We will use proof by contradiction to demonstrate that if this is not true, then there must exist a point (x, y) at which the optimal Lagrange multiplier $\lambda_i(x, y)$ is zero for some i.

Suppose $y^*(x)$ is not unique, then the optimal solution set must be the face F of a polyhedron. Denote ∂F as the relative boundary of F (for specific definition, see Definition 2.11 on Page 28 in Drusvyatskiy (2020)). For any two points y and y' in $F \setminus \partial F$, they have the same active constraints. Due to linearity of g and h, we have $\nabla_y g(x, y) = \nabla_y g(x, y')$ and $\nabla_y h_i(x, y) = \nabla_y h_i(x, y')$ for any i. Combining these facts with the LICQ assumption, we get $\lambda_i(x, y) = \lambda_i(x, y')$. For any point y'' in ∂F, the active index set of $y \in F \setminus \partial F$ is necessarily a proper subset of the active index set of y''. By LICQ assumption, we can prove $\lambda_i(x, y) = \lambda_i(x, y'')$ for any active i as follows: denote $\mathcal{I}^*(x, y)$, $\mathcal{I}^*(x, y'')$ as the set of active indices at (x, y) and (x, y''), respectively. Consider the following KKT condition at (x, y) and (x, y'')

$$\nabla_y g(x, y) + \sum_{i \in \mathcal{I}^*(x,y)} \lambda_i(x, y) \nabla_y h_i(x, y) = 0 \tag{E.1}$$

$$\nabla_y g(x, y'') + \sum_{i \in \mathcal{I}^*(x,y'')} \lambda_i(x, y'') \nabla_y h_i(x, y'') = 0 \tag{E.2}$$

Note that $\mathcal{I}^*(x, y)$ is a proper subset of $\mathcal{I}^*(x, y'')$, and $\nabla_y g(x, y) = \nabla_y g(x, y'')$, $\nabla_y h_i(x, y) = \nabla_y h_i(x, y'')$ since $g(x, y)$, $h_i(x, y)$ is linear in y for any i, by subtracting (E.1) from (E.2), we have:

$$\sum_{i \in \mathcal{I}^*(x,y'')} (\lambda_i(x, y'') - \lambda_i(x, y)) \nabla_y h_i(x, y'') = 0.$$

By LICQ assumption, the gradients $\nabla_y h_i(x, y'')$ are independent for $i \in \mathcal{I}^*(x, y'')$. Thus, we obtain that $\lambda_i(x, y) = \lambda_i(x, y'')$ for any $i \in \mathcal{I}^*(x, y'')$. Since $\mathcal{I}^*(x, y)$ is a proper subset of $\mathcal{I}^*(x, y'')$, there exists an index $i \in \mathcal{I}^*(x, y'')$ but $i \notin \mathcal{I}^*(x, y)$. Therefore, we know $\lambda_i(x, y) = 0$, which further means $\lambda_i(x, y'') = \lambda_i(x, y) = 0$. However, i is active at (x, y''). This contradicts the assumption that x is an **SCSC** point, thus proving the uniqueness of $y^*(x)$. By the argument above, we can also conclude $\lambda_i(x, y)$ is not a function of $y \in y^*(x)$.

(b)$\Rightarrow$(a): Assuming $y^*(x)$ is unique, we want to show x is an **SCSC** point. From the proof of **(a)$\Rightarrow$(b)**, we know that $\lambda_i(x, y)$ is independent of the choice of $y \in y^*(x)$. With out loss of generality, we denote $\lambda_i(x, y) = \lambda_i(x)$. We will use proof by contradiction to demonstrate that if this is not true, we can construct a vector v such that we can prove $h_i(x, y^*(x) + sv) \leq 0$ for any index i and s small enough. This means $(x, y^*(x) + sv)$ is feasible. We can also prove $g(x, y^*(x) + sv) = g(x, y^*(x))$. This contradicts the uniqueness of $y^*(x)$.

Given that $y^*(x)$ is the unique optimal solution, it must be an extreme point of a polyhedron. Without loss of generality, we assume $y^*(x)$ is determined by $h_i(x, y^*(x)) = 0$, $i = 1, ..., p$, i.e. the

40

first p constraints are active. Since $y^*(x)$ is unique, we have $p \geq m$. By assuming the LICQ, we also know $p \leq m$. Thus p exactly equals m. By complementary slackness, the multipliers satisfy $\lambda_i(x) = 0$ for $i = m+1, ..., k$. We now prove that $\lambda_i(x) > 0$ for each $i \in \{1, ..., m\}$. Suppose, for contradiction, that $\lambda_i(x) = 0$ for some $i \in \{1, ..., m\}$. Without loss of generality, assume $\lambda_i(x) = 0$ for $i \in \{1, ..., l\}$, where $l \leq m$. According to KKT conditions, we have

$$\nabla_y g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_y h_i(x, y^*(x))$$

$$= \begin{cases} \nabla_y g(x, y^*(x)) + \sum_{i=l+1}^{m} \lambda_i(x) \nabla_y h_i(x, y^*(x)) & \text{if } l < m \\ \nabla_y g(x, y^*(x)) & \text{if } l = m \end{cases} \tag{E.3}$$

$$= 0.$$

For notational convenience, we suppose that $l < m$. The case that $l = m$ is similar. Therefore,

$$\nabla_y g(x, y^*(x)) \in \text{span}\{\nabla_y h_i(x, y^*(x)) : i = l+1, ..., m\},$$

where $\text{span}\{\nabla_y h_i(x, y^*(x)) : i = l+1, ..., m\}$ is the subspace spanned by $\nabla_y h_i(x, y^*(x))$ for $i = l+1, ..., m$. Now consider the subspace $\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\}$. The LICQ assumption means these vectors are independent, thus the dimension of this subspace is l. Let us denote the matrix formed by these gradients as

$$A = [\nabla_y h_1(x, y^*(x)), \cdots, \nabla_y h_l(x, y^*(x))]^\top.$$

By Gordan's theorem, either $Av < 0$ has a solution v, or $A^\top w = 0$ has a nonzero solution w with $w \geq 0$. Note that by LICQ assumption, $\nabla_y h_1(x, y^*(x)), \cdots, \nabla_y h_l(x, y^*(x))$ are linear independent, so $A^\top w = 0$ has only the zero solution. Consequently, there exists a vector $v \neq 0$ such that $v^\top \nabla_y h_i(x, y^*(x)) < 0$ for any $i \in \{1, ..., m\}$. Treating $\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., m\}$ as the entire space, we consider the orthogonal complement of the subspace $\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\}$, denoted as $(\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\})^\perp$. Denote the sum of this space and the space spanned by v as $(\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\})^\perp + \text{span}\{v\}$. The dimension of $(\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\})^\perp + \text{span}\{v\}$ is $m - l + 1$. Treating $\mathbb{R}^m$ as the entire space, we consider the orthogonal complement of the subspace $\text{span}\{\nabla_y h_i(x, y^*(x)) : i = l+1, ..., m\}$, denoted as $(\text{span}\{\nabla_y h_i(x, y^*(x)) : i = l+1, ..., m\})^\perp$. Its dimension is l. Then, the intersection of two subspaces

$$\left\{ (\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\})^\perp + \text{span}\{v\} \right\} \cap (\text{span}\{\nabla_y h_i(x, y^*(x)) : i = l+1, ..., m\})^\perp \neq \emptyset$$

since the sum of their dimensions is greater that m. Choose a vector $\hat{v} \neq 0$ in the intersection of the two subspaces. We decompose $\hat{v} = v_1 + v_2$, where $v_1 \in (\text{span}\{\nabla_y h_i(x, y^*(x)) : i = 1, ..., l\})^\perp$ and $v_2 \in \text{span}\{v\}$, so there exist a constant α such that $v_2 = \alpha v$. We can assume $\alpha \geq 0$, otherwise consider $-\hat{v}$. Therefore, we have

$$\hat{v}^\top \nabla_y h_i(x, y^*(x)) = v_2^\top \nabla_y h_i(x, y^*(x)) = \alpha v^\top \nabla_y h_i(x, y^*(x)) \leq 0$$

for any $i \in \{1, ..., l\}$, which implies

$$h_i(x, y^*(x) + s\hat{v}) \leq 0 \text{ for any } i \in \{1, ..., l\}, \text{ and } s > 0 \tag{E.4}$$

by the linearity of $h_i(x, y)$. Note that $\hat{v} \in (\text{span}\{\nabla_y h_i(x, y^*(x)) : i = l + 1, ..., m\})^\perp$, we also have $\hat{v}^\top \nabla_y h_i(x, y^*(x)) = 0$ for any $i \in \{l + 1, ..., m\}$, which means

$$h_i(x, y^*(x) + s\hat{v}) = 0 \text{ for any } i \in \{l + 1, ..., m\} \text{ and } s \in \mathbb{R}. \tag{E.5}$$

Since $i \in \{m + 1, ..., k\}$ is inactive, we know $h_i(x, y^*(x)) < 0$ for any $i \in \{m + 1, ..., k\}$. By the continuity of $h_i(x, y)$ we can find a constant $S > 0$ such that

$$h_i(x, y^*(x) + s\hat{v}) < 0 \text{ for any } 0 < s \leq S. \tag{E.6}$$

Combining the fact $\hat{v}^\top \nabla_y h_i(x, y^*(x)) = 0$ and (E.3), we obtain $\hat{v}^\top \nabla_y g(x, y^*(x)) = 0$, and thus

$$g(x, y^*(x) + s\hat{v}) = g(x, y^*(x)) \text{ for any } s \in \mathbb{R}. \tag{E.7}$$

(E.4), (E.5), (E.6) means $y^*(x) + s\hat{v}$ remains feasible for any $0 < s \leq S$. (E.7) implies $y^*(x) + s\hat{v}$ is still optimal for any $0 < s \leq S$. These contradicts the uniqueness of $y^*(x)$. Therefore, we conclude that $\lambda_i(x) > 0$ for any $i \in \{1, ..., m\}$, i.e. x is an **SCSC** point.

(b)$\Rightarrow$(c): Assuming $y^*(x)$ is unique, we will prove $\nabla_x y^*(x)$ exists. Since we have proved that (b) is equivalent to (a), x is also an **SCSC** point. By Lemma B.3, i.e. the continuity of $\lambda_i(x)$, there exists a neighborhood U of x such that any point in U is an **SCSC** point, and their active indices are the same. Without loss of generality, suppose $y^*(x)$ is the solution of following equations

$$\begin{pmatrix} h_1(x, y^*(x)) \\ \vdots \\ h_p(x, y^*(x)) \end{pmatrix} = 0.$$

for any x in the neighborhood. Since $y^*(x)$ is unique, we have $p \geq m$. According to LICQ assumption, we also know $p \leq m$. Therefore, p exactly equals m, and $(\nabla_y h_1(x, y^*(x)), ..., h_m(x, y^*(x)))$ is invertivble. By the Implicit Function Theorem, $y^*(x)$ is differentiable at x.

(c)$\Rightarrow$(b): Assuming $\nabla_x y^*(x)$ exists, we will show $y^*(x)$ is unique. The existence of the gradient $\nabla_x y^*(x)$ implicitly assumes differentiability at the point $y^*(x)$, which in turn requires that $y^*(x)$ be well-defined and single-valued. Therefore, $y^*(x)$ is unique.

E.2 Proof of Theorem 3.3

We first establish the following lemma:

Lemma E.1. *Suppose Assumption 3.1 and 3.3 hold and $y^*(x)$ is unique, then we have $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$.*

Proof. The KKT condition of the original problem is

$$\nabla_y g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_y h_i(x, y^*(x)) = 0. \tag{E.8}$$

We first claim that under linear case, the notations in (E.8) is also meaningful because

1. By Remark 3.2, $\lambda_i(x, y)$ is independent of the choice of $y \in y^*(x)$. Without loss of generality we denote $\lambda_i(x) = \lambda_i(x, y)$;

2. Although $y^*(x)$ may represent a set, given that $h_i(x, y)$ and $g(x, y)$ are all linear in y, $\nabla_y h_i(x, y)$ is independent of the choice of y. Without loss of generality we use the notation $\nabla_y h_i(x, y^*(x))$ and $\nabla_y g(x, y^*(x))$.

By subtracting (E.8) from optimal condition equation of barrier reformulation

$$\nabla_y g(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) = 0,$$

we derive the following expression

$$\nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) + \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y^*(x)) - \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \right) = 0.$$

By adding and subtracting $\sum_{i=1}^{k} \lambda_i(x) \nabla_y h_i(x, y_t^*(x))$, we obatin

$$
\begin{aligned}
0 =& \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y^*(x)) - \lambda_i(x) \nabla_y h_i(x, y_t^*(x)) \right) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y_t^*(x)) - \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \right) \\
=& \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \\
&+ \sum_{i=1}^{k} \lambda_i(x) \left(\nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \right) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right) \nabla_y h_i(x, y_t^*(x)).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\left\| \sum_{i=1}^{k} \left(\lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right) \nabla_y h_i(x, y_t^*(x)) \right\| \\
&\leq \| \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \| + \left\| \sum_{i=1}^{k} \lambda_i(x) \left(\nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \right) \right\| \\
&\leq \| \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \| + \sum_{i=1}^{k} \lambda_i(x) \| \nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \| \quad \text{(E.9)} \\
&\overset{(i)}{=} 0,
\end{aligned}
$$

where (i) comes from the following equations due to the linearity of $g(x, y)$ and $h_i(x, y)$ for any i

$$
\begin{aligned}
\| \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \| &= 0, \\
\| \nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \| &= 0.
\end{aligned}
$$

Next we prove that $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$ for any i, i.e. for any $\epsilon > 0$, there exists T such that for any $0 < t \leq T$, we have

$$\left| \lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right| \leq \epsilon.$$

Without loss of generality, we prove the case $i = 1$. If this claim were false, there exists $\epsilon_0 > 0$ such that for any $T > 0$, we can find $0 < t \leq T$ satisfies

$$\left| \lambda_1(x) - \frac{t}{-h_1(x, y_t^*(x))} \right| \geq \epsilon_0.$$

This implies we can find a sequence t_j, such that

$$\left| \lambda_1(x) - \frac{t_j}{-h_1(x, y_{t_j}^*(x))} \right| \geq \epsilon_0.$$

for any j and $\lim_{j \to \infty} t_j = 0$. Set

$$\mu_j^i = \lambda_i(x) - \frac{t_j}{-h_i(x, y_{t_j}^*(x))}$$

and let $i_j^* = \arg\max_i |\mu_j^i|$, we can choose a subsequence of $\{t_j\}_{j=1}^\infty$, still denoted as $\{t_j\}_{j=1}^\infty$, such that $i_j^* = i^*$ for any j, where index i^* is a fixed index. It holds that $|\mu_j^{i^*}| \geq |\mu_j^1| \geq \epsilon_0$. Moreover, we have $|\mu_j^i|/|\mu_j^{i^*}| \leq 1$ for any index i. Therefore, we can select a subsequence, still denoted as $\{t_j\}_{j=1}^\infty$, such that $\lim_{j \to \infty} |\mu_j^i|/|\mu_j^{i^*}| =: \hat{\mu}^i$ exists for any i. By (E.9), we obtain

$$\left\| \sum_{i=1}^k \left(\lambda_i(x) - \frac{t_j}{-h_i(x, y_{t_j}^*(x))} \right) \nabla_y h_i(x, y_{t_j}^*(x)) \right\| = 0 \tag{E.10}$$

Let j go to infinity, we obtain

$$\sum_{i=1}^k \hat{\mu}^i \nabla_y h_i(x_0, y^*(x_0)) = 0.$$

Recall (3.2) says

$$\|y - y^*(x)\| \leq \frac{1}{\tau(x)\|\nabla_y g(x, y^*(x))\|} \left(g(x, y) - g(x, y^*(x)) \right).$$

According to Lemma B.1, which shows $g(x, y_t^*) - g(x, y^*(x)) \leq kt$, we have

$$\|y_{t_j}^*(x) - y^*(x)\| \leq \frac{1}{\tau(x)\|\nabla_y g(x, y^*(x))\|} \left(g(x, y_{t_j}^*(x)) - g(x, y^*(x)) \right) \leq \frac{kt_j}{\tau(x)\|\nabla_y g(x, y^*(x))\|}, \tag{E.11}$$

where $\tau(x)$ is the lower bound of the cosine of the angle between $y - y^*(x)$ and the $\nabla_y g(x, y^*(x))$. The right hands of (E.11) is finite because when $y^*(x)$ is unique, $\tau(x)\|\nabla_y g(x, y^*(x))\| \neq 0$. Therefore, $\lim_{j \to \infty} y_{t_j}^*(x) = y^*(x)$. This implies

$$|\hat{\mu}^i| = \lim_{j \to \infty} \frac{|\mu_j^i|}{|\mu_j^{i^*}|} \leq \frac{|\lambda_i(x) - \frac{0}{-h_i(x, y^*(x))}|}{\epsilon_0} = 0$$

44

for any $i \notin \mathcal{I}^*(x_0)$. Dividing both sides of (E.10) with $\mu_j^{i^*}$ and let j go to infinity, we have

$$0 = \lim_{j \to \infty} \frac{\left\| \sum_{i=1}^k \left(\lambda_i(x) - \frac{t_j}{-h_i(x, y_{t_j}^*(x))} \right) \nabla_y h_i(x, y_{t_j}^*(x)) \right\|}{\mu_j^{i^*}}$$

$$= \sum_{i=1}^k \hat{\mu}^i \nabla_y h_i(x, y^*(x))$$

$$\overset{(i)}{=} \sum_{i \in \mathcal{I}^*(x_0)} \hat{\mu}^i \nabla_y h_i(x, y^*(x))$$

In (i) we utilize the face that $\hat{\mu}^i = 0$ for any $i \notin \mathcal{I}^*(x_0)$. Since $\hat{\mu}^{i^*} = \lim_{j \to \infty} |\mu_j^{i^*}|/|\mu_j^{i^*}| = 1$, we also have $i^* \in \mathcal{I}^*(x)$. However, by the LICQ assumption, $\nabla_y h_i(x, y^*(x))$ are linear independent for $i \in \mathcal{I}^*(x_0)$, which means $\hat{\mu}^i$ for any $i \in \mathcal{I}^*(x_0)$. Here is a contradiction, so $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$ for any i. $\qquad \square$

Lemma E.2. *Suppose Assumption 3.1 and Assumption 3.3 hold. If x is an* **SCSC** *point, then there are exact m constraints active and the following matrix is invertible:*

$$\sum_{i=1}^k \lambda_i^2(x) \nabla_y h_i(x, y^*(x)) \left(\nabla_y h_i(x, y^*(x)) \right)^\top.$$

Proof. By Proposition 3.2, x is an **SCSC** point means $y^*(x)$ is unique. Since the lower-level problem is a linear program, $y^*(x)$ is a vertex of a polyhedron, so there are at least m constraints active. According to LICQ assumption, there are at most m constraints active. Without loss of generality, assume $h_i(x, y^*(x))$ is active for $i = 1, ..., m$. We also know $\nabla_y h_i(x, y^*(x))$ are linear independent for $i = 1, ..., m$ by LICQ assumption. Therefore, the following matrix

$$\sum_{i=1}^k \lambda_i^2(x) \nabla_y h_i(x, y^*(x)) \left(\nabla_y h_i(x, y^*(x)) \right)^\top = \sum_{i=1}^m \lambda_i^2(x) \nabla_y h_i(x, y^*(x)) \left(\nabla_y h_i(x, y^*(x)) \right)^\top.$$

is invertible. $\qquad \square$

Next, we prove the relation of Jacobians in linear case when t approaches 0.

Proof of Theorem 3.3: We fix an **SCSC** point x_0 and prove this theorem at this point. We intend to present our proof in two steps. First, we will compute the limit of $\nabla_x y_t^*(x_0)$ when t goes to 0. Second, we will compute $\nabla_x y^*(x_0)$ directly, and then show that they are equal.

Step 1. In this step, we will use the standard process to compute the Jacobian $\nabla_x y_t^*(x_0)$ as follows

$$\nabla_x y_t^*(x_0) = (\nabla_{yy} \widetilde{g}_t(x_0, y_t^*(x_0)))^{-1} \nabla_{yx} \widetilde{g}_t(x_0, y_t^*(x_0)) \tag{E.12}$$
$$= (t \nabla_{yy} \widetilde{g}_t(x_0, y_t^*(x_0)))^{-1} t \nabla_{yx} \widetilde{g}_t(x_0, y_t^*(x_0)).$$

where $0 < t \le T$ with some constant T. Next we will leverage Lemma E.1 and Lemma E.2 to compute the limit

$$\lim_{t \to 0} (t \nabla_{yy} \widetilde{g}_t(x_0, y_t^*(x_0)))^{-1} t \nabla_{yx} \widetilde{g}_t(x_0, y_t^*(x_0)).$$

Since $\widetilde{g}_t(x, y)$ is strongly convex in y for any x, by the optimal condition of the barrier reformulated lower-level problem, we have

$$\nabla_y \widetilde{g}_t(x, y_t^*(x)) = 0 \quad \forall x.$$

Note that $\widetilde{g}_t$ is two times continuously differentiable by Assumption 3.1(1), to get the expression of $\nabla_x y_t^*(x_0)$, we take the gradient with respect to x at point x_0 on both sides, yielding

$$0 = \nabla_{xy}\widetilde{g}_t(x_0, y_t^*(x_0))$$

$$\overset{(i)}{=} \nabla_x \left(\nabla_y g(x_0, y_t^*(x_0)) + \sum_{i=1}^k \frac{t\nabla_y h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right)$$

$$\overset{(ii)}{=} \nabla_{xy}^2 g(x_0, y_t^*(x_0)) + (\nabla_x y_t^*(x_0))^\top \nabla_{yy}^2 g(x_0, y_t^*(x_0))$$

$$+ \sum_{i=1}^k \frac{t\nabla_x h_i(x_0, y_t^*(x_0)) (\nabla_y h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))} + (\nabla_x y_t^*(x_0))^\top \sum_{i=1}^k \frac{t\nabla_y h_i(x_0, y_t^*(x_0)) (\nabla_y h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))}$$

$$+ \sum_{i=1}^k \frac{t\nabla_{xy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} + (\nabla_x y_t^*(x_0))^\top \sum_{i=1}^k \frac{t\nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))},$$

where (i) is by the definition of $\widetilde{g}_t(x, y)$ and (ii) is by direct computation. Rearranging terms, we obtain

$$\left(\nabla_{yy}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^k \frac{t\nabla_y h_i(x_0, y_t^*(x_0)) (\nabla_y h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))} + \sum_{i=1}^k \frac{t\nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right) \nabla_x y_t^*(x_0)$$

$$= - \left(\nabla_{yx}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^k \frac{t\nabla_y h_i(x_0, y_t^*(x_0)) (\nabla_x h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))} + \sum_{i=1}^k \frac{t\nabla_{yx}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right). \tag{E.13}$$

Lemma E.1 and Lemma B.1 tell us $\lim_{t\to 0} t/(-h_i(x_0, y_t^*(x_0))) = \lambda_i(x_0)$ for any i and $\lim_{t\to 0} y_t^*(x_0) = y^*(x_0)$. Thus

$$\lim_{t\to 0} \left(t\nabla_{yy}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^k \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) (\nabla_y h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))} + t\sum_{i=1}^k \frac{t\nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right)$$

$$\overset{(i)}{=} \sum_{i=1}^k \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) (\nabla_y h_i(x_0, y^*(x_0)))^\top, \tag{E.14}$$

and

$$\lim_{t\to 0} \left(t\nabla_{yx}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^k \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) (\nabla_x h_i(x_0, y_t^*(x_0)))^\top}{h_i^2(x_0, y_t^*(x_0))} + t\sum_{i=1}^k \frac{t\nabla_{yx}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right)$$

$$\overset{(ii)}{=} \sum_{i=1}^k \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) (\nabla_x h_i(x_0, y^*(x_0)))^\top, \tag{E.15}$$

where (i), (ii) are because

$$\lim_{t\to 0} t\nabla_{yy}^2 g(x_0, y_t^*(x_0)) = \lim_{t\to 0} t \cdot \nabla_{yy}^2 g(x_0, y^*(x_0)) = 0,$$

$$\lim_{t\to 0} t\nabla_{yx}^2 g(x_0, y_t^*(x_0)) = \lim_{t\to 0} t \cdot \nabla_{yx}^2 g(x_0, y^*(x_0)) = 0,$$

$$\lim_{t\to 0} t\frac{t\nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} = \lim_{t\to 0} t \cdot \lambda_i(x_0) \nabla_{yy}^2 h_i(x_0, y^*(x_0)) = 0,$$

46

$$\lim_{t \to 0} t \frac{t \nabla_{yx}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} = \lim_{t \to 0} t \cdot \lambda_i(x_0) \nabla_{yx}^2 h_i(x_0, y^*(x_0)) = 0.$$

By Lemma E.2, the following matrix is invertible

$$\sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\nabla_y h_i(x_0, y^*(x_0))\right)^\top.$$

Therefore, again applying

$$\lim_{t \to 0} \sum_{i=1}^{k} \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top}{h_i^2(x_0, y_t^*(x_0))} = \sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\nabla_y h_i(x_0, y^*(x_0))\right)^\top,$$

we can find a constant T small enough such that

$$\sum_{i=1}^{k} \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top}{h_i^2(x_0, y_t^*(x_0))}$$

is invertible for any $0 < t \le T$. This implies the invertibility of

$$t\nabla_{yy}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^{k} \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top}{h_i^2(x_0, y_t^*(x_0))} + t \sum_{i=1}^{k} \frac{t \nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))}.$$

for any $0 < t \le T$. By (E.13), we derive the expression of $\nabla_x y_t^*(x_0)$ for any $0 < t \le T$

$$\nabla_x y_t^*(x_0) = - \left[t\nabla_{yy}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^{k} \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top}{h_i^2(x_0, y_t^*(x_0))} + t \sum_{i=1}^{k} \frac{t \nabla_{yy}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right]^{-1}$$
$$\times \left[t\nabla_{yx}^2 g(x_0, y_t^*(x_0)) + \sum_{i=1}^{k} \frac{t^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_x h_i(x_0, y_t^*(x_0))\right)^\top}{h_i^2(x_0, y_t^*(x_0))} + t \sum_{i=1}^{k} \frac{t \nabla_{yx}^2 h_i(x_0, y_t^*(x_0))}{-h_i(x_0, y_t^*(x_0))} \right]$$

From (E.14) and (E.15), we have

$$\lim_{t \to 0} \nabla_x y_t^*(x_0) = \lim_{t \to 0} - \left(\sum_{i=1}^{k} \left(\frac{t}{-h_i(x_0, y_t^*(x_0))} \right)^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top \right)^{-1}$$
$$\times \left(\sum_{i=1}^{k} \left(\frac{t}{-h_i(x_0, y_t^*(x_0))} \right)^2 \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_x h_i(x_0, y_t^*(x_0))\right)^\top \right)$$
$$= - \left(\sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_y h_i(x_0, y_t^*(x_0))\right)^\top \right)^{-1}$$
$$\times \left(\sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y_t^*(x_0)) \left(\nabla_x h_i(x_0, y_t^*(x_0))\right)^\top \right)$$

Step 2. In this step, we will establish that

$$\nabla_x y^*(x_0) = - \left(\sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\nabla_y h_i(x_0, y^*(x_0))\right)^\top \right)^{-1}$$
$$\times \left(\sum_{i=1}^{k} \lambda_i^2(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\nabla_x h_i(x_0, y^*(x_0))\right)^\top \right)$$

47

$$= - \left(\sum_{i=1}^{k} \lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \right)^{\top} \right)^{-1}$$
$$\times \left(\sum_{i=1}^{k} \lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\lambda_i(x_0) \nabla_x h_i(x_0, y^*(x_0)) \right)^{\top} \right)$$

By the continuity of $h_i(x, y)$ and Lemma B.3, i.e. the continuity of $\lambda_i(x)$, we can select a neighborhood U of x_0 such that the active index set for any point in U is the same as that of x_0. According to Proposition 3.2, x is an **SCSC** point, which means $y^*(x)$ is unique. By Lemma E.2, there are exact m constrains active at point x_0, we assume $h_i(x_0, y^*(x_0)) = 0$ for $i = 1, ..., m$, and $i = m+1, ..., k$ are inactive. Therefore, $y^*(x)$ is the solution of the following m equations

$$\begin{pmatrix} h_1(x, y) \\ \vdots \\ h_m(x, y) \end{pmatrix} = 0.$$

for any x in U. Since $\lambda_i(x_0) > 0$ for any active index at point x_0 with $i = 1, ..., m$, these equations are equivalent to

$$\begin{pmatrix} \lambda_1(x_0) h_1(x, y) \\ \vdots \\ \lambda_m(x_0) h_m(x, y) \end{pmatrix} = 0.$$

Replace y by $y^*(x)$ in these equations, we have

$$\begin{pmatrix} \lambda_1(x_0) h_1(x, y^*(x)) \\ \vdots \\ \lambda_m(x_0) h_m(x, y^*(x)) \end{pmatrix} = 0.$$

These hold for any x in U. Take gradient for x on both sides, we obtain

$$\begin{pmatrix} \lambda_1(x_0) \nabla_x^{\top} h_1(x, y^*(x)) \\ \vdots \\ \lambda_m(x_0) \nabla_x^{\top} h_m(x, y^*(x)) \end{pmatrix} + \begin{pmatrix} \lambda_1(x_0) \nabla_y^{\top} h_1(x, y^*(x)) \\ \vdots \\ \lambda_m(x_0) \nabla_y^{\top} h_m(x, y^*(x)) \end{pmatrix} \nabla_x y^*(x) = 0$$

for any x in U. Multiply the matrix $(\lambda_1(x_0) \nabla_y h_1(x, y^*(x)), \cdots, \lambda_m(x_0) \nabla_y h_m(x, y^*(x)))$ to both sides and moving terms. Note that the index $i = m+1, ..., k$ are inactive at point x_0, then $\lambda_i(x_0) = 0$ for $i = m+1, ..., k$. We finalize

$$\nabla_x y^*(x_0) = - \left(\sum_{i=1}^{k} \lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \right)^{\top} \right)^{-1}$$
$$\times \left(\sum_{i=1}^{k} \lambda_i(x_0) \nabla_y h_i(x_0, y^*(x_0)) \left(\lambda_i(x_0) \nabla_x h_i(x_0, y^*(x_0)) \right)^{\top} \right) \qquad \text{(E.16)}$$
$$= \lim_{t \to 0} \nabla_x y_t^*(x_0).$$

This completes the proof.

E.3 Proof of Theorem 3.4

Lemma E.3. *Suppose Assumption 3.1 and 3.2 hold, then we have $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$ uniformly for any i.*

Proof. The KKT condition of original problem is

$$\nabla_y g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_y h_i(x, y^*(x)) = 0. \tag{E.17}$$

By subtracting (E.17) from optimal condition equation of barrier reformulation, given below,

$$\nabla_y g(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) = 0,$$

we derive the following expression

$$\nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) + \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y^*(x)) - \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \right) = 0.$$

It follows that

$$
\begin{aligned}
0 =& \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y^*(x)) - \lambda_i(x) \nabla_y h_i(x, y_t^*(x)) \right) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) \nabla_y h_i(x, y_t^*(x)) - \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \right) \\
=& \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \\
&+ \sum_{i=1}^{k} \lambda_i(x) \left(\nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \right) \\
&+ \sum_{i=1}^{k} \left(\lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right) \nabla_y h_i(x, y_t^*(x)) \\
=& \left\| \sum_{i=1}^{k} \left(\lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right) \nabla_y h_i(x, y_t^*(x)) \right\| \\
\leq& \| \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \| + \left\| \sum_{i=1}^{k} \lambda_i(x) \left(\nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \right) \right\| \\
\leq& \| \nabla_y g(x, y^*(x)) - \nabla_y g(x, y_t^*(x)) \| + \sum_{i=1}^{k} |\lambda_i(x)| \, \| \nabla_y h_i(x, y^*(x)) - \nabla_y h_i(x, y_t^*(x)) \| \tag{E.18}
\end{aligned}
$$

Next we prove that $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$ uniformly for any i, i.e. for any $\epsilon > 0$, there exists T such that for any $x \in \mathcal{X}$ and $0 < t \leq T$, we have

$$\left| \lambda_i(x) - \frac{t}{-h_i(x, y_t^*(x))} \right| \leq \epsilon.$$

Without loss of generality, we consider the case $i = 1$. We prove this by contradiction. If there exists $\epsilon_0 > 0$ such that for any $T > 0$, we can find $x \in \mathcal{X}$ and $0 < t \le T$ satisfies

$$\left| \lambda_1(x) - \frac{t}{-h_1(x, y_t^*(x))} \right| \ge \epsilon_0.$$

This implies we can find a sequence (x_j, t_j), such that the following holds for any j

$$\left| \lambda_1(x_j) - \frac{t_j}{-h_1(x_j, y_{t_j}^*(x_j))} \right| \ge \epsilon_0,$$

and also we have $\lim_{j \to \infty} t_j = 0$. Since $\mathcal{X}$ is compact, we can assume $\lim_{j \to \infty} x_j = x_0$. Set

$$\mu_j^i = \lambda_i(x_j) - \frac{t_j}{-h_i(x_j, y_{t_j}^*(x_j))}$$

and let $i_j^* = \arg\max_i |\mu_j^i|$, we can choose a subsequence of $\{(x_j, t_j)\}_{j=1}^{\infty}$, still denoted as $\{(x_j, t_j)\}_{j=1}^{\infty}$, such that $i_j^* = i^*$ for any j, where index i^* is a fixed index. It holds that $|\mu_j^{i^*}| \ge |\mu_j^1| \ge \epsilon_0$. Moreover, we have $|\mu_j^i| / |\mu_j^{i^*}| \le 1$ for any index i. Therefore, we can select a subsequence, still denoted as $\{(x_j, t_j)\}_{j=1}^{\infty}$, such that $\lim_{j \to \infty} |\mu_j^i| / |\mu_j^{i^*}| =: \hat{\mu}^i$ exists for any i. Apply (E.18) to (x_j, t_j), we obtain

$$\left\| \sum_{i=1}^{k} \left(\lambda_i(x_j) - \frac{t_j}{-h_i(x_j, y_{t_j}^*(x_j))} \right) \nabla_y h_i(x_j, y_{t_j}^*(x_j)) \right\|$$
$$\le \left\| \nabla_y g(x_j, y^*(x_j)) - \nabla_y g(x_j, y_{t_j}^*(x_j)) \right\| + \sum_{i=1}^{k} |\lambda_i(x_j)| \left\| \nabla_y h_i(x_j, y^*(x_j)) - \nabla_y h_i(x_j, y_{t_j}^*(x_j)) \right\|$$

$$\tag{E.19}$$

It is worth noting that

$$\lim_{j \to \infty} \left(\nabla_y h_i(x_j, y_{t_j}^*(x_j)) - \nabla_y h_i(x_0, y^*(x_0)) \right) = 0 \tag{E.20}$$

$$\lim_{j \to \infty} \left(\nabla_y g(x_j, y_{t_j}^*(x_j)) - \nabla_y g(x_0, y^*(x_0)) \right) = 0. \tag{E.21}$$

This is because, by Lemma B.1, $y_t^*(x)$ uniformly converges to $y^*(x)$. Under strongly convex setting, we also know $y^*(x)$ continuously depends on x. According to Lemma E.6, $y_{t_j}^*(x_j)$ converges to $y^*(x_0)$, which implies $(x_j, y_{t_j}^*(x_j))$ converges to $(x_0, y^*(x_0))$.

Note that $\lambda_i(x_j)$ is uniform bounded by Lemma B.3, the right-hand side of (E.19) converges to 0. Dividing both sides of (E.19) with μ_j^* and let j go to infinity, it follows that

$$\sum_{i=1}^{k} \hat{\mu}^i \nabla_y h_i(x_0, y^*(x_0)) = 0.$$

Applying Lemma B.1 to (x_j, t_j), we have

$$\| y_{t_j}^*(x_j) - y^*(x_j) \| \le \sqrt{\frac{2}{\mu_g} k t_j}.$$

50

This means $\lim_{j\to\infty} y^*_{t_j}(x_j) = y^*(x_0)$. Therefore, the following holds

$$|\hat{\mu}^i| = \lim_{j\to\infty}\frac{|\mu^i_j|}{|\mu^{i^*}_j|} \leq \frac{|\lambda_i(x_0) - \frac{0}{-h_i(x_0,y^*(x_0))}|}{\epsilon_0} = 0.$$

Thus, we have

$$\sum_{i\in\mathcal{I}^*(x_0)} \hat{\mu}^i \nabla_y h_i(x_0, y^*(x_0)) = 0.$$

Since $\hat{\mu}^{i^*} = \lim_{j\to\infty} |\mu^{i^*}_j|/|\mu^{i^*}_j| = 1$, we also have $i \in \mathcal{I}^*(x_0)$. However, by the LICQ assumption, $\nabla_y h_i(x_0, y^*(x_0))$ are linear independent for $i \in \mathcal{I}^*(x_0)$, which means $\hat{\mu}^i = 0$ for any $i \in \mathcal{I}^*(x_0)$. Here is a contradiction, so $t/(-h_i(x, y^*_t(x)))$ converges to $\lambda_i(x)$ uniformly for any i.

$\square$

Lemma E.4. *Suppose Assumption 3.1 and Assumption 3.2 hold, and x is an **SCSC** point. Then $\lim_{t\to 0} \nabla_x y^*_t(x)$ exists.*

Proof. Denote $\mathcal{I}^*(x)$ be the set of active index at $y^*(x)$. Since LICQ assumption is satisfied at $y^*(x)$, $\{\nabla_y h_i(x, y^*(x))\}_{i\in\mathcal{I}^*(x)}$ are linearly independent. By Lemma B.1 and Lemma B.2, $y^*_t(x)$ continuously converges to $y^*(x)$, so we can assume that t is sufficiently small such that $\{\nabla_y h_i(x, y^*_t(x))\}_{i\in\mathcal{I}^*(x)}$ are linearly independent. Following (E.12) in the proof of Theorem 3.3, we derive

$$\nabla_x y^*_t(x) = -\left(\nabla^2_{yy}g(x, y^*_t(x)) + \sum_{i=1}^k \frac{1}{t}\frac{t^2}{h^2_i(x, y^*_t(x))}\nabla_y h_i(x, y^*_t(x))\left(\nabla_y h_i(x, y^*_t(x))\right)^\top + \sum_{i=1}^k \frac{t\nabla^2_{yy}h_i(x, y^*_t(x))}{-h_i(x, y^*_t(x))}\right)^{-1}$$

$$\times \left(\nabla^2_{yx}g(x, y^*_t(x)) + \sum_{i=1}^k \frac{1}{t}\frac{t^2}{h^2_i(x, y^*_t(x))}\nabla_y h_i(x, y^*_t(x))\left(\nabla_x h_i(x, y^*_t(x))\right)^\top + \sum_{i=1}^k \frac{t\nabla^2_{yx}h_i(x, y^*_t(x))}{-h_i(x, y^*_t(x))}\right)$$

Unlike the linear case, handling the strongly convex case becomes extremely challenging. The reason is that in the linear case, the limit of a few terms in the above expression (e.g., $A^k_t(x)$ and $D_t(x)$ to be defined later) are invertible, but this is no longer the case without linearity. Therefore, it is necessary to develop a new set of tools.

Denote the following terms as

$$v^j_t(x) = \frac{t}{-h_j(x, y^*_t(x))}\nabla_y h_j(x, y^*_t(x)) \tag{E.22}$$

$$v^j(x) = \lambda_j(x)\nabla_y h_j(x, y^*(x)) \tag{E.23}$$

$$V^j_t(x) = \text{span}\{\frac{t}{-h_i(x, y^*_t(x))}\nabla_y h_i(x, y^*_t(x)) : i \in \{1, ..., j\} \cap \mathcal{I}^*(x)\} \tag{E.24}$$

$$V^j = \text{span}\{\lambda_i \nabla_y h_i(x, y^*(x)) : i \in \{1, ..., j\} \cap \mathcal{I}^*(x)\}$$

$$A^j_t(x) = \sum_{i=1}^j \frac{1}{t}(v^j_t(x))(v^j_t(x))^\top$$

$$= \sum_{i=1}^j \frac{1}{t}\frac{t^2}{h^2_i(x, y^*_t(x))}\nabla_y h_i(x, y^*_t(x))\left(\nabla_y h_i(x, y^*_t(x))\right)^\top, \quad j = 1, ..., k$$

$$B_t(x) = \nabla^2_{yy}g(x, y^*_t(x)) + \sum_{i=1}^k \frac{t\nabla^2_{yy}h_i(x, y^*_t(x))}{-h_i(x, y^*_t(x))}$$

51

$$B_t^j(x) = A_t^j(x) + B_t(x) \tag{E.25}$$

$$B(x) = \nabla_{yy}^2 g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_{yy}^2 h_i(x, y^*(x))$$

$$C_t(x) = \nabla_{yx}^2 g(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{t \nabla_{yx}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))}$$

$$C(x) = \nabla_{yx}^2 g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_{yx}^2 h_i(x, y^*(x))$$

$$D_t(x) = \sum_{i=1}^{k} \frac{1}{t} \frac{t^2}{h_i^2(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \left(\nabla_x h_i(x, y_t^*(x)) \right)^\top, \tag{E.26}$$

where $\mathcal{I}^*(x)$ denotes the set of **active index** at the point $(x, y^*(x))$.

By the above notations, we have

$$\nabla_x y_t^*(x) = \left(A_t^k(x) + B_t(x) \right)^{-1} \left(C_t(x) + D_t(x) \right). \tag{E.27}$$

We will demonstrate the convergence of the above two terms separately.
Step 1. To prove that the following limit exists,

$$\lim_{t \to 0} \left(A_t^k(x) + B_t(x) \right)^{-1} C_t(x)$$

we will show $\lim_{t \to 0} C_t(x)$ exists and $\lim_{t \to 0} \left(A_t^k(x) + B_t(x) \right)^{-1}$ exists as a linear transformation.

By Lemma E.3, i.e. $\lim_{t \to 0} t/(-h_i(x, y_t^*(x))) = \lambda_i(x)$, and Lemma B.1, which shows $\lim_{t \to 0} y_t^*(x) = y^*(x)$, we have

$$\lim_{t \to 0} C_t(x) = \lim_{t \to 0} \nabla_{yx}^2 g(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{t \nabla_{yx}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))}$$

$$= \nabla_{yx}^2 g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla_{yx}^2 h_i(x, y^*(x)).$$

Now we prove $\lim_{t \to 0} (A_t^k(x) + B_t(x))^{-1}$ exists, which is the key point of this proof. We proceed by induction.

If $j = 1$ is an active index, note that

$$A_t^j(x) = \sum_{i=1}^{j} \frac{1}{t} v_t^i(x) (v_t^i(x))^\top.$$

Applying the Sherman-Morrison formula, we obtain

$$(A_t^1(x) + B_t(x))^{-1} = \left(\frac{1}{t} v_t^1(x) \left(v_t^1(x) \right)^\top + B_t(x) \right)^{-1}$$

$$= B_t(x)^{-1} - \frac{B_t(x)^{-1} v_t^1(x) v_t^1(x)^\top B_t(x)^{-1}}{t^2 + v_t^1(x)^\top B_t(x)^{-1} v_t^1(x)}.$$

To prove $\lim_{t\to 0}\left(A_t^k(x)+B_t(x)\right)^{-1}$ exists, it is sufficient to show that $\lim_{t\to 0}B_t(x)$ and $\lim_{t\to 0}v_t^1(x)$ exist, and $\lim_{t\to 0}v_t^1(x)^\top B_t(x)^{-1}v_t^1(x)>0$. Applying Lemma E.3, and Lemma B.1 again, we obtain

$$\lim_{t\to 0}v_t^1(x)=\lim_{t\to 0}\frac{t}{-h_1(x,y_t^*(x))}\cdot\lim_{t\to 0}\nabla_y h_1(x,y_t^*(x))=\lambda_1(x)\nabla_y h_1(x,y^*(x))=v^1(x).$$

Since $B(x)\succeq\nabla_{yy}^2 g(x,y^*(x))\succ 0$, it follows that

$$\lim_{t\to 0}B_t(x)^{-1}=\left(\nabla_{yy}^2 g(x,y^*(x))+\sum_{i=1}^k\lambda_i(x)\nabla_{yy}^2 h_i(x,y^*(x))\right)^{-1}=B(x)^{-1}\succ 0,$$

Under the LICQ assumption and given that we have assumed index 1 is active, we have $v^1(x)\neq 0$. Therefore, $\lim_{t\to 0}v_t^1(x)^\top B_t(x)^{-1}v_t^1(x)=v^1(x)^\top B(x)^{-1}v^1(x)>0$, implying

$$\lim_{t\to 0}(A_t^1(x)+B_t(x))^{-1}=B(x)^{-1}-\frac{B(x)^{-1}v^1(x)v^1(x)^\top B(x)^{-1}}{v^1(x)^\top B(x)^{-1}v^1(x)}.$$

If $j=1$ is an inactive index, we obtain

$$\lim_{t\to 0}A_t^1(x)$$
$$=\lim_{t\to 0}\frac{1}{t}\frac{t^2}{h_1^2(x,y_t^*(x))}\nabla_y h_1(x,y_t^*(x))\left(\nabla_y h_1(x,y_t^*(x))\right)^\top$$
$$=\lim_{t\to 0}\frac{t}{h_1^2(x,y_t^*(x))}\nabla_y h_1(x,y_t^*(x))\left(\nabla_y h_1(x,y_t^*(x))\right)^\top$$
$$=0.$$

Therefore,

$$\lim_{t\to 0}(A_t^1(x)+B_t(x))^{-1}=B(x)^{-1}.$$

Next, we prove the existence of $\lim_{t\to 0}(A_t^k(x)+B_t(x))^{-1}$ by induction. For $j=1$, we have proved the existence of $\lim_{t\to 0}(A_t^j(x)+B_t(x))^{-1}$. Suppose when $j=\tilde{j}$, $\lim_{t\to 0}(A_t^{\tilde{j}}(x)+B_t(x))^{-1}=:G^{\tilde{j}}(x)$ exists, where $G^{\tilde{j}}(x)$ is some quantity to be specified soon. We will prove that for $j=\tilde{j}+1$, $\lim_{t\to 0}(A_t^{\tilde{j}+1}(x)+B_t(x))^{-1}=\lim_{t\to 0}(B_t^{\tilde{j}}(x))^{-1}=:G^{\tilde{j}+1}(x)$ also exists, by considering two cases, depending on whether $j+1$ is active or not.

Case 1: $j=\tilde{j}+1$ is an active index. Again using the Sherman-Morrison formula, we find

$$(A_t^{\tilde{j}+1}(x)+B_t(x))^{-1}=(\frac{1}{t}v_t^{\tilde{j}+1}(x)v_t^{\tilde{j}+1}(x)^\top+A_t^{\tilde{j}}(x)+B_t(x))^{-1}$$
$$\overset{(i)}{=}(\frac{1}{t}v_t^{\tilde{j}+1}(x)v_t^{\tilde{j}+1}(x)^\top+B_t^{\tilde{j}}(x))^{-1}$$
$$=B_t^{\tilde{j}}(x)^{-1}-\frac{B_t^{\tilde{j}}(x)^{-1}v_t^{\tilde{j}+1}(x)v_t^{\tilde{j}+1}(x)^\top B_t^{\tilde{j}}(x)^{-1}}{t^2+v_t^{\tilde{j}+1}(x)^\top B_t^{\tilde{j}}(x)^{-1}v_t^{\tilde{j}+1}(x)}. \tag{E.28}$$

(i) is because the notation in (E.25). Note that $\lim_{t\to 0}v_t^{\tilde{j}+1}(x)=\lambda_{\tilde{j}+1}(x)\nabla_y h_{\tilde{j}+1}(x,y^*(x))$ exists, and by the induction hypothesis, $\lim_{t\to 0}B_t^{\tilde{j}}(x)^{-1}=G^{\tilde{j}}(x)$ exists. Then to show that the limit of (E.28) exists, it is sufficient to prove that

$$\lim_{t\to 0}v_t^{\tilde{j}+1}(x)^\top B_t^{\tilde{j}}(x)^{-1}v_t^{\tilde{j}+1}(x)=v^{\tilde{j}+1}(x)^\top G^{\tilde{j}}(x)v^{\tilde{j}+1}(x)>0. \tag{E.29}$$

53

Towards this end, let us first define $w_t^{\widetilde{j}+1}(x) := B_t^{\widetilde{j}}(x)^{-1}v_t^{\widetilde{j}+1}(x)$, then it suffices to first show that $w^{\widetilde{j}+1}(x) = \lim_{t\to 0} w_t^{\widetilde{j}+1}(x)$ exists, then $v^{\widetilde{j}+1}(x)^\top w^{\widetilde{j}+1}(x) > 0$ as desired.

To proceed, let us first observe that the following two limits exist:

$$\lim_{t\to 0}(A_t^{\widetilde{j}}(x) + B_t(x))^{-1} = \lim_{t\to 0} B_t^{\widetilde{j}}(x)^{-1} := G^{\widetilde{j}}(x) \tag{E.30}$$

$$\lim_{t\to 0} v_t^{\widetilde{j}+1}(x) = \lambda_{\widetilde{j}+1}(x)\nabla_y h_{\widetilde{j}+1}(x, y^*(x)),$$

which represents that $\lim_{t\to 0} w_t^{\widetilde{j}+1}(x) = \lim_{t\to 0}(A_t^{\widetilde{j}}(x) + B_t(x))^{-1}v_t^{\widetilde{j}+1}(x)$ also exists.

Next we calculate $w^{\widetilde{j}+1}(x) = \lim_{t\to 0} w_t^{\widetilde{j}+1}(x)$ indirectly. We decompose $v_t^{\widetilde{j}+1}(x)$ and $w_t^{\widetilde{j}+1}(x)$ as $v_t^{\widetilde{j}+1}(x) = \hat{v}_t + \hat{v}_t^\perp$ and $w_t^{\widetilde{j}+1}(x) = \hat{w}_t + \hat{w}_t^\perp$ respectively, with $\hat{v}_t, \hat{w}_t \in V_t^{\widetilde{j}}(x)$ and $\hat{v}_t^\perp, \hat{w}_t^\perp \in (V_t^{\widetilde{j}}(x))^\perp$. Here $V_t^{\widetilde{j}}(x)$ defined at (E.24) is

$$V_t^j(x) = \text{span}\{\frac{t}{-h_i(x, y_t^*(x))}\nabla_y h_i(x, y_t^*(x)) : i \in \{1, ..., j\} \cap \mathcal{I}^*(x)\}.$$

A key observation is, since the term $A_t^{\widetilde{j}}(x)$ in $B_t^{\widetilde{j}}(x)$ has the potential to approach infinity, $B_t^{\widetilde{j}}(x)$ will stretch vectors in the subspace $V_t^{\widetilde{j}}(x)$ to infinite length, hence its inverse will compress vectors in $V_t^{\widetilde{j}}(x)$ to zero. We will demonstrate that the limit of $\hat{w}_t$ is zero, and then compute the limit of $\hat{w}_t^\perp$.

We will first show that the decomposition vectors $\hat{v}_t$, $\hat{w}_t$, $\hat{v}_t^\perp$, $\hat{w}_t^\perp$ also have limit when t goes to 0. Note that the projection map has a matrix representation:

$$\hat{w}_t = \text{proj}_{V_t^{\widetilde{j}}(x)} w_t^{\widetilde{j}+1}(x) = F_t(x)(F_t(x)^\top F_t(x))^{-1}F_t(x)^\top w_t^{\widetilde{j}+1}(x),$$

where $F_t(x) := \{[v_t^i(x)] : i \in \{1, ..., \widetilde{j}\} \cap \mathcal{I}^*(x)\}$ is the matrix with columns as the vectors $v_t^i(x)$ for $i \in \{1, ..., \widetilde{j}\} \cap \mathcal{I}^*(x)$. According to Lemma E.3, and Lemma B.1, $v_t^i(x)$ converges to $v^i(x)$ for $i = 1, ..., \widetilde{j}$. By LICQ assumption, $F(x)^\top F(x)$ is invertible, where $F(x) := \{[v^i(x)] : i \in \{1, ..., \widetilde{j}\} \cap \mathcal{I}^*(x)\}$ is the matrix with columns as the vectors $v^i(x)$ for $i \in \{1, ..., \widetilde{j}\} \cap \mathcal{I}^*(x)$. Note that $\lim_{t\to 0} F_t(x) = F(x)$, we confirm

$$\lim_{t\to 0} F_t(x)(F_t(x)^\top F_t(x))^{-1}F_t(x)^\top = F(x)(F(x)^\top F(x))^{-1}F(x)^\top$$

exists. Therefore $\hat{w}_t$ converges, and the convergence of $\hat{w}_t^\perp = w_t^{\widetilde{j}+1}(x) - \hat{w}_t$ follows. The convergence of $\hat{v}$ and $\hat{v}_t^\perp$ can also be obtained by a similar process. Denote $\hat{w} = \lim_{t\to 0} \hat{w}_t$, $\hat{w}^\perp = \lim_{t\to 0} \hat{w}_t^\perp$ and $\hat{v} = \lim_{t\to 0} \hat{v}_t$, $\hat{v}^\perp = \lim_{t\to 0} \hat{v}_t^\perp$.

Note that

$$\hat{v}_t + \hat{v}_t^\perp = v_t^{\widetilde{j}+1}(x)$$
$$= \left(A_t^{\widetilde{j}}(x) + B_t(x)\right) w_t^{\widetilde{j}+1}(x)$$
$$= \left(A_t^{\widetilde{j}}(x) + B_t(x)\right) \left(\hat{w}_t + \hat{w}_t^\perp\right)$$
$$= A_t^{\widetilde{j}}(x)\hat{w}_t + A_t^{\widetilde{j}}(x)\hat{w}_t^\perp + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^\perp, \tag{E.31}$$

We will compute convergence of some terms in (E.31) which are needed to prvove (E.29)

1. **Claim:** $\lim_{t\to 0} A_t^{\widetilde{j}}(x)\hat{w}_t^\perp = 0$.

   Proof. This is because

   $$\lim_{t\to 0} A_t^{\widetilde{j}}(x)\hat{w}_t^\perp = \lim_{t\to 0} \sum_{i=1}^{\widetilde{j}} \frac{1}{t} v_t^i(x) v_t^i(x)^\top \hat{w}_t^\perp$$

   $$\stackrel{(i)}{=} \lim_{t\to 0} \sum_{i\in\{1,...,\widetilde{j}\}\setminus \mathcal{I}^*(x)} \frac{1}{t} v_t^i(x) v_t^i(x)^\top \hat{w}_t^\perp$$

   $$= \lim_{t\to 0} \sum_{i\in\{1,...,\widetilde{j}\}\setminus \mathcal{I}^*(x)} \frac{t}{h_i^2(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \nabla_y h_i(x, y_t^*(x))^\top \hat{w}_t^\perp$$

   $$\stackrel{(ii)}{=} 0,$$

   where (i) is because $v_t^i(x) \in V_t^{\widetilde{j}}(x)$, $\hat{w}_t^\perp \in (V_t^{\widetilde{j}}(x))^\perp$, and $v_t^i(x)\hat{w}_t^\perp = 0$ for any $i \in \{1,...,\widetilde{j}\} \cap \mathcal{I}^*(x)$, and (ii) is because $\lim_{t\to 0}[t/h_i^2(x, y_t^*(x)) = 0/h_i^2(x, y^*(x))] = 0$ for any $i \notin \mathcal{I}^*(x)$;

2. **Claim:** $\lim_{t\to 0} \hat{w}_t = 0$.

   Proof. If $\lim_{t\to 0} \hat{w}_t =: \zeta \neq 0$, it must be that $\zeta \in V^{\widetilde{j}}(x)$ since $\hat{w}_t \in V_t^{\widetilde{j}}(x)$ for any t. Consequently,

   $$\lim_{t\to 0} \|A_t^{\widetilde{j}}(x)\hat{w}_t\| \stackrel{(i)}{=} \lim_{t\to 0} \left\| \sum_{i\in\{1,...,\widetilde{j}\}\cap \mathcal{I}^*(x)} \frac{1}{t} v_t^i(x) v_t^i(x)^\top \hat{w}_t \right\| = \infty, \tag{E.32}$$

   where (i) is because $\lim_{t\to 0}[t/h_i^2(x, y_t^*(x))] = 0$ for any $i \notin \mathcal{I}^*(x)$ and also $\lim_{t\to 0} \hat{w}_t^\perp = \hat{w}^\perp$ exists. (E.32) implies that $\lim_{t\to 0} \|A_t^{\widetilde{j}}(x)\hat{w}_t + A_t^{\widetilde{j}}(x)\hat{w}_t^\perp + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^\perp\| = \infty$, since $\lim_{t\to 0} \|A_t^{\widetilde{j}}(x)\hat{w}_t^\perp + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^\perp\|$ exists as all three terms in $\|A_t^{\widetilde{j}}(x)\hat{w}_t^\perp + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^\perp\|$ have limits. However, $\lim_{t\to 0} \|\hat{v}_t + \hat{v}_t^\perp\| = \|v^{\widetilde{j}}(x)\| < \infty$. Combining these facts with (E.31), we have

   $$\infty > \lim_{t\to 0} \|\hat{v}_t + \hat{v}_t^\perp\| = \lim_{t\to 0} \|A_t^{\widetilde{j}}(x)\hat{w}_t + A_t^{\widetilde{j}}(x)\hat{w}_t^\perp + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^\perp\| = \infty.$$

   Here is a contradiction, so we conclude $\lim_{t\to 0} \hat{w}_t = 0$;

3. **Claim:** $\lim_{t\to 0} B_t(x)\hat{w}_t = 0$.

   Proof. According to Lemma E.3, and Lemma B.1, it is easy to see $\lim_{t\to 0} B_t(x) = B(x)$. Note that $\lim_{t\to 0} \hat{w}_t = 0$, we can directly get $\lim_{t\to 0} B_t(x)\hat{w}_t = 0$.

Based on the above results, we are able to compute $w^{\widetilde{j}+1}(x) = \lim_{t\to 0} w_t^{\widetilde{j}+1}(x) = \lim_{t\to 0} \hat{w}_t^\perp = \hat{w}^\perp$ as follows:

1. Consider a linear transformation $\mathscr{B}(x)$ restricted to $(V^{\widetilde{j}}(x))^\perp$ as follows

   $$\mathscr{B}(x)(v) := \mathsf{proj}_{(V^{\widetilde{j}}(x))^\perp} B(x)v$$

   for any $v \in (V^{\widetilde{j}}(x))^\perp$. We show that $\mathscr{B}(x)$ is positive definite. Suppose there exists a non-zero vector $\eta \in (V^{\widetilde{j}}(x))^\perp$ such that $\eta^\top \mathscr{B}(x)(\eta) \leq 0$. Decomposing $B(x)\eta = \eta_1 + \eta_2$, where $\eta_1 \in V^{\widetilde{j}}(x)$ and $\eta_2 \in (V^{\widetilde{j}}(x))^\perp$, then $\eta^\top \mathscr{B}(x)(\eta) = \eta^\top \mathsf{proj}_{(V^{\widetilde{j}}(x))^\perp} B(x)\eta = \eta^\top \eta_2 \leq 0$. Hence $\eta^\top B(x)\eta = \eta^\top(\eta_1 + \eta_2) = \eta^\top \eta_2 \leq 0$, contradicting the fact that $B(x) \succeq \nabla_{yy} g(x, y^*(x))$. Therefore, $\mathscr{B}(x)$ must be a positive definite linear transformation from $(V^{\widetilde{j}}(x))^\perp$ to $(V^{\widetilde{j}}(x))^\perp$;

2. Consider the inverse of $\mathscr{B}(x)$ on $(V^{\tilde{j}}(x))^{\perp}$, denoted as $(\mathscr{B}(x))^{-1}$. We extend the linear transformation $(\mathscr{B}(x))^{-1} : (V^{\tilde{j}}(x))^{\perp} \to (V^{\tilde{j}}(x))^{\perp}$ to a linear transformation $\widetilde{\mathscr{B}}(x)$ defined on $\mathbb{R}^m$ as follows

$$\widetilde{\mathscr{B}}(x)(v) = \begin{cases} 0 & \text{if } v \in V^{\tilde{j}}(x) \\ (\mathscr{B}(x))^{-1}(v) & \text{if } v \in \left(V^{\tilde{j}}(x)\right)^{\perp}. \end{cases} \tag{E.33}$$

and $\widetilde{\mathscr{B}}(x)$ remains positive definite on $(V^{\tilde{j}}(x))^{\perp}$, as $(\mathscr{B}(x))^{-1}$ is positive definite;

3. We will show that $\lim_{t\to 0} w_t^{\tilde{j}+1}(x) = \widetilde{\mathscr{B}}(x)(v^{\tilde{j}+1})$. By calculation, we have

$$
\begin{aligned}
\hat{v} + \hat{v}^{\perp} &= \lim_{t\to 0} \hat{v}_t + \hat{v}_t^{\perp} \\
&= \lim_{t\to 0} v_t^{\tilde{j}+1}(x) \\
&= \lim_{t\to 0} \left(A_t^{\tilde{j}}(x) + B_t(x) \right) w_t^{\tilde{j}+1}(x) \\
&= \lim_{t\to 0} \left(A_t^{\tilde{j}}(x) + B_t(x) \right) \left(\hat{w}_t + \hat{w}_t^{\perp} \right) \\
&= \lim_{t\to 0} \left(A_t^{\tilde{j}}(x)\hat{w}_t + A_t^{\tilde{j}}(x)\hat{w}_t^{\perp} + B_t(x)\hat{w}_t + B_t(x)\hat{w}_t^{\perp} \right) \\
&\overset{(i)}{=} \lim_{t\to 0} A_t^{\tilde{j}}(x)\hat{w}_t + B(x)\hat{w}^{\perp} \\
&= \lim_{t\to 0} \sum_{i\in\{1,...,\tilde{j}\}} \frac{t}{h_i^2(x, y_t^*(x))} \nabla_y h_j(x, y_t^*(x)) \nabla_y h_j(x, y_t^*(x))^{\top} \hat{w}_t + B(x)\hat{w}^{\perp} \\
&\overset{(ii)}{=} \lim_{t\to 0} \sum_{i\in\{1,...,\tilde{j}\}\cap \mathcal{I}^*(x)} \frac{t}{h_i^2(x, y_t^*(x))} \nabla_y h_j(x, y_t^*(x)) \nabla_y h_j(x, y_t^*(x))^{\top} \hat{w}_t + B(x)\hat{w}^{\perp} \\
&= \sum_{i\in\{1,...,\tilde{j}\}\cap \mathcal{I}^*(x)} \left(\lim_{t\to 0} \frac{t}{h_i^2(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x))^{\top} \hat{w}_t^{\perp} \right) \nabla_y h_i(x, y^*(x)) + B(x)\hat{w}^{\perp},
\end{aligned}
$$

where (i) is because $\lim_{t\to 0} A_t^{\tilde{j}}(x)\hat{w}_t^{\perp} = 0$, $\lim_{t\to 0} B_t(x)\hat{w}_t = 0$, $\lim_{t\to 0} B_t(x) = B(x)$ which have been shown in the claims we proved before, and (ii) is because $\lim_{t\to 0} t/h_i^2(x, y_t^*(x)) = 0$ for any $i \notin \mathcal{I}^*(x)$. Note that the vector

$$\sum_{i\in\{1,...,\tilde{j}\}\cap \mathcal{I}^*(x)} \left(\lim_{t\to 0} \frac{t}{h_i^2(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x))^{\top} \hat{w}_t^{\perp} \right) \nabla_y h_i(x, y^*(x))$$

is in the subspace $V^{\tilde{j}}(x)$, then $\hat{v}^{\perp} = \mathsf{proj}_{(V^{\tilde{j}}(x))^{\perp}} B(x)\hat{w}^{\perp} = \mathscr{B}(x)(\hat{w}^{\perp})$, implying $\lim_{t\to 0} w_t^{\tilde{j}+1}(x) = \hat{w}^{\perp} = (\mathscr{B}(x))^{-1}(\hat{v}^{\perp}) = \widetilde{\mathscr{B}}(x)(v^{\tilde{j}+1})$.

To conclude, we get the result that $w^{\tilde{j}+1}(x) = \widetilde{\mathscr{B}}(x)(v^{\tilde{j}+1})$. Under the LICQ assumption and the assumption that $\tilde{j} + 1$ is active, we have $\hat{v}^{\perp} = \mathsf{proj}_{(V^{\tilde{j}}(x))^{\perp}} v^{\tilde{j}+1}(x) \neq 0$. Otherwise, if $\mathsf{proj}_{(V^{\tilde{j}}(x))^{\perp}} v^{\tilde{j}+1}(x) = 0$, then $v^{\tilde{j}+1}(x) \in V^{\tilde{j}}(x)$, which implies $v^{\tilde{j}+1}(x)$ can be expressed as the linear combination of $v^i(x)$ where $i \in \mathcal{I}(x) \cap \{1, ..., \tilde{j}\}$. Note that $v^i(x) = \lambda_i(x)\nabla_y h_i(x, y^*(x))$, and

$\lambda_i(x) > 0$ for any $i \in \mathcal{I}(x)$. Thus $\nabla_y h_{\widetilde{j}+1}(x, y^*(x))$ can be expressed as the linear combination of $\nabla_y h_i(x, y^*(x))$ where $i \in \mathcal{I}(x) \cap \{1, ..., \widetilde{j}\}$. This contradicts the LICQ assumption.

Hence,

$$
\begin{aligned}
\lim_{t \to 0} v_t^{\widetilde{j}+1}(x)^\top B_t^{\widetilde{j}}(x)^{-1} v_t^{\widetilde{j}+1}(x) &= (\hat{v} + \hat{v}^\perp)^\top (\hat{w} + \hat{w}^\perp) \\
&= (\hat{v} + \hat{v}^\perp)^\top \hat{w}^\perp \\
&\stackrel{(i)}{=} \hat{v}^{\perp\top}(\mathscr{B}(x))^{-1}(\hat{v}^\perp) \tag{E.34} \\
&\stackrel{(i)}{>} 0, \tag{E.35}
\end{aligned}
$$

where (i) is because $\hat{v}^\top \hat{w}^\perp = 0$, and (ii) is because $B_t^{\widetilde{j}}(x)^{-1}$ is positive definite in $\left(V^{\widetilde{j}}(x)\right)^\perp$. It follows that

$$
\begin{aligned}
\lim_{t \to 0}(A_t^{\widetilde{j}+1}(x) + B_t(x))^{-1} &= \lim_{t \to 0} B_t^{\widetilde{j}}(x)^{-1} - \frac{B_t^{\widetilde{j}}(x)^{-1} v_t^{\widetilde{j}+1}(x) v_t^{\widetilde{j}+1}(x)^\top B_t^{\widetilde{j}}(x)^{-1}}{t^2 + v_t^{\widetilde{j}+1}(x)^\top B_t^{\widetilde{j}}(x)^{-1} v_t^{\widetilde{j}+1}(x)} \\
&= \widetilde{\mathscr{B}}(x) - \frac{\widetilde{\mathscr{B}}(x)\left(\hat{v}^\perp(x)\right)\left(\widetilde{\mathscr{B}}(x)\left(\hat{v}^\perp(x)\right)\right)^\top}{\hat{v}^{\perp\top}(\mathscr{B}(x))^{-1}(\hat{v}^\perp)}
\end{aligned}
$$

exists. We have proved the existence of $\lim_{t \to 0}(A_t^{\widetilde{j}+1}(x) + B_t(x))^{-1}$ when $j = \widetilde{j} + 1$ is an active index.

Case 2: $j = \widetilde{j} + 1$ is an inactive index. In this case, we have

$$
\begin{aligned}
\lim_{t \to 0}(A_t^{\widetilde{j}+1}(x) + B_t(x))^{-1} &= \lim_{t \to 0}\left(\frac{t}{h_{\widetilde{j}+1}^2(x, y_t^*(x))}\nabla_y h_{\widetilde{j}+1}(x, y_t^*(x))\left(\nabla_y h_{\widetilde{j}+1}(x, y_t^*(x))\right)^\top + B_t^{\widetilde{j}}(x)\right)^{-1} \\
&= \lim_{t \to 0}\left(tE_t^{\widetilde{j}+1}(x) + B_t^{\widetilde{j}}(x)\right)^{-1},
\end{aligned}
$$

where
$$
E_t^{j}(x) := \frac{1}{h_j^2(x, y_t^*(x))}\nabla_y h_j(x, y_t^*(x))\left(\nabla_y h_j(x, y_t^*(x))\right)^\top.
$$

Note that the following limit exists

$$
\lim_{t \to 0} E_t^{\widetilde{j}+1}(x) = \frac{1}{h_{\widetilde{j}+1}^2(x, y^*(x))}\nabla_y h_{\widetilde{j}+1}(x, y^*(x))\left(\nabla_y h_{\widetilde{j}+1}(x, y^*(x))\right)^\top.
$$

Further, by inductive assumption, the following limit exists

$$
\lim_{t \to 0} B_t^{\widetilde{j}^{-1}}(x) = \lim_{t \to 0}(A_t^{\widetilde{j}}(x) + B_t(x))^{-1}.
$$

Further, assuming t is sufficiently small, applying Taylor expansion

$$
\begin{aligned}
(B_t^{\widetilde{j}}(x) + tE_t^{\widetilde{j}+1}(x))^{-1} =& B_t^{\widetilde{j}}(x)^{-1} - tB_t^{\widetilde{j}}(x)^{-1}E_t^{\widetilde{j}+1}(x)B_t^{\widetilde{j}}(x)^{-1} \\
&+ t^2 B_t^{\widetilde{j}}(x)^{-1}E_t^{\widetilde{j}+1}(x)B_t^{\widetilde{j}}(x)^{-1}E_t^{\widetilde{j}+1}(x)B_t^{\widetilde{j}}(x)^{-1} - ...
\end{aligned}
$$

$$= B_t^{\widetilde{j}-1}(x) + \mathcal{O}(t).$$

Thus, $\lim_{t\to 0}(A_t^{\widetilde{j}+1}(x) + B_t(x))^{-1} = \lim_{t\to 0}\left(B_t^{\widetilde{j}}(x)^{-1} + \mathcal{O}(t)\right) = G^{\widetilde{j}}(x)$ which is defined in (E.30).

To conclude, we have proved

$$\lim_{t\to 0}\left(A_t^k(x) + B_t(x)\right)^{-1} C_t(x)$$

exists regardless of whether j is active or inactive.

Step 2. Now go back to (E.27), we will show that $\lim_{t\to 0}(A_t^k(x) + B_t(x))^{-1}D_t(x)$ exists. By the notation (E.26), we have

$$D_t(x) = \sum_{i=1}^{k}\frac{1}{t}\left(\frac{t}{-h_j(x, y_t^*(x))}\right)^2 \nabla_y h_j(x, y_t^*(x))\nabla_x h_j(x, y_t^*(x))$$

$$= \sum_{i=1}^{k}\frac{1}{t}\frac{t}{-h_j(x, y_t^*(x))}\nabla_y h_j(x, y_t^*(x))\left(\frac{t}{-h_j(x, y_t^*(x))}\nabla_x h_j(x, y_t^*(x))\right)^\top$$

Note that by Lemma E.3

$$\lim_{t\to 0}\left(\frac{t}{-h_j(x, y_t^*(x))}\nabla_x h_j(x, y_t^*(x))\right)^\top = (\lambda_i(x)\nabla_x h_j(x, y^*(x)))^\top,$$

then it is sufficient to prove that the following limit exists for any j:

$$\lim_{t\to 0}(A_t^k(x) + B_t(x))^{-1}\frac{1}{t}\frac{t}{-h_j(x, y_t^*(x))}\nabla_y h_j(x, y_t^*(x)).$$

Without loss of generality, assume $j = k$. Otherwise, simply swap the positions of the j-th and k-th constraints.

If k is an inactive index, then we have

$$\lim_{t\to 0}\frac{1}{t}\frac{t}{-h_k(x, y_t^*(x))}\nabla_y h_k(x, y_t^*(x)) = \lim_{t\to 0}\frac{1}{-h_k(x, y_t^*(x))}\nabla_y h_k(x, y_t^*(x)) = \frac{1}{-h_k(x, y^*(x))}\nabla_y h_k(x, y^*(x)).$$

It follows that

$$\lim_{t\to 0}(A_t^k(x) + B_t(x))^{-1}\frac{1}{t}\frac{t}{-h_k(x, y_t^*(x))}\nabla_y h_k(x, y_t^*(x))$$

$$= \lim_{t\to 0}(A_t^k(x) + B_t(x))^{-1}\cdot\lim_{t\to 0}\frac{1}{t}\frac{t}{-h_k(x, y_t^*(x))}\nabla_y h_k(x, y_t^*(x))$$

$$= \lim_{t\to 0}(A_t^k(x) + B_t(x))^{-1}\frac{1}{-h_k(x, y^*(x))}\nabla_y h_k(x, y^*(x))$$

$$= G^k(x)\frac{1}{-h_k(x, y^*(x))}\nabla_y h_k(x, y^*(x)),$$

where $G^k(x)$ is defined in (E.30).

If k is an active index, then we have

$$(A_t^k(x) + B_t(x))^{-1}\frac{1}{t}\frac{t}{-h_k(x, y_t^*(x))}\nabla_y h_k(x, y_t^*(x))$$

$$\overset{(i)}{=}(A_t^k(x) + B_t(x))^{-1}\frac{1}{t}v_t^k(x)$$

$$= \left(\frac{1}{t}v_t^k(x)v_t^k(x)^\top + B_t^{k-1}(x)\right)^{-1}\frac{1}{t}v_t^k(x)$$

$$= \left(v_t^k(x)v_t^k(x)^\top + tB_t^{k-1}(x)\right)^{-1}v_t^k(x)$$

$$\overset{(i)}{=}\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x) - \frac{\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x)v_t^k(x)^\top\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x)}{1 + v_t^k(x)^\top\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x)}$$

$$= \frac{\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x)}{1 + v_t^k(x)^\top\left(tB_t^{k-1}(x)\right)^{-1}v_t^k(x)}$$

$$= \frac{\left(A_t^{k-1}(x) + B_t(x)\right)^{-1}v_t^k(x)}{t + v_t^k(x)^\top\left(A_t^{k-1}(x) + B_t(x)\right)^{-1}v_t^k(x)}. \tag{E.36}$$

In (i) we plug in $\frac{t}{-h_k(x,y_t^*(x))}\nabla_y h_k(x, y_t^*(x)) = v_t^k(x)$, and (ii) is by the Sherman-Morrison formula. This converges is because: as t converge to 0, by replacing $\tilde{j}$ with $k-1$ in (E.34), we obtain that $\lim_{t\to 0} v_t^k(x)^\top\left(A_t^{k-1}(x) + B_t(x)\right)^{-1}v_t^k(x)$ is positive.

Finally $\lim_{t\to 0}\nabla_x y_t^*(x)$ exists since

$$\nabla_x y_t^*(x) = (A_t^k(x) + B_t(x))^{-1}(C_t(x) + D_t(x))$$

and we proved the convergence of the two parts in the above two steps. $\qquad\square$

The following two lemmas are used in the proof of Lemma E.7.

Lemma E.5. *For a family of vector-valued function $\{f_t(x)\}_{t\in\mathbb{R}^+}$ defined on a compact set U, if $f_t(x)$ point-wisely, but not uniformly, converges to $f(x)$ when t goes to 0, then there exists a sequence $\{(x_j, t_j)\}_{j=1}^\infty$ such that x_j converges to $\hat{x}$ and t_j converges to 0 but $\lim_{j\to\infty} f_{t_j}(x_j)$ does not exist.*

Proof. By ϵ-δ Language, $f_t(x)$ does not converges to $f(x)$ uniformly means there exists a constant ϵ_0 such that for any T, there exists a point x and t_1, $t_2 \leq T$ satisfying $\|f_{t_1}(x) - f_{t_2}(x)\| \geq \epsilon_0$. We construct the sequence (t_j, x_j) as follows:

- For any $j \geq 1$, set $T_j = 1/j$. There exists $t'_{j,1}$, $t'_{j,2} \leq T_j$ and a point x'_j such that

$$\|f_{t'_{j,1}}(x'_j) - f_{t'_{j,2}}(x'_j)\| \geq \epsilon_0;$$

- Since U is compact, choose a subsequence of $\{(x'_j, t'_{j,1}, t'_{j,2})\}_{j=1}^\infty$, still denoted as $\{(x'_j, t'_{j,1}, t'_{j,2})\}_{j=1}^\infty$, such that $\lim_{j\to\infty} x'_j = \hat{x}$ for some point $\hat{x}$;

- For any $j \geq 1$, set $x_{2j-1} = x_{2j} = x'_j$ and $t_{2j-1} = t'_{j,1}$, $t_{2j} = t'_{j,2}$. Then we have

$$\|f_{t_{2j-1}}(x_{2j-1}) - f_{t_{2j}}(x_{2j})\| \geq \epsilon_0, \quad \forall j \geq 1. \tag{E.37}$$

It is not hard to see that $\lim_{j\to\infty} x_j = \lim_{j\to\infty} x'_j = \hat{x}_j$. However, $\{f_{t_j}(x_j)\}_{j=1}^\infty$ is not a Cauchy sequence by (E.37). We conclude that $\lim_{j\to\infty} f_{t_j}(x_j)$ does not exist. $\qquad\square$

Lemma E.6. *Suppose $\{f_t(x)\}_{t\in\mathbb{R}^+}$ is a family of matrix-valued functions defined on U. If $f_t(x)$ uniformly converges to a continuous function $f(x)$, and $\{(x_j, t_j)\}_{j=1}^\infty$ is a sequence such that $\lim_{j\to\infty} t_j = 0$ and $\lim_{j\to\infty} x_j = \hat{x}$ for some $\hat{x}$, then $\lim_{j\to\infty} f_{t_j}(x_j) = f(\hat{x})$.*

Proof. By uniformly convergence, for any $\epsilon/2 > 0$, there exists T such that $\|f_t(x) - f(x)\| \le \epsilon/2$ for any $0 < t \le T$ and $x \in U$. Since $\lim_{j\to\infty} t_j = 0$, we can find J_1 such that $t_j \le T$ for any $j \ge J_1$. Note that $f(x)$ is continuous, for any $\epsilon/2 > 0$, there exists δ such that $\|f(x) - f(\hat{x})\| \le \epsilon/2$ for any x satisfies $\|x - \hat{x}\| \le \delta$. According to $\lim_{j\to\infty} x_j = \hat{x}$, there exists T_2 such that $\|x_j - \hat{x}\| \le \delta$ for any $0 < t \le T_2$. To conclude, for any $\epsilon > 0$, we can find $J = \max\{J_1, J_2\}$ such that for any $j \ge J$ we have

$$\|f_{t_j}(x_j) - f(\hat{x})\| \le \|f_{t_j}(x_j) - f(x_j)\| + \|f(x_j) - f(\hat{x})\| \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

The proof is complete. $\qquad\square$

Lemma E.7. *Suppose Assumption 3.1 and Assumption 3.2 hold, and x is an **SCSC** point. Then there exists a compact neighborhood U of x such that $\nabla_x y_t^*(x)$ converges uniformly in U. Compact neighborhood is a compact set containing an open ball centered at x.*

Proof. We fix an **SCSC** point x_0 and prove this result at this point. We select a compact neighborhood U of x_0 satisfying the following properties:

- The active index set for any point in U is the same as that of x_0, and $\lambda_i(x) > 0$ for each active index as long as U is small enough, because $\lambda_i(x)$ is continuous, and x_0 is an **SCSC** point;

- There exists $H > 0$ such that the inequality $h_i(x, y^*(x)) \le -H < 0$ holds for any inactive index i and any x in U. This can be satisfied when H and U are small enough because $\lambda_i(x)$, $h_i(x, y^*(x))$ are continuous in x, and x_0 is an **SCSC** point;

- There exists a constant T such that the vectors $\nabla_y h_i(x, y_t^*(x))$, where $i \in \{1, ..., k\} \cap \mathcal{I}^*(x_0)$, are linear independent for any $0 < t \le T$. This because the LICQ assumption is satisfied at point $(x, y^*(x))$ and $\nabla_y h_i(x, y_t^*(x))$ converges to $\nabla_y h_i(x, y^*(x))$ uniformly by lemma E.3.

Using the notation in lemma E.4, we have

$$\nabla_x y_t^*(x) = (A_t^k(x) + B_t(x))^{-1}(C_t(x) + D_t(x)).$$

To prove $\nabla_x y_t^*(x)$ uniformly converges in U, it is sufficient to prove that $\left(A_t^k(x) + B_t(x)\right)^{-1} C_t(x)$ and $(A_t(x) + B_t(x))^{-1} D_t(x)$ converge uniformly in U separately.

Before detailed proof, we give the following uniform convergence which can all be directly obtained from Lemma B.1 and E.3, i.e. $y_t^*(x)$ converges to $y^*(x)$ uniformly, and $t/(-h_i(x, y_t^*(x)))$ converges to $\lambda_i(x)$ uniformly

$$\lim_{t\to 0} v_t^j(x) = \lim_{t\to 0} \frac{t}{-h_j(x, y_t^*(x))} \nabla_y h_j(x, y_t^*(x)) = \lambda_j(x) \nabla_y h_j(x, y^*(x)); \qquad (E.38)$$

$$\lim_{t\to 0} B_t(x) = \lim_{t\to 0} \nabla_{yy}^2 g(x, y_t^*(x)) + \sum_{i=1}^k \frac{t}{-h_j(x, y_t^*(x))} \nabla_{yy}^2 h_i(x, y_t^*(x))$$

$$= \nabla^2_{yy} g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla^2_{yy} h_i(x, y^*(x)); \tag{E.39}$$

$$\lim_{t \to 0} C_t(x) = \lim_{t \to 0} \nabla^2_{yx} g(x, y^*_t(x)) + \sum_{i=1}^{k} \frac{t}{-h_j(x, y^*_t(x))} \nabla^2_{yx} h_i(x, y^*_t(x))$$

$$= \nabla^2_{yx} g(x, y^*(x)) + \sum_{i=1}^{k} \lambda_i(x) \nabla^2_{yx} h_i(x, y^*(x)). \tag{E.40}$$

Step 1: We will prove $\left(A^k_t(x) + B_t(x)\right)^{-1} C_t(x)$ converges uniformly in U by contradiction. If it does not uniformly converge, we will identify a subsequence (x_j, t_j) such that x_j converges to $\hat{x} \in U$, and t_j converges to 0, but $\left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} C_{t_j}(x_j)$ does not converge to $\widetilde{\mathscr{B}}(\hat{x})(C(\hat{x}))$, where $\widetilde{\mathscr{B}}$ is defined in (E.33). However, we will show that for that kind of sequence, we can always have $\lim_{j \to \infty} \left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} C_{t_j}(x_j) = \widetilde{\mathscr{B}}(\hat{x})(C(\hat{x}))$, which leads a contradiction.

Denoting $C_t(x) := [C_{t,1}(x), ..., C_{t,n}(x)]$. Note that

$$\left(A^k_t(x) + B_t(x)\right)^{-1} C_t(x) = \left[\left(A^k_t(x) + B_t(x)\right)^{-1} C_{t,1}(x), \cdots, \left(A^k_t(x) + B_t(x)\right)^{-1} C_{t,n}(x)\right],$$

it is sufficient to prove that $\lim_{t \to 0} \left(A^k_t(x) + B_t(x)\right)^{-1} C_{t,i}(x)$ converges uniformly for any i. Since $C_t(x)$ converges uniformly by (E.40), we denote $C_i(x) := \lim_{t \to 0} C_{t,i}(x)$.

By our design of U, the active index set $\mathcal{I}^*(x) = \mathcal{I}^*(x_0)$ remains unchanged for any x in U. According to (E.33) in the proof of Lemma E.4, we know $\lim_{t \to 0} \left(A^k_t(x) + B_t(x)\right)^{-1} C_i(x)$ can be computed as follows:

$$\lim_{t \to 0} \left(A^k_t(x) + B_t(x)\right)^{-1} C_i(x) = \widetilde{\mathscr{B}}(x)(C_i(x)) = \begin{cases} 0 & \text{if } C_i(x) \in V^k(x_0) \\ \mathscr{B}(x)^{-1}(C_i(x)) & \text{if } C_i(x) \in \left(V^k(x_0)\right)^{\perp}. \end{cases}$$

Now we prove that $\left(A^k_t(x) + B_t(x)\right)^{-1} C_{t,i}(x)$ converges to $\widetilde{\mathscr{B}}(x)(C_i(x))$ uniformly for any i. If this does not hold for some i, by Lemma E.5, there exists a sequence $\{(x_j, t_j)\}_{j=1}^{\infty}$ such that $0 < t_j \leq T$ for any j, and $\lim_{j \to \infty} t_j = 0$, $\lim_{j \to \infty} x_j = \hat{x}$ for some point $\hat{x} \in U$, but

$$\lim_{t \to 0} \left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} C_{t_j,i}(x_j)$$

does not exist. We will demonstrate that $\left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} C_{t_j,i}(x_j)$ has a uniform bound for j so that we can find a subsequence, still denoted as $\{(x_j, t_j)\}_{j=1}^{\infty}$, such that

$$\lim_{j \to \infty} \left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} C_{t_j,i}(x_j) \tag{E.41}$$

exists but not equal $\widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x}))$. Then we will show that the limit of (E.41) must be $\widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x}))$, induces a contradiction.

Note that

$$A^k_{t_j}(x_j) + B_{t_j}(x_j) \succeq B_{t_j}(x_j) \succeq \nabla^2_{yy} g(x_j, y^*_{t_j}(x_j))$$

and $g(x, y)$ is μ_g-strongly convex. Thus, we have

$$\left(A^k_{t_j}(x_j) + B_{t_j}(x_j)\right)^{-1} \preceq \frac{1}{\mu_g} I$$

is uniformly bounded. Note that $C_{t,i}(x)$ converges to $C_i(x)$ uniformly, $C_{t,i}(x)$ can be seen as a continuous map defined on $[0,T] \times \mathcal{X}$. Therefore, $C_{t,i}(x)$ also has a uniform bound for any t and x. Therefore, $\left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right)^{-1} C_{t_j,i}(x_j)$ has a uniform bound for any j.

We choose a subsequence, still denoted as $\{(x_j, t_j)\}_{j=1}^\infty$, such that

$$\lim_{j \to \infty} w_{t_j}(x_j) := \lim_{j \to \infty} \left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right)^{-1} C_{t_j,i}(x_j) = w$$

but $w \neq \widetilde{\mathcal{B}}(\hat{x})(C_i(\hat{x}))$.

Decompose $C_{t_j,i}(x_j) = \hat{C}_j + \hat{C}_j^\perp$ and $w_{t_j}(x_j) = \hat{w}_j + \hat{w}_j^\perp$, where $\hat{C}_j, \hat{w}_j \in V_{t_j}^k(x_j)$ and $\hat{C}_j^\perp$, $\hat{w}_j^\perp \in \left(V_{t_j}^k(x_j) \right)^\perp$. Here by notation given in (E.24)

$$V_t^j(x) = \operatorname{span}\{ \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) : i \in \{1, ..., j\} \cap \mathcal{I}^*(x) \}$$

is from (E.24). We want to show that $\hat{C}_j, \hat{C}_j^\perp, \hat{w}_j$ and $\hat{w}_j^\perp$ all converges when j goes to infinity. Note that

$$\hat{C}_j = \operatorname{proj}_{V_{t_j}^k(x_j)} C_{t_j,i}(x_j)$$
$$= F_{t_j}(x_j) \left(F_{t_j}(x_j)^\top F_{t_j}(x_j) \right)^{-1} F_{t_j}(x_j)^\top C_{t_j,i}(x_j)$$

where $F_{t_j}(x_j)$ is the matrix with columns as the vectors $v_{t_j}^i(x_j) = (t_j/(-h_i(x_j, y_{t_j}^*(x_j)))) \nabla_y h_i(x_j, y_{t_j}^*(x_j))$ for $i \in \mathcal{I}^*(x)$. Since we have the uniform convergence by (E.40), and clearly $C_i(x)$ is continuous in x, by Lemma E.6, we get $\lim_{j \to \infty} C_{t_j,i}(x_j) = \hat{C}_i(\hat{x})$. It is sufficient to prove

$$F_{t_j}(x_j) \left(F_{t_j}(x_j)^\top F_{t_j}(x_j) \right)^{-1} F_{t_j}(x_j)^\top$$

also converges. According to (E.38), we know $v_t^i(x)$ converges to $\lambda_i(x) \nabla_y h_i(x, y^*(x))$ uniformly. Note that $\lambda_i(x) \nabla_y h_i(x, y^*(x))$ is continuous in x. By Lemma E.6, $\lim_{j \to \infty} v_{t_j}^i(x_j) = v^i(\hat{x})$. This implies $\lim_{j \to \infty} F_{t_j}(x_j) = F(\hat{x})$ where $F(\hat{x})$ is the matrix with columns as the vectors $\lambda_i(\hat{x}) \nabla_y h_i(\hat{x}, y^*(\hat{x}))$ for $i \in \mathcal{I}^*(x)$. By LICQ assumption $F(\hat{x})^\top F(\hat{x})$ is invertible, we obtain $F_{t_j}(x_j) \left(F_{t_j}(x_j)^\top F_{t_j}(x_j) \right)^{-1} F_{t_j}(x_j)^\top$ converges to $F(\hat{x}) \left(F(\hat{x})^\top F(\hat{x}) \right)^{-1} F(\hat{x})^\top$ when j goes to infinity. Therefore,

$$\begin{aligned}
\lim_{j \to \infty} \hat{C}_j &= \lim_{j \to \infty} \left(\operatorname{proj}_{V_{t_j}^k(x_j)} C_{t_j,i}(x_j) \right) \\
&= \lim_{j \to \infty} \left(F_{t_j}(x_j) \left(F_{t_j}(x_j)^\top F_{t_j}(x_j) \right)^{-1} F_{t_j}(x_j)^\top C_{t_j,i}(x_j) \right) \\
&= \lim_{j \to \infty} \left(F_{t_j}(x_j) \left(F_{t_j}(x_j)^\top F_{t_j}(x_j) \right)^{-1} F_{t_j}(x_j)^\top \right) \cdot \lim_{j \to \infty} \left(C_{t_j,i}(x_j) \right) \\
&= F(\hat{x}) \left(F(\hat{x})^\top F(\hat{x}) \right)^{-1} F(\hat{x})^\top C_i(\hat{x}) \\
&= \operatorname{proj}_{V^k(\hat{x})} C_i(\hat{x}) \\
&=: \hat{C}.
\end{aligned}$$

By the same method, $\lim_{j \to \infty} \hat{w}_j =: \hat{w}$, $\lim_{j \to \infty} \hat{C}_j^\perp =: \hat{C}^\perp = C_i(\hat{x}) - \hat{C}$, and $\lim_{j \to \infty} \hat{w}_j^\perp =: \hat{w}^\perp = w - \hat{w}$ all exist.

Next, we will show that

$$\lim_{j\to\infty} \left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right)^{-1} (C_{t_j,i}(x_j)) = \lim_{j\to\infty} w_{t_j}(x_j) = \widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x})),$$

to contradict our assumption that

$$\lim_{j\to\infty} \left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right)^{-1} (C_{t_j,i}(x_j)) \neq \widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x})).$$

Note that

$$
\begin{aligned}
\hat{C}_j + \hat{C}_j^{\perp} &= C_{t_j,i}(x_j) \\
&= \left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right) w_{t_j}(x_j) \\
&= \left(A_{t_j}^k(x_j) + B_{t_j}(x_j) \right) \left(w_j + w_j^{\perp}+ \right) \\
&= A_{t_j}^k(x_j)\hat{w}_j + A_{t_j}^k(x_j)\hat{w}j^{\perp} + B_{t_j}(x_j)\hat{w}_j + B_{t_j}(x_j)\hat{w}_j^{\perp}.
\end{aligned}
$$

Similar to the proof of Lemma E.4, we will prove the following three results:

1. **Claim:** $\lim_{j\to\infty} A_{t_j}^k(x_j)\hat{w}_j^{\perp} = 0.$

   Proof. This is because

   $$
   \begin{aligned}
   \lim_{j\to\infty} A_{t_j}^k(x_j)\hat{w}_j^{\perp} &= \lim_{j\to\infty} \sum_{i=1}^{k} \frac{1}{t} v_{t_j}^i(x_j) v_{t_j}^i(x_j)^{\top} \hat{w}_j^{\perp} \\
   &\overset{(i)}{=} \lim_{j\to\infty} \sum_{i\in\{1,\dots,k\}\setminus\mathcal{I}^*(x_0)} \frac{1}{t} v_{t_j}^i(x_j) v_{t_j}^i(x_j)^{\top} \hat{w}_j^{\perp} \\
   &= \lim_{j\to\infty} \sum_{i\in\{1,\dots,k\}\setminus\mathcal{I}^*(x_0)} \frac{t}{h_i^2(x_j, y_{t_j}^*(x_j))} \nabla_y h_i(x_j, y_{t_j}^*(x_j)) \nabla_y h_i(x_j, y_{t_j}^*(x_j))^{\top} \hat{w}_j^{\perp} \\
   &\overset{(ii)}{=} 0,
   \end{aligned}
   $$

   where (i) is because $v_{t_j}^j(x_j) \in V_t^j(x)$, $\hat{w}_j^{\perp} \in (V_t^j(x))^{\perp}$, and then $v_{t_j}^j(x_j)\hat{w}_j^{\perp} = 0$ for any $i \in \{1,\dots,k\} \cap \mathcal{I}^*(x_0)$, and (ii) is because

   $$\lim_{j\to\infty} (t_j/h_i^2(x_j, y_{t_j}^*(x_j))) = (\lim_{j\to\infty} t_j)/h_i^2(\hat{x}, y^*(\hat{x})) \leq 0/H^2 = 0, \quad \forall i \notin \mathcal{I}^*(x_0).$$

   Here H, we defined before, is the lower bound of $-h_i(x, y^*(x))$ with x in U and inactive index i;

2. **Claim:** $\lim_{j\to\infty} \hat{w}_j = 0.$

   Proof. We prove this by contradiction. If $\lim_{j\to\infty} \hat{w}_j =: \zeta \neq 0$, it must be that $\zeta \in V^k(\hat{x})$ since $\hat{w}_j \in V_{t_j}^k(x_j)$ for any j. Consequently,

   $$\lim_{j\to\infty} \|A_{t_j}^k(x_j)\hat{w}_j\| \overset{(i)}{=} \lim_{j\to\infty} \left\| \sum_{i\in\{1,\dots,k\}\cap\mathcal{I}^*(x_0)} \frac{1}{t_j} v_{t_j}^i(x_j) v_{t_j}^i(x_j)^{\top} \hat{w}_j \right\| = \infty,$$

   where (i) is because $\lim_{j\to\infty}(t_j/h_i^2(x_j, y_{t_j}^*(x_j))) = 0$ for any $i \notin \mathcal{I}^*(x_0)$ which we have proven in the first Claim.

63

By (E.39), we have $B_t(x)$ converges to $B(x)$ uniformly. Therefore, by the continuity of $B(x)$ in x and Lemma E.6, $\lim_{j\to\infty} B_{t_j}(x_j) = B(\hat{x})$, implying that $\lim_{j\to\infty} \|A_{t_j}^k(x_j)\hat{w}_j^\perp + B_{t_j}(x_j)\hat{w}_j + B_{t_j}(x_j)\hat{w}_j^\perp\| < \infty$. Thus $\lim_{j\to\infty} \|A_{t_j}^k(x_j)\hat{w}_j + A_{t_j}^k(x_j)\hat{w}_j^\perp + B_{t_j}(x_j)\hat{w}_j + B_{t_j}(x_j)\hat{w}_j^\perp\| = \infty$. However, $\lim_{j\to\infty} \|\hat{C}_j + \hat{C}_j^\perp\| = \|C_i(\hat{x})\| < \infty$ since $C_i(x)$ is a continuous map defined on $\mathcal{X}$. This induces the following contradiction

$$\infty > \lim_{j\to\infty} \|\hat{C}_j + \hat{C}_j^\perp\| = \lim_{j\to\infty} \|A_{t_j}^k(x_j)\hat{w}_j + A_{t_j}^k(x_j)\hat{w}_j^\perp + B_{t_j}(x_j)\hat{w}_j + B_{t_j}(x_j)\hat{w}_j^\perp\| = \infty.$$

Therefore, we conclude $\lim_{j\to\infty} \hat{w}_j = 0$;

3. **Claim:** $\lim_{j\to\infty} B_{t_j}(x_j)\hat{w}_j = 0$.

   Proof. Note that $\lim_{j\to\infty} B_{t_j}(x_j) = B(\hat{x})$ and $\lim_{j\to\infty} \hat{w}_j = 0$, we can directly get $\lim_{j\to\infty} B_{t_j}(x_j)\hat{w}_j = 0$.

Given that $A_{t_j}^k(x_j)\hat{w}_j \in V_{t_j}^k(x_j)$ for any j, we can obtain

$$\begin{aligned}
\hat{C}^\perp &= \lim_{j\to\infty} \hat{C}_j^\perp \\
&= \lim_{j\to\infty} \mathsf{proj}_{\left(V_{t_j}^k(x_j)\right)^\perp} \left(A_{t_j}^k(x_j)\hat{w}_j + A_{t_j}^k(x_j)\hat{w}_j^\perp + B_{t_j}(x_j)\hat{w}_j + B_{t_j}(x_j)\hat{w}_j^\perp\right) \\
&= \mathsf{proj}_{\left(V^k(\hat{x})\right)^\perp} \left(B(\hat{x})\hat{w}^\perp\right),
\end{aligned}$$

where $\hat{x}$ is the limit of x_j. It follows that

$$\lim_{j\to\infty} \left(A_{t_j}^k(x_j) + B_{t_j}(x_j)\right)^{-1} (C_{t_j,i}(x_j)) = \lim_{j\to\infty} w_{t_j}(x_j) = \lim_{j\to\infty} \hat{w}_j + \hat{w}_j^\perp = \hat{w}^\perp = \widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x})).$$

This contradicts our assumption that

$$\lim_{j\to\infty} \left(A_{t_j}^k(x_j) + B_{t_j}(x_j)\right)^{-1} (C_{t_j,i}(x_j)) \neq \widetilde{\mathscr{B}}(\hat{x})(C_i(\hat{x})).$$

So we proved the first part that $\left(A_t^k(x) + B_t(x)\right)^{-1} C_t(x)$ converges uniformly.

Step 2: Now we prove the second part that $(A_t(x) + B_t(x))^{-1} D_t(x)$ converges uniformly. Recall that

$$D_t(x) = \sum_{i=1}^k \frac{1}{t} \frac{t^2}{h_i^2(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \left(\nabla_x h_i(x, y_t^*(x))\right)^\top.$$

We can obtain the uniform convergence of $(t/(-h_i(x, y_t^*(x)))) \left(\nabla_x h_i(x, y_t^*(x))\right)^\top$ similar to (E.38). $(A_t(x) + B_t(x))^{-1} D_t(x)$ can be rewritten as

$$\begin{aligned}
&\left(A_t^k(x) + B_t(x)\right)^{-1} D_t(x) \\
&= \left((A_t^k(x) + B_t(x))^{-1} \frac{1}{t} \frac{t}{-h_j(x, y_t^*(x))} \nabla_y h_j(x, y_t^*(x))\right) \left(\frac{t}{-h_j(x, y_t^*(x))} \nabla_x h_i(x, y_t^*(x))\right)^\top
\end{aligned}$$

Then it is sufficient to prove that

$$(A_t^k(x) + B_t(x))^{-1} \frac{1}{t} \frac{t}{-h_j(x, y_t^*(x))} \nabla_y h_j(x, y_t^*(x))$$

64

converges uniformly for any j. Without loss of generality, in the following proof, we only consider the case $j = k$. Otherwise, simply swap the positions of the j-th and k-th constraints.

If k is an inactive index, by the assumption at the beginning that $h_k(x, y^*(x)) \leq -H$, we have

$$\frac{1}{t} \frac{t}{-h_k(x, y_t^*(x))} \nabla_y h_k(x, y_t^*(x)) = \frac{1}{-h_k(x, y_t^*(x))} \nabla_y h_k(x, y_t^*(x))$$

converges to $(1/(-h_k(x, y^*(x)))) \nabla_y h_k(x, y^*(x))$ uniformly. We can obtain the uniform convergence of the following term

$$(A_t^k(x) + B_t(x))^{-1} \frac{1}{t} \frac{t}{-h_k(x, y_t^*(x))} \nabla_y h_k(x, y_t^*(x))$$

by replacing $C_{t,i}(x)$ with $(1/(-h_k(x, y_t^*(x)))) \nabla_y h_k(x, y_t^*(x))$ in Step 1,

If k is an active index, following (E.36) in the proof of lemma E.4, we have

$$(A_t^k(x) + B_t(x))^{-1} \frac{1}{t} \frac{t}{-h_k(x, y_t^*(x))} \nabla_y h_k(x, y_t^*(x))$$

$$= \frac{\left(A_t^{k-1}(x) + B_t(x)\right)^{-1} v_t^k(x)}{t + v_t^k(x)^\top \left(A_t^{k-1}(x) + B_t(x)\right)^{-1} v_t^k(x)},$$

where we define $v_t^k(x)$ in (E.22) as

$$v_t^k(x) = \frac{t}{-h_k(x, y_t^*(x))} \nabla_y h_k(x, y_t^*(x)).$$

We just need to prove $\left(A_t^{k-1}(x) + B_t(x)\right)^{-1} v_t^k(x)$ converges uniformly, and

$$\lim_{t \to 0} v_t^k(x)^\top \left(A_t^{k-1}(x) + B_t(x)\right)^{-1} v_t^k(x) \neq 0$$

for any x in U. The first result has already been proven in Step 1 just by replacing k with $k-1$ and $C_{t,i}(x)$ with $v_t^k(x)$. The second result has also been proven in (E.34) by replacing $\widetilde{j}$ with $k-1$.

To conclude, we proved that $\nabla_x y_t^*(x)$ converges uniformly in U. $\qquad \square$

Next, we prove the relation of Jacobians under the strongly convex setting when t approaches 0. **Proof of Theorem 3.4:** We fix an **SCSC** point x_0 and prove this theorem at this point. It is sufficient to prove that $\lim_{t \to 0} \frac{\partial (y_t^*(x))_i}{\partial x_j} = \frac{\partial (y^*(x))_i}{\partial x_j}$ at **SCSC** point x_0 for any i, j. By Lemma B.1, we have $y_t^*(x_0)$ converges to $y^*(x_0)$, and by Lemma E.7 there exists a neighborhood U of x_0 such that $\nabla_x y_t^*(x)$ converges uniformly in U, we can interchange the order of limit and derivatives

$$\lim_{t \to 0} \frac{\partial (y_t^*(x_0))_i}{\partial x_j} = \frac{\partial \lim_{t \to 0} (y_t^*(x_0))_i}{\partial x_j} = \frac{\partial (y^*(x_0))_i}{\partial x_j}.$$

Therefore, we completed the proof.

F Convergence Analysis of Algorithms

F.1 Proof of Proposition 4.1

Through directly computing the Hessian, we get

$$\nabla^2_{yy}\widetilde{g}_t(x,y) = \nabla^2_{yy}g(x,y) + t\sum_{i=1}^{k}\left(\frac{\nabla^2_{yy}h_i(x,y)}{-h_i(x,y)} + \frac{\nabla_y h_i(x,y)\nabla_y h_i(x,y)^\top}{h_i^2(x,y)}\right).$$

If $g(x,y)$ is μ_g-strongly convex in y, then

$$\nabla^2_{yy}\widetilde{g}_t(x,y) \succeq \mu_g I,$$

which means $\widetilde{g}_t(x,y)$ is μ_g-strongly convex.

If $h_i(x,y)$ are linear in y for any i, define $A(x) = (\nabla_y h_1(x,y),\cdots,\nabla_y h_k(x,y))$. In the proof in Proposition 3.1, we have proven that $A(x)A(x)^\top$ is positive for any x. The smallest eigenvalue $\sigma(x)$ is continuous depends on x, then we obtain that $\sigma = \min_{x\in\mathcal{X}}\sigma(x) > 0$. It is also easy to see $H = \sup_{i\in\{1,\ldots,k\},x\in\mathcal{X},y\in\mathcal{Y}_x} -h_i(x,y) < \infty$. Therefore

$$\begin{aligned}\nabla^2_{yy}\widetilde{g}_t(x,y) &\succeq t\sum_{i=1}^{k}\frac{\nabla_y h_i(x,y)\nabla_y h_i(x,y)^\top}{h_i^2(x,y)} \\ &\succeq t\frac{\sum_{i=1}^{k}\nabla_y h_i(x,y)\nabla_y h_i(x,y)^\top}{H^2} \\ &= t\frac{A(x)^\top A(x)}{H^2} \\ &\succeq t\frac{\sigma}{H^2}I\end{aligned}$$

which means $\widetilde{g}_t(x,y)$ is $t\frac{\sigma}{H^2}$-strongly convex.

F.2 Proof of Theorem 4.1

Denote $\overline{y} = \arg\min_{y\in\mathcal{Y}(x)}\{\max_{i\in\{1,\ldots,k\}}h_i(x,y)\}$. From our assumption

$$\arg\min_{y\in\mathcal{Y}(x)}\{\max_{i\in\{1,\ldots,k\}}h_i(x,y)\} \leq -d,$$

it is clear that $h_i(x,\overline{y}) \leq -d$ for any i. Our main idea of the proof is to reduce the problem to a one-dimensional case.

If $y_t^*(x) = \overline{y}$, then let $m = d$. Then we directly have $h_i(x,y_t^*(x)) \leq -d$ for any i.

If $y_t^*(x) \neq \overline{y}$. Consider the line $y_s = \overline{y} + s\boldsymbol{v}$, where $\boldsymbol{v} = \frac{y_t^*(x)-\overline{y}}{\|y_t^*(x)-\overline{y}\|}$ and define

$$\begin{aligned}\widetilde{g}_x(s) &:= \widetilde{g}_t(x,y_s) \\ &= g(x,\overline{y}+s\boldsymbol{v}) - t\sum_{i=1}^{k}\log\left(-h_i(x,\overline{y}+s\boldsymbol{v})\right).\end{aligned}$$

It is not hard to see $s^* := \|y_t^*(x) - \overline{y}\|$ is the minimizer of $\widetilde{g}_x(s)$. We evaluate the bound of $h_j(x,y_t^*(x))$ for any index j based on the range of values for $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}}$, where $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}}$ is the directional derivative of $h_j(x,\cdot)$ along direction $\boldsymbol{v}$ at point y_{s^*}.

Case 1: If $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}} \le 0$, by convexity of $h_j(x,y_s)$ in s, $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_s}$ is monotonically increasing in s, thus we can get $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_s} \le 0$ for any $s \in [0,s^*]$. Therefore, we have

$$h_j(x,y_{s^*}) - h_j(x,\overline{y}) = \int_0^{s^*} \frac{d}{ds}h_j(x,\overline{y}+s\boldsymbol{v})ds \le 0,$$

implying $h_j(x,y_{s^*}) \le h_j(x,\overline{y}) \le -d$.

Case 2: If $0 < \frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}} \le \frac{d}{4R}$, where R is from Assumption 4.3(3), denote $s_j^* = \operatorname{argmin}_{\{s:y_s\in\mathcal{Y}(x)\}} h_j(x,y_s)$. From the convexity of $h_j(x,y_s)$ in s, we know $s^* \ge s_j^*$. Otherwise, by convexity of $h_j(x,y_s)$ in s, $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_s}$ is monotonically increasing in s, $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_s} \ge 0$ for any $s \in [s^*,s_j^*]$, implying $h_j(x,y_{s^*}) \le h_j(x,y_{s_j^*})$. This contradicts $s_j^* = \operatorname{argmin}_{\{s:y_s\in\mathcal{Y}(x)\}} h_j(x,y_s)$. Therefore, we obtain

$$h_j(x,y_{s^*}) - h_j(x,y_{s_j^*}) = \int_{s_j^*}^{s^*} \frac{d}{ds}h_j(x,\overline{y}+s\boldsymbol{v})ds \overset{(i)}{\le} (s^*-s_j^*)\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}} \overset{(ii)}{\le} \frac{d}{2}.$$

In (i) we used the result that $\frac{d}{ds}h_j(x,\overline{y}+s\boldsymbol{v}) \le \frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}}$ for any $s_j^* \le s \le s^*$. This is because, by convexity of $h_j(x,y_s)$ in s, $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_s}$ is monotonically increasing in s. In (ii) we use the fact that $s^* - s_j^* \le 2R$ from Assumption 4.3(3) and $\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}} \le \frac{d}{4R}$. Note that $h_j(x,y_{s_j^*}) \le h_j(x,\overline{y}) \le -d$, we have $h_j(x,y_t^*(x)) = h_j(x,y_{s^*}) \le -\frac{d}{2}$.

Case 3: If $\frac{d}{4R} < \frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}}$, we divide the index set $I = \{1,...,k\}$ into two subsets, $I_1 = \{i_1 \in I : \frac{d}{d\boldsymbol{v}}h_{i_1}(x,y)|_{y=y_{s^*}} > 0\}$ and $I_2 = \{i_2 \in I : \frac{d}{d\boldsymbol{v}}h_{i_2}(x,y)|_{y=y_{s^*}} \le 0\}$. Note that the optimality condition of $\widetilde{g}_x(s)$ is $\frac{d}{ds}\widetilde{g}_x(s)|_{s=s^*} = 0$, which can be expressed as

$$\frac{d}{d\boldsymbol{v}}g(x,y)|_{y=y_{s^*}} + t\sum_{i=1}^k \frac{\frac{d}{d\boldsymbol{v}}h_i(x,y)|_{y=y_{s^*}}}{-h_i(x,y_{s^*})} = 0. \tag{F.1}$$

Rearranging the terms of (F.1), note that $-h_j(x,y_{s^*}) = -h_j(x,y_t^*(x)) > 0$, we obtain

$$\frac{td}{-4Rh_j(x,y_{s^*})} < t\frac{\frac{d}{d\boldsymbol{v}}h_j(x,y)|_{y=y_{s^*}}}{-h_j(x,y_{s^*})} = -\frac{d}{d\boldsymbol{v}}g(x,y)|_{y=y_{s^*}} - t\sum_{i\in I\setminus\{j\}} \frac{\frac{d}{d\boldsymbol{v}}h_i(x,y)|_{y=y_{s^*}}}{-h_i(x,y_{s^*})}$$

$$\overset{(i)}{\le} L_g - t\sum_{i_1\in I_1\setminus\{j\}} \frac{\frac{d}{d\boldsymbol{v}}h_{i_1}(x,y)|_{y=y_{s^*}}}{-h_{i_1}(x,y_{s^*})} - t\sum_{i_2\in I_2\setminus\{j\}} \frac{\frac{d}{d\boldsymbol{v}}h_{i_2}(x,y)|_{y=y_{s^*}}}{-h_{i_2}(x,y_{s^*})}$$

$$\overset{(ii)}{\le} L_g - t\sum_{i_2\in I_2\setminus\{j\}} \frac{\frac{d}{d\boldsymbol{v}}h_{i_2}(x,y)|_{y=y_{s^*}}}{-h_{i_2}(x,y_{s^*})}$$

$$\overset{(iii)}{\le} L_g + \frac{tkL_h}{d}.$$

In (i) we utilize $-\frac{d}{d\boldsymbol{v}}g(x,y)|_{y=y_{s^*}} \le L_g$ by Assumption 4.4(3). (ii) is due to

$$-t\sum_{i_1\in I_1\setminus\{j\}} \frac{\frac{d}{d\boldsymbol{v}}h_{i_1}(x,y)|_{y=y_{s^*}}}{-h_{i_1}(x,y_{s^*})} \le 0.$$

In (iii) we use the fact that I_2 is designed such that for all $i_2 \in I_2$, Case 1 is satisfied, so $-h_{i_2}(x,y_{s^*}) \le d$; further we have used the fact that for $i_2 \in I_2$, $0 < -\frac{d}{d\boldsymbol{v}}h_{i_2}(x,y)|_{y=y_{s^*}} \le L_h$ where the second inequality comes from Assumption 4.4(6). Note that $t \le T$, we have

$$h_j(x,y_{s^*}) \le -t\frac{d^2}{4dRL_g + 4RTkL_h}.$$

To conclude, the following holds for any j

$$h_j(x, y_{s^*}) \leq -\min\{t\frac{d^2}{4dRL_g + 4RTkL_h}, \frac{d}{2}\}.$$

It is noteworthy that we do not need the convexity of $g(x, y)$ and $\widetilde{g}_t(x, y)$ in this proof.

F.3 Proof of Proposition 4.2

Through directly computing the Hessian, we get

$$\nabla^2_{yy}\widetilde{g}_t(x, y) = \nabla^2_{yy}g(x, y) + t\sum_{i=1}^k \left(\frac{\nabla^2_{yy}h_i(x, y)}{-h_i(x, y)} + \frac{\nabla_y h_i(x, y)\nabla_y h_i(x, y)^\top}{h_i^2(x, y)} \right).$$

Note that we only consider $y \in \mathcal{Y}_m(x)$, i.e. $h_i(x, y) \leq -m$ for any i. Therefore, we have the following estimate

$$\left\| \nabla^2_{yy}\widetilde{g}_t(x, y) \right\| \leq \left\| \nabla^2_{yy}g(x, y) \right\| + t\sum_{i=1}^k \left(\frac{\left\| \nabla^2_{yy}h_i(x, y) \right\|}{-h_i(x, y)} + \frac{\left\| \nabla_y h_i(x, y) \right\|^2}{h_i^2(x, y)} \right)$$

$$\overset{(i)}{\leq} \overline{L}_g + \frac{tk\overline{L}_h}{m} + \frac{tkL_h^2}{m^2},$$

where (i) is from the assumption that $h_i(x, y) \leq -m$, and Assumption 4.4 that $\left\| \nabla^2_{yy}g(x, y) \right\| \leq \overline{L}_g$, $\left\| \nabla^2_{yy}h_i(x, y) \right\| \leq \overline{L}_h$ and $\left\| \nabla_y h_i(x, y) \right\| \leq L_h$ for any i.

F.4 Proof of Lemma 4.1

By direct computing

$$\left\| \nabla^2_{yy}\widetilde{g}_t(x_1, y_1) - \nabla^2_{yy}\widetilde{g}_t(x_2, y_2) \right\|$$

$$= \left\| \nabla^2_{yy}g(x_1, y_1) + t\sum_{i=1}^k \left(\frac{\nabla^2_{yy}h_i(x_1, y_1)}{-h_i(x_1, y_1)} + \frac{\nabla_y h_i(x_1, y_1)\nabla_y h_i(x_1, y_1)^\top}{h_i^2(x_1, y_1)} \right) \right.$$

$$\left. - \left(\nabla^2_{yy}g(x_2, y_2) + t\sum_{i=1}^k \left(\frac{\nabla^2_{yy}h_i(x_2, y_2)}{-h_i(x_2, y_2)} + \frac{\nabla_y h_i(x_2, y_2)\nabla_y h_i(x_2, y_2)^\top}{h_i^2(x_2, y_2)} \right) \right) \right\|$$

$$= \left\| (\nabla^2_{yy}g(x_1, y_1) - \nabla^2_{yy}g(x_2, y_2)) + t\sum_{i=1}^k \left(\frac{\nabla^2_{yy}h_i(x_1, y_1)}{-h_i(x_1, y_1)} - \frac{\nabla^2_{yy}h_i(x_2, y_2)}{-h_i(x_2, y_2)} \right) \right.$$

$$\left. + t\sum_{i=1}^k \left(\frac{\nabla_y h_i(x_1, y_1)\nabla_y h_i(x_1, y_1)^\top}{h_i^2(x_1, y_1)} - \frac{\nabla_y h_i(x_2, y_2)\nabla_y h_i(x_2, y_2)^\top}{h_i^2(x_2, y_2)} \right) \right\|$$

$$= \left\| (\nabla^2_{yy}g(x_1, y_1) - \nabla^2_{yy}g(x_2, y_2)) + t\sum_{i=1}^k \frac{h_i(x_1, y_1)\nabla^2_{yy}h_i(x_2, y_2) - h_i(x_2, y_2)\nabla^2_{yy}h_i(x_1, y_1)}{h_i(x_1, y_1)h_i(x_2, y_2)} \right.$$

$$\left. + t\sum_{i=1}^k \frac{h_i^2(x_2, y_2)\nabla_y h_i(x_1, y_1)\nabla_y h_i(x_1, y_1)^\top - h_i^2(x_1, y_1)\nabla_y h_i(x_2, y_2)\nabla_y h_i(x_2, y_2)^\top}{h_i^2(x_1, y_1)h_i^2(x_2, y_2)} \right\|. \quad \text{(F.2)}$$

We estimate three terms of (F.2) separately. For the first term, we utilize Assumption 4.4(5) and obtain

$$\left\| \nabla^2_{yy}g(x_1, y_1) - \nabla^2_{yy}g(x_2, y_2) \right\| \leq \overline{\overline{L}}_g \|(x_1, y_1) - (x_2, y_2)\|.$$

68

For the second, we have

$$\left\| t\sum_{i=1}^{k} \frac{h_i(x_1,y_1)\nabla_{yy}^2 h_i(x_2,y_2) - h_i(x_2,y_2)\nabla_{yy}^2 h_i(x_1,y_1)}{h_i(x_1,y_1)h_i(x_2,y_2)} \right\|$$

$$\leq t\sum_{i=1}^{k} \left\| \frac{h_i(x_1,y_1)\nabla_{yy}^2 h_i(x_2,y_2) - h_i(x_2,y_2)\nabla_{yy}^2 h_i(x_1,y_1)}{h_i(x_1,y_1)h_i(x_2,y_2)} \right\|$$

$$= t\sum_{i=1}^{k} \left\| \frac{h_i(x_1,y_1)\nabla_{yy}^2 h_i(x_2,y_2) - h_i(x_1,y_1)\nabla_{yy}^2 h_i(x_1,y_1) + h_i(x_1,y_1)\nabla_{yy}^2 h_i(x_1,y_1) - h_i(x_2,y_2)\nabla_{yy}^2 h_i(x_1,y_1)}{h_i(x_1,y_1)h_i(x_2,y_2)} \right\|$$

$$\leq t\sum_{i=1}^{k} \left(\left\| \frac{\nabla_{yy}^2 h_i(x_2,y_2) - \nabla_{yy}^2 h_i(x_1,y_1)}{h_i(x_2,y_2)} \right\| + \left\| \frac{\nabla_{yy}^2 h_i(x_1,y_1)\,(h_i(x_1,y_1) - h_i(x_2,y_2))}{h_i(x_1,y_1)h_i(x_2,y_2)} \right\| \right)$$

$$\overset{(i)}{\leq} tk\left(\frac{\overline{\overline{L}}_h}{m} + \frac{\overline{L}_h L_h}{m^2} \right) \|(x_1,y_1) - (x_2,y_2)\|, \tag{F.3}$$

where (i) is because $h_i(x_1,y_1),\, h_i(x_2,y_2) \leq -m$, $\|\nabla_{yy}^2 h_i(x_2,y_2) - \nabla_{yy}^2 h_i(x_1,y_1)\| \leq \overline{\overline{L}}_h\|(x_1,y_1) - (x_2,y_2)\|$, $\|\nabla_{yy}^2 h_i(x_1,y_1)\| \leq \overline{L}_h$, and $|h_i(x_1,y_1) - h_i(x_2,y_2)| \leq L_h\|(x_1,y_1) - (x_2,y_2)\|$.

For the final term, we have

$$\left\| t\sum_{i=1}^{k} \frac{h_i^2(x_2,y_2)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_2,y_2)\nabla_y h_i(x_2,y_2)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$\leq t\sum_{i=1}^{k} \left\| \frac{h_i^2(x_2,y_2)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_2,y_2)\nabla_y h_i(x_2,y_2)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$= t\sum_{i=1}^{k} \left(\left\| \frac{h_i^2(x_2,y_2)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\| \right.$$

$$+ \left\| \frac{h_i^2(x_1,y_1)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_2,y_2)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$+ \left. \left\| \frac{h_i^2(x_1,y_1)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_2,y_2)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_2,y_2)\nabla_y h_i(x_2,y_2)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\| \right) \tag{F.4}$$

$$\overset{(i)}{\leq} \left(2tk\frac{L_h^3}{m^3} + 2tk\frac{L_h \overline{L}_h}{m^2} \right) \|(x_1,y_1) - (x_2,y_2)\|.$$

In (i), when evaluating the first term, we use the following estimate

$$\left\| \frac{h_i^2(x_2,y_2)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top - h_i^2(x_1,y_1)\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$= \left\| \frac{(h_i(x_2,y_2) + h_i(x_1,y_1))(h_i(x_2,y_2) - h_i(x_1,y_1))\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top}{h_i^2(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$= \left\| \frac{(h_i(x_2,y_2) - h_i(x_1,y_1))\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top}{h_i^2(x_1,y_1)h_i(x_2,y_2)} + \frac{(h_i(x_2,y_2) - h_i(x_1,y_1))\nabla_y h_i(x_1,y_1)\nabla_y h_i(x_1,y_1)^\top}{h_i(x_1,y_1)h_i^2(x_2,y_2)} \right\|$$

$$\leq 2tk\frac{L_h^3}{m^3}\|(x_1,y_1) - (x_2,y_2)\|.$$

In conclusion, we have

$$\left\|\nabla^2_{yy}\widetilde{g}_t(x_1,y_1) - \nabla^2_{yy}\widetilde{g}_t(x_2,y_2)\right\|$$

$$\leq \left(\overline{\overline{L}}_g + tk\left(\frac{\overline{\overline{L}}_h}{m} + \frac{\overline{L}_hL_h}{m^2} + \frac{2L_h^3}{m^3} + \frac{2L_h\overline{L}_h}{m^2}\right)\right)\|(x_1,y_1) - (x_2,y_2)\|.$$

We can also derive that $\|\nabla^2_{xy}\widetilde{g}_t(x_1,y_1) - \nabla^2_{xy}\widetilde{g}_t(x_2,y_2)\| \leq \overline{\overline{L}}_{\widetilde{g}_t,m}\|(x_1,y_1) - (x_2,y_2)\|$ using the same process.

F.5 Proof of Lemma 4.2

By Assumption 4.4(6), i.e. the Lipschitz continuity of $h_i(x,y)$, we have $h_i(x+\Delta x, y) \leq -\frac{d}{2}$ for any i and $x+\Delta x$, where $\|\Delta x\| \leq \frac{d}{2L_h}$. This implies $\min_{y\in\mathcal{Y}(x)}\{\max_{i\in\{1,\ldots,k\}} h_i(x+\Delta x, y)\}\} \leq -\frac{d}{2}$ for any point in the ball $B_x(\frac{d}{2L_h})$. According to Theorem 4.1, we know $h_i(x+\Delta x, y_t^*(x+\Delta x))$ has a local upper bound denoted as $-m^{loc}$ in the ball $B_x(\frac{d}{2L_h})$. m^{loc} can be computed by replacing d with $\frac{d}{2}$ in Theorem 4.1. This completes the first part of the proof.

Before proceeding with the proof, we provide the following estimate

$$\left\|\nabla^2_{yx}\widetilde{g}_t(x,y)\right\| \leq \left\|\nabla^2_{yx}g(x,y)\right\| + t\sum_{i=1}^k\left(\frac{\left\|\nabla^2_{yx}h_i(x,y)\right\|}{-h_i(x,y)} + \frac{\|\nabla_yh_i(x,y)\|\,\|\nabla_xh_i(x,y)\|}{h_i^2(x,y)}\right)$$

$$\leq \overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2} \tag{F.5}$$

similar as the proof in Proposition 4.2.

By direct computation, for any x_1, x_2 in the ball $B_x(\frac{d}{2L_h})$, we have

$$\left\|\nabla_x\widetilde{\phi}_t(x_1) - \nabla_x\widetilde{\phi}_t(x_2)\right\|$$

$$= \left\|\nabla_xf(x_1,y_t^*(x_1)) - \nabla^2_{xy}\widetilde{g}_t(x_1,y_t^*(x_1))(\nabla^2_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\nabla_yf(x_1,y_t^*(x_1))\right.$$
$$\left. -\nabla_xf(x_2,y_t^*(x_2)) + \nabla^2_{xy}\widetilde{g}_t(x_2,y_t^*(x_2))(\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1}\nabla_yf(x_2,y_t^*(x_2))\right\|$$

$$\leq \left\|(\nabla_xf(x_1,y_t^*(x_1)) - \nabla_xf(x_2,y_t^*(x_2)))\right\|$$
$$+ \left\|\nabla^2_{xy}\widetilde{g}_t(x_2,y_t^*(x_2))(\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1}(\nabla_yf(x_2,y_t^*(x_2)) - \nabla_yf(x_1,y_t^*(x_1)))\right\|$$
$$+ \left\|\nabla^2_{xy}\widetilde{g}_t(x_2,y_t^*(x_2))((\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1} - (\nabla^2_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1})\nabla_yf(x_1,y_t^*(x_1))\right\|$$
$$+ \left\|(\nabla^2_{xy}\widetilde{g}_t(x_2,y_t^*(x_2)) - \nabla^2_{xy}\widetilde{g}_t(x_1,y_t^*(x_1)))(\nabla^2_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\nabla_yf(x_1,y_t^*(x_1))\right\|$$

$$\overset{(i)}{\leq} \left(\overline{L}_f + \overline{L}_f\frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\left(\frac{1}{\mu_{\widetilde{g}_t}}\right)^2\overline{\overline{L}}_{\widetilde{g}_t,m^{loc}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\frac{1}{\mu_{\widetilde{g}_t}}\overline{\overline{L}}_{\widetilde{g}_t,m^{loc}}\right)$$
$$\times \|(x_1,y_t^*(x_1)) - (x_2,y_t^*(x_2))\|. \tag{F.6}$$

In (i) we estimate $\|(\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1} - (\nabla^2_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\|$ using the following inequality

$$\|A^{-1} - B^{-1}\| = \|A^{-1}(A - B)B^{-1}\| \leq \|A^{-1}\| \cdot \|B^{-1}\| \cdot \|A - B\|,$$

and $\|(\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1}\| \leq 1/\mu_{\widetilde{g}_t}$ from Proposition 4.1.

Now we need to bound $\|(x_1,y_t^*(x_1)) - (x_2,y_t^*(x_2))\|$. Note that $\|y_t^*(x_1) - y_t^*(x_2)\| \leq \|\nabla_xy_t^*(x)\|_{loc}\|x_1 - x_2\|$, where $\|\nabla_xy_t^*(x)\|_{loc}$ is the upper bound of norm of the Jacobian in the ball $B_x(\frac{d}{2L_h})$. We need to evaluate $\|\nabla_xy_t^*(x)\|_{loc}$. By optimal condition of $\widetilde{g}_t(x,y)$, we know $\nabla_y\widetilde{g}_t(x,y_t^*(x))=0$. This implies

$$0 = \nabla_x(\nabla_y\widetilde{g}_t(x,y_t^*(x))) = \nabla^2_{xy}\widetilde{g}_t(x,y_t^*(x)) + (\nabla_xy_t^*(x))^\top\nabla^2_{yy}\widetilde{g}_t(x,y_t^*(x)).$$

This following

$$\|\nabla_x y_t^*(x)\| = \|\nabla_{xy}^2 \widetilde{g}_t(x,y)(\nabla_{yy}^2 \widetilde{g}_t(x,y))^{-1}\|$$
$$= \|\nabla_{xy}^2 \widetilde{g}_t(x,y)\|\|\nabla_{yy}^2 \widetilde{g}_t(x,y)\|^{-1}$$
$$\overset{(i)}{\leq} \frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right),$$

where in (i) we use $\|(\nabla_{yy}^2 \widetilde{g}_t(x_2, y_t^*(x_2)))^{-1}\| \leq 1/\mu_{\widetilde{g}_t}$ and

$$\left\|\nabla_{yx}^2 \widetilde{g}_t(x,y)\right\| \leq \overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}$$

Therefore, we obtain

$$\|y_t^*(x_1) - y_t^*(x_2)\| \leq \frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right)\|x_1 - x_2\|.$$

To conclude, we obtain

$$\|\nabla_x \widetilde{\phi}_t(x_1) - \nabla_x \widetilde{\phi}_t(x_2)\|$$
$$\leq \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\left(\frac{1}{\mu_{\widetilde{g}_t}}\right)^2 \overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\frac{1}{\mu_{\widetilde{g}_t}}\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}\right)$$
$$\times \|(x_1, y_t^*(x_1)) - (x_2, y_t^*(x_2))\|$$
$$\leq \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\left(\frac{1}{\mu_{\widetilde{g}_t}}\right)^2 \overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right) + L_f\frac{1}{\mu_{\widetilde{g}_t}}\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}}\right)$$
$$\times \left(1 + \frac{1}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2}\right)\right)\|x_1 - x_2\|.$$

The proof is completed.

F.6   Proof of Lemma 4.3

Now we apply Lemma 4.2 to the reference point x_s. Replacing d in Lemma 4.2 with d_s, it follows that for any x in the ball $B_{x_s}(d_s/(2L_h))$, we have $h_i(x, y_t^*(x)) \leq -m^{loc}$. As stated in the first paragraph of proof in Lemma 4.2, the constant $-m^{loc}$ can be computed by replacing d in Theorem 4.1 with $d_s/2$. Here we use the notation $m(d_s/2) := m^{loc}$ to emphasize the dependency of m on d_s. Since the feasibility check must terminate when d is between $\frac{D}{2}$ to D, d_s is greater than $D/2$ for any s. From Theorem 4.1, we know that $m(d)$ is monotonically increasing with respect to d. It follows that $m^{loc} = m(d_s/2) \geq m(D/4) =: M^*$. Note that

$$\overline{\overline{L}}_{\widetilde{g}_t, m} = \overline{\overline{L}}_g + tk\left(\frac{\overline{\overline{L}}_h}{m} + \frac{\overline{L}_h L_h}{m^2} + \frac{2L_h^3}{m^3} + \frac{2L_h\overline{L}_h}{m^2}\right),$$

which is monotonically decreasing with respect to m, we have $\overline{\overline{L}}_{\widetilde{g}_t, m^{loc}} \leq \overline{\overline{L}}_{\widetilde{g}_t, M^*}$. Therefore, the following holds

$$\overline{L}_{\widetilde{\phi}_t}^{loc} = \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,m^{loc}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,m^{loc}} \right)$$

$$\times \left(1 + \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m^{loc}} + \frac{tkL_h^2}{(m^{loc})^2} \right) \right)$$

$$\leq \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M^*} + \frac{tkL_h^2}{(M^*)^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,M^*} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M^*} + \frac{tkL_h^2}{(M^*)^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,M^*} \right)$$

$$\times \left(1 + \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M^*} + \frac{tkL_h^2}{(M^*)^2} \right) \right)$$

$$=: \overline{L}_{\widetilde{\phi}_t}.$$

By the design of m in Theorem 4.1, we have $M^* = m(D/4) = \mathcal{O}(t)$. Note that

$$\overline{\overline{L}}_{\widetilde{g}_t,M^*} = \overline{\overline{L}}_g + tk \left(\frac{\overline{\overline{L}}_h}{M^*} + \frac{\overline{L}_h L_h}{(M^*)^2} + \frac{2L_h^3}{(M^*)^3} + \frac{2L_h \overline{L}_h}{(M^*)^2} \right) = \mathcal{O}(1/t^2), \tag{F.7}$$

we conclude $\overline{L}_{\widetilde{\phi}_t} = \mathcal{O}(1/t^4)$.

F.7 Proof of Lemma 4.4

According to Theorem 4.2, we have $\|y_t^*(x_s) - \hat{y}_s\| \leq \epsilon$. Note that $h_i(x, y_t^*(x)) \leq -m_s$ and $h_i(x, \hat{y}_s) \leq -m_s$ for any i. By direct computation

$$\left\| \nabla_x \widetilde{\phi}_t(x_s) - \hat{\nabla}_x \widetilde{\phi}_t(x_s) \right\|$$

$$= \| \nabla_x f(x_s, y_t^*(x_s)) - \nabla_{xy}^2 \widetilde{g}_t(x_s, y_t^*(x_s))(\nabla_{yy}^2 \widetilde{g}_t(x_s, y_t^*(x_s)))^{-1} \nabla_y f(x_s, y_t^*(x_s))$$

$$- \nabla_x f(x_s, \hat{y}_s) + \nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1} \nabla_y f(x_s, \hat{y}_s) \|$$

$$\leq \| (\nabla_x f(x_s, y_t^*(x)) - \nabla_x f(x_s, \hat{y}_s)) \|$$

$$+ \left\| \nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1}(\nabla_y f(x_s, \hat{y}_s) - \nabla_y f(x_s, y_t^*(x_s))) \right\|$$

$$+ \left\| \nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)((\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1} - (\nabla_{yy}^2 \widetilde{g}_t(x_s, y_t^*(x_s)))^{-1}) \nabla_y f(x_s, y_t^*(x_s)) \right\|$$

$$+ \left\| (\nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s) - \nabla_{xy}^2 \widetilde{g}_t(x_s, y_t^*(x_s)))(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1} \nabla_y f(x_s, y_t^*(x_s)) \right\|$$

$$\leq \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,m_s} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,m_s} \right)$$

$$\times \| (x, y_t^*(x_s)) - (x, \hat{y}_s) \|$$

$$\leq \left(\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,m_s} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,m_s} \right) \epsilon_s$$

$$= \overline{L}'_{\widetilde{\phi}_t,m_s} \epsilon_s. \tag{F.8}$$

F.8 Proof of Theorem 4.3

Before proceeding, we establish the following result

Lemma F.1. *The coefficient* $\overline{L}'_{\widetilde{\phi}_t,m_s}$ *defined in Lemma 4.4 has an upper bound*

$$\overline{L}'_{\widetilde{\phi}_t} = \overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M} + \frac{tkL_h^2}{M^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,M} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M} + \frac{tkL_h^2}{M^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,M}.$$

Here M is from Corollary 4.1, and $\overline{\overline{L}}_{\widetilde{g}_t,M}$ can be computed by replacing m from Lemma 4.1 with M.

Proof. From Corollary 4.1, we have $m_s \geq M$. In view of Lemma 4.4, $\overline{\overline{L}}_{\widetilde{g}_t,m}$ is monotonically increases with respect to m, so $\overline{\overline{L}}_{\widetilde{g}_t,m_s} \leq \overline{\overline{L}}_{\widetilde{g}_t,M}$. Therefore, we obtain

$$
\begin{aligned}
\overline{L}'_{\widetilde{\phi}_t,m_s} &= \overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,m_s} \left(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,m_s} \\
&\leq \overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M} + \frac{tkL_h^2}{M^2} \right) + L_f \left(\frac{1}{\mu_{\widetilde{g}_t}} \right)^2 \overline{\overline{L}}_{\widetilde{g}_t,M} \left(\overline{L}_g + \frac{tk\overline{L}_h}{M} + \frac{tkL_h^2}{M^2} \right) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,M} \\
&= \overline{L}'_{\widetilde{\phi}_t}.
\end{aligned}
$$

This complete the proof. $\qquad\square$

Next, we provide the proof of Theorem 4.3. For any point $x_s \in \mathcal{X}$, the lower-level algorithm will guarantee (by Theorem 4.2):

$$
\|\hat{y}_s - y_t^*(x_s)\| \leq \epsilon_s \tag{F.9}
$$

within $\mathcal{O}(\kappa \log(1/\epsilon_s))$ numbers of lower-level gradient oracles, where $\hat{y}_s$ is the output of the lower-level algorithm. According to Lemma 4.4, $1/\epsilon_s = 4\overline{L}'_{\widetilde{\phi}_t,m_s}/\epsilon \leq 4\overline{L}'_{\widetilde{\phi}_t}/\epsilon$, the iteration numbers of lower-level solver is at most $\mathcal{O}(\kappa \log(4\overline{L}'_{\widetilde{\phi}_t}/\epsilon))$.

To proceed with the proof, we first give an upper bound of $\|\hat{\nabla}_x \widetilde{\phi}_t(x_s)\|$. Note that

$$
\hat{\nabla}_x \widetilde{\phi}_t(x_s) = \nabla_x f(x_s, \hat{y}_s) - \nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1} \nabla_y f(x_s, \hat{y}_s),
$$

and we have (by Assumption 4.4 and Proposition 4.1)

$$
\begin{aligned}
&\|\nabla_x f(x_s, \hat{y}_s)\| \leq L_f, \ \|\nabla_y f(x_s, \hat{y}_s)\| \leq L_f, \ \|\nabla_y h(x_s, \hat{y}_s)\| \leq L_h; \\
&\|\nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)\| \leq \overline{L}_g, \ \|\nabla_{xy}^2 \widetilde{h}_t(x_s, \hat{y}_s)\| \leq \overline{L}_h, \ \|(\nabla_{yy}^2 \widetilde{g}_t(x_s, \hat{y}_s))^{-1}\| \leq 1/\mu_{\widetilde{g}_t}.
\end{aligned}
$$

Further, note that

$$
\nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s) = \nabla_{xy}^2 g(x_s, \hat{y}_s) + t \sum_{i=1}^k \left(\frac{\nabla_{xy}^2 h_i(x_s, \hat{y}_s)}{-h_i(x_s, \hat{y}_s)} + \frac{\nabla_x h_i(x_s, \hat{y}_s) \nabla_y h_i(x_s, \hat{y}_s)^\top}{h_i^2(x_s, \hat{y}_s)} \right).
$$

In view of Corollary 4.1 we know $h_i(x_s, \hat{y}_s) \leq -M$, thus

$$
\|\nabla_{xy}^2 \widetilde{g}_t(x_s, \hat{y}_s)\| \leq \overline{L}_g + tk\frac{\overline{L}_h}{M} + tk\frac{L_h^2}{M^2}.
$$

Combining the previous estimates, we reach to the conclusion that

$$
\|\hat{\nabla}_x \widetilde{\phi}_t(x_s)\| \leq L_f + \frac{L_f}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + tk\frac{\overline{L}_h}{M} + tk\frac{L_h^2}{M^2} \right).
$$

From Lemma 4.2, we know that $\widetilde{\phi}_t$ is locally Lipschitz smooth with parameter $\overline{L}_{\widetilde{\phi}_t,s}$ for any x such that $\|x - x_s\| \leq d_s/(2L_h)$. Since we take $\eta_s = \min\{d_s/(2L_h) \cdot 1/\|\hat{\nabla}_x \widetilde{\phi}_t(x_s)\|, 1/\overline{L}_{\widetilde{\phi}_t,s}\}$ so that $\|x_{s+1} - x_s\| \leq d_s/(2L_h)$, therefore we know from Lipschitzness that

$$
\widetilde{\phi}_t(x_{s+1}) - \widetilde{\phi}_t(x_s) \leq \nabla \widetilde{\phi}_t(x_s)^\top (x_{s+1} - x_s) + \frac{\overline{L}_{\widetilde{\phi}_t,s}}{2} \|x_{s+1} - x_s\|^2
$$

73

where $\overline{L}_{\widetilde{\phi}_t,s}$ is specified by Lemma 4.2, i.e.

$$\overline{L}_{\widetilde{\phi}_t,s} = (\overline{L}_f + \overline{L}_f \frac{1}{\mu_{\widetilde{g}_t}}(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2}) + L_f(\frac{1}{\mu_{\widetilde{g}_t}})^2 \overline{\overline{L}}_{\widetilde{g}_t,m_s}(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2}) + L_f \frac{1}{\mu_{\widetilde{g}_t}} \overline{\overline{L}}_{\widetilde{g}_t,m_s})$$
$$\times (1 + \frac{1}{\mu_{\widetilde{g}_t}}(\overline{L}_g + \frac{tk\overline{L}_h}{m_s} + \frac{tkL_h^2}{m_s^2}))$$

where m_s and d_s are outputs of Algorithm 2, also $\overline{L}_{\widetilde{\phi}_t,s}$ is upper bounded by $\overline{L}_{\widetilde{\phi}_t}$ (see Lemma 4.3).

Therefore we have

$$\widetilde{\phi}_t(x_{s+1}) - \widetilde{\phi}_t(x_s) \leq \nabla_x \widetilde{\phi}_t(x_s)^\top (x_{s+1} - x_s) + \frac{\overline{L}_{\widetilde{\phi}_t,s}}{2} \|x_{s+1} - x_s\|^2$$
$$\overset{(i)}{\leq} (\frac{\overline{L}_{\widetilde{\phi}_t,s}}{2} - \frac{1}{\eta_s}) \|x_{s+1} - x_s\|^2 + (\nabla \widetilde{\phi}_t(x_s) - \hat{\nabla}\widetilde{\phi}_t(x_s))^\top (x_{s+1} - x_s)$$
$$\leq (\frac{\overline{L}_{\widetilde{\phi}_t,s}}{2} - \frac{1}{\eta_s}) \|x_{s+1} - x_s\|^2 + \eta_s \|\nabla\widetilde{\phi}_t(x_s) - \hat{\nabla}\widetilde{\phi}_t(x_s)\|^2 + \frac{1}{4\eta_s} \|x_{s+1} - x_s\|^2$$
$$\overset{(ii)}{\leq} -\frac{1}{4\eta_s} \|x_{s+1} - x_s\|^2 + \eta_s \left(\frac{\epsilon}{4}\right)^2$$

(F.10)

where (i) is by the property of projection, namely

$$(x_s - \eta_s \hat{\nabla}_x \widetilde{\phi}_t(x_s) - x_{s+1})^\top (x - x_{s+1}) \leq 0, \ \forall x \in \mathcal{X}$$

and taking $x = x_s$ gives

$$\hat{\nabla}_x \widetilde{\phi}_t(x_s)^\top (x_{s+1} - x_s) \leq -\frac{1}{\eta_s} \|x_s - x_{s+1}\|^2.$$

In (ii) we utilize the following result from Lemma 4.4

$$\|\nabla_x \widetilde{\phi}_t(x_s) - \hat{\nabla}\widetilde{\phi}_t(x_s)\| \leq \epsilon_s \overline{L}'_{\widetilde{\phi}_t,m_s} \leq \frac{\epsilon}{4}.$$

Now suppose

$$\frac{1}{\eta_s^2} \|x_{s+1} - x_s\|^2 \leq \frac{\epsilon^2}{4},$$

then the proof is already finished. Otherwise we have $-\frac{1}{4\eta_s} \|x_{s+1} - x_s\|^2 + \eta_s \left(\frac{\epsilon}{4}\right)^2 \leq 0$, and by (F.10) we know that the algorithm is a descent algorithm. Re-arranging the terms we get:

$$\frac{1}{\eta_s^2} \|x_{s+1} - x_s\|^2 \leq \frac{4(\widetilde{\phi}_t(x_s) - \widetilde{\phi}_t(x_{s+1}))}{\eta_s} + \frac{\epsilon^2}{4}.$$

It remains to lower bound η_s so that the right-hand side of the above inequality is upper bounded. Note that we have defined

$$\eta_s = \min\{1, \frac{d_s}{2L_h} \frac{1}{\|\hat{\nabla}_x \widetilde{\phi}_t(x_s)\|}, \frac{1}{\overline{L}_{\widetilde{\phi}_t,s}}\}$$

where

$$\frac{d_s}{2L_h} \frac{1}{\|\hat{\nabla}_x \widetilde{\phi}_t(x_s)\|} \geq \frac{D}{4L_h} \frac{1}{L_f + \frac{L_f}{\mu_{\widetilde{g}_t}} \left(\overline{L}_g + tk\frac{\overline{L}_h}{M} + tk\frac{L_h^2}{M^2}\right)}$$

(D is a constant given in Remark 4.3, and $d_s \geq D/2$ from Corollary 4.1) and

$$\frac{1}{\overline{L}_{\widetilde{\phi}_t,s}} \geq \frac{1}{\overline{L}_{\widetilde{\phi}_t}}$$

where $\overline{L}_{\widetilde{\phi}_t}$ is given in Lemma 4.3. Thus we have

$$\frac{1}{\eta_s^2}\|x_{s+1} - x_s\|^2 \leq \frac{4(\widetilde{\phi}_t(x_s) - \widetilde{\phi}_t(x_{s+1}))}{\zeta} + \frac{\epsilon^2}{4}.$$

where

$$\zeta = \min\left\{1, \frac{D}{4L_h}\frac{1}{L_f + \frac{L_f}{\mu_{\widetilde{g}_t}}\left(\overline{L}_g + tk\frac{\overline{L}_h}{M} + tk\frac{L_h^2}{M^2}\right)}, \frac{1}{\overline{L}_{\widetilde{\phi}_t}}\right\}. \tag{F.11}$$

By setting $\max_{x\in\mathcal{X}} \widetilde{\phi}_t(x) = \widetilde{\Phi} < +\infty$ and $S = \frac{64\widetilde{\Phi}}{5\zeta\epsilon^2} = \mathcal{O}(\frac{1}{\zeta\epsilon^2})$, we derive

$$\min_{s=0,\dots,S-1}\frac{1}{\eta_s}\|x_s - x_{s+1}\| \leq \frac{3\epsilon}{4}$$

by telescoping sum for $s = 0,\dots,S$. Then

$$\min_{s=0,\dots,S-1}\frac{1}{\eta_s}\|x_s - \mathsf{proj}_{\mathcal{X}}(x_s - \eta_s\nabla_x\widetilde{\phi}_t(x_s))\|$$

$$\leq \min_{s=0,\dots,S-1}\frac{1}{\eta_s}\Big(\|x_s - \mathsf{proj}_{\mathcal{X}}(x_s - \eta_s\nabla_x\widetilde{\phi}_t(x_s))\|$$

$$+ \|\mathsf{proj}_{\mathcal{X}}(x_s - \eta_s\nabla_x\widetilde{\phi}_t(x_s)) - \mathsf{proj}_{\mathcal{X}}(x_s - \eta_s\nabla_x\widetilde{\phi}_t(x_s))\|\Big)$$

$$\leq \frac{3\epsilon}{4} + \max_{s=0,\dots,S-1}\frac{1}{\eta_s}\|\mathsf{proj}_{\mathcal{X}}(\eta_s\hat{\nabla}_x\widetilde{\phi}_t(x_s) - \eta_s\nabla_x\widetilde{\phi}_t(x_s))\|$$

$$\leq \frac{3\epsilon}{4} + \max_{s=0,\dots,S-1}\|\hat{\nabla}_x\widetilde{\phi}_t(x_s) - \nabla_x\widetilde{\phi}_t(x_s)\|$$

$$\leq \frac{3\epsilon}{4} + \frac{\epsilon}{4}$$

$$= \epsilon.$$

Next, we prove the number of iterations for the inner loop is $\widetilde{\mathcal{O}}(t^{-0.5})$. Recall that, from the beginning of the proof, the number of iterations for the inner loop is at most $\mathcal{O}(\kappa\log(4\overline{L}'_{\widetilde{\phi}_t}/\epsilon))$ from Theorem 4.2, where $\kappa = \sqrt{\overline{L}_{\widetilde{g}_t,m_s}/\mu_{\widetilde{g}_t}}$. Since $\overline{L}_{\widetilde{g}_t,m_s}$ can be computed by replacing m in Proposition 4.2 with m_s, and $m_s = \mathcal{O}(t)$ in view of Theorem 4.1, we have $\overline{L}_{\widetilde{g}_t,m_s} = \mathcal{O}(t^{-1})$, implying $\kappa = \mathcal{O}(t^{-0.5})$. According to Lemma 4.4, $\overline{L}'_{\widetilde{\phi}_t}$ is polynomial in t. We conclude that $\mathcal{O}(\kappa\log(4\overline{L}'_{\widetilde{\phi}_t}/\epsilon)) = \widetilde{\mathcal{O}}(t^{-0.5})$.

F.9 Explaination for Remark 4.6

This is because, by the beginning of the proof for Theorem 4.3, the number of iterations for the inner loop is at most $\mathcal{O}(\kappa\log(4\overline{L}'_{\widetilde{\phi}_t}/\epsilon))$, where $\kappa = (\overline{L}_{\widetilde{g}_t,m_s}/\mu_{\widetilde{g}_t})^{1/2}$. Since $\overline{L}_{\widetilde{g}_t,m_s}$ can be computed by replacing m in Proposition 4.2 with m_s, and $m_s = \mathcal{O}(t)$ by Theorem 4.1, we have $\overline{L}_{\widetilde{g}_t,m_s} = \mathcal{O}(t^{-1})$, implying $\kappa = \mathcal{O}(t^{-0.5})$. According to Lemma F.1, $\overline{L}'_{\widetilde{\phi}_t}$ is polynomial in t. We conclude that $\mathcal{O}(\kappa\log(4\overline{L}'_{\widetilde{\phi}_t}/\epsilon)) = \widetilde{\mathcal{O}}(t^{-0.5})$.

F.10 Proof of Theorem 4.4

By Theorem 4.3 we know that for each given t_i

$$\frac{1}{\eta_{t_i}}\|x_i - \text{proj}_{\mathcal{X}}(x_i - \eta_{t_i}\nabla_x\widetilde{\phi}_{t_i}(x_i))\| \leq \epsilon_i \tag{F.12}$$

where η_{t_i} is the stepsize η_s as specified in Theorem 4.3. Note that $\eta_{t_i} \leq 1$, also since (F.12) is decreasing when η_{t_i} increases, see Bertsekas (1997, Lemma 2.3.1), we conclude that:

$$\|x_i - \text{proj}_{\mathcal{X}}(x_i - \nabla_x\widetilde{\phi}_{t_i}(x_i))\| \leq \epsilon_i$$

Now that $i \to \infty$ and $x_i \to x^*$, since x^* is **SCSC** point, there exists a neighborhood of x^* where all points are **SCSC** points (from Lemma B.3 we know the multiplier is continuous and from Definition 3.1 we know **SCSC** means the multiplier is positive, which implies an open set). Therefore we could without loss of generality say that x_i are also **SCSC** points. Consequently Theorem 3.3 and 3.4 imply

$$\nabla_x\widetilde{\phi}_{t_i}(x_i) \to \nabla_x\phi(x^*).$$

Now if we take $i \to \infty$, we know $x_i \to x^*$, $\epsilon_i \to 0$. By

$$\|x_i - \text{proj}_{\mathcal{X}}(x_i - \nabla_x\widetilde{\phi}_{t_i}(x_i))\| \leq \epsilon_i$$

we conclude that

$$x^* = \text{proj}_{\mathcal{X}}(x^* - \nabla_x\phi(x^*))$$

which means that x^* is the stationary point of ϕ.

F.11 Proof of Lemma C.1

By direct computation, we derive

$$\nabla_x y_t^*(x) = - \left(\nabla_{yy}^2 g(x, y_t^*(x)) + \sum_{i=1}^{k}\frac{1}{t}\frac{t^2}{h_i^2(x, y_t^*(x))}\nabla_y h_i(x, y_t^*(x))\left(\nabla_y h_i(x, y_t^*(x))\right)^\top + \sum_{i=1}^{k}\frac{t\nabla_{yy}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))} \right)^{-1}$$

$$\times \left(\nabla_{yx}^2 g(x, y_t^*(x)) + \sum_{i=1}^{k}\frac{1}{t}\frac{t^2}{h_i^2(x, y_t^*(x))}\nabla_y h_i(x, y_t^*(x))\left(\nabla_x h_i(x, y_t^*(x))\right)^\top + \sum_{i=1}^{k}\frac{t\nabla_{yx}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))} \right).$$

Denote

$$A_t(x, y) = \nabla_{yy}^2 g(x, y) + \sum_{i=1}^{k}\frac{t\nabla_{yy}^2 h_i(x, y)}{-h_i(x, y)}, \ y \in \text{int}\mathcal{Y}(x) \tag{F.13}$$

$$B_t(x, y) = \nabla_{yx}^2 g(x, y) + \sum_{i=1}^{k}\frac{t\nabla_{yx}^2 h_i(x, y)}{-h_i(x, y)}, \ y \in \text{int}\mathcal{Y}(x) \tag{F.14}$$

$$v_t^i(x, y) = \frac{t}{-h_i(x, y)}\nabla_y h_i(x, y), \ y \in \text{int}\mathcal{Y}(x) \tag{F.15}$$

$$u_t^i(x, y) = \frac{t}{-h_i(x, y)}\nabla_x h_i(x, y), \ y \in \text{int}\mathcal{Y}(x), \tag{F.16}$$

where $\text{int}\mathcal{Y}(x)$ is the interior of $\mathcal{Y}(x)$. These terms are well-defined only when $y \in \text{int}\mathcal{Y}(x)$ because $h_i(x, y)$ may be 0 for some index i when $y \in \mathcal{Y}(x) \setminus \text{int}\mathcal{Y}(x)$. Note that $y_t^*(x) \in \text{int}\mathcal{Y}(x)$, then

$$\nabla_x y_t^*(x) = \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k}\frac{1}{t}v_t^i(x, y_t^*(x))\left(v_t^i(x, y_t^*(x))\right)^\top \right)^{-1}$$

76

$$\times \left(B_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(u_t^i(x, y_t^*(x)) \right)^\top \right). \tag{F.17}$$

We first show that

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} B_t(x, y_t^*(x)) \right\|$$

is uniformly bounded. By Corollary 4.1, $h_i(x, y_t^*(x)) \leq -M$, where $M = \min\{M_1 t, M_2\}$ for some constant M_1 and M_2. This implies

$$t/(-h_i(x, y_t^*(x))) \leq \max\{1/M_1, 1/M_2\} \tag{F.18}$$

for any i since $0 < t \leq 1$. By Assumption 4.4(4)(7), we have $\|\nabla_{yx}^2 g(x, y_t^*(x))\| \leq \overline{L}_g$ and $\|\nabla_{yx}^2 h_i(x, y_t^*(x))\| \leq \overline{L}_h$. Thus

$$\|B_t(x, y_t^*(x))\| = \left\| \nabla_{yx}^2 g(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{t \nabla_{yx}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))} \right\| \leq \overline{L}_g + k\overline{L}_h \max\{1/M_1, 1/M_2\}.$$

Note that

$$\left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \preceq (\nabla_{yy}^2 g(x, y_t^*(x)))^{-1} \preceq \frac{1}{\mu_g} I, \tag{F.19}$$

we conclude that

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} B_t(x, y_t^*(x)) \right\| \leq \frac{1}{\mu_g} \left(\overline{L}_g + k\overline{L}_h \max\{1/M_1, 1/M_2\} \right).$$

To prove the uniform boundedness of $\|\nabla_x y_t^*(x)\|$, it is sufficient to show that

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(u_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right\| \tag{F.20}$$

is uniformly bounded for any fixed $\hat{j}$. We will prove this on the interval $(0, \widetilde{T}]$ and then on $[\widetilde{T}, 1]$ for some constant $\widetilde{T}$. The value of $\widetilde{T}$ will be determined at the end of the proof. We prove the uniform boundedness on $(0, \widetilde{T}]$ by two steps: denoting

$$\mathcal{X}_1 = \{x \in \mathcal{X} : h_{\hat{j}}(x, y^*(x)) \text{ active}\}, \tag{F.21}$$

we will first show that $\nabla_x y_t^*(x)$ is uniformly bounded on an open neighborhood of $\mathcal{X}_1$, and then demonstrate that $\nabla_x y_t^*(x)$ is also bounded outside this open neighborhood.

We declare that in this proof, all mentions of openness refer to openness in the sense of the subspace topology.

Step 1: We first show the uniform boundedness of (F.20) on an open neighborhood of $\mathcal{X}_1$ as follows: we will prove that for any x in $\mathcal{X}_1$, there exists an open neighborhood W_x of x such that (F.20) is bounded above in this open neighborhood W_x. These open neighborhoods form an open cover of the compact set $\mathcal{X}_1$, so we can select a finite subcover $\{W_i\}_{i=1}^{p}$. We will show the uniform boundedness of (F.20) on $\bigcup_{i=1}^{p} W_i$.

Now we provide the detailed proof. Set $-h_{\hat{j}}(x, y_t^*(x)) = \alpha$, then

$$\left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top\right)^{-1} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(u_t^{\hat{j}}(x, y_t^*(x))\right)^\top$$

$$\overset{(i)}{=} \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top\right)^{-1} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(\frac{t}{-h_i(x, y_t^*(x))} \nabla_x h_{\hat{j}}(x, y_t^*(x))\right)^\top$$

$$\overset{(ii)}{=} \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top\right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x)) \left(\nabla_x h_{\hat{j}}(x, y_t^*(x))\right)^\top.$$

(F.22)

In (i) we plug in $u_t^{\hat{j}}(x, y_t^*(x) = \frac{t}{-h_i(x, y_t^*(x))} \nabla_x h_{\hat{j}}(x, y_t^*(x))$, and in (ii) we utilize $-h_{\hat{j}}(x, y_t^*(x)) = \alpha$. Note that $\|\nabla_x h_{\hat{j}}(x, y_t^*(x))\|$ is uniformly bounded above from Assumption 4.4(6), then it is sufficient to prove that the following term has uniform bound

$$\left\|\left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top\right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x))\right\|.$$

Denote

$$C_t(x, y_t^*(x)) = A_t(x, y_t^*(x)) + \sum_{i \neq \hat{j}, i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top.$$

It is clear that $C_t(x, y_t^*(x)) \succeq \mu_g I$ is positive definite. We have

$$\left\|\left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x))\right)^\top\right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x))\right\|$$

$$= \left\|\left(C_t(x, y_t^*(x)) + \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top\right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x))\right\|$$

$$\overset{(i)}{=} \left\|\alpha^{-1}\left(C_t^{-1}(x, y_t^*(x)) v_t^{\hat{j}}(x, y_t^*(x)) - \frac{C_t^{-1}(x, y_t^*(x)) v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x)) v_t^{\hat{j}}(x, y_t^*(x))}{t + \left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x)) v_t^{\hat{j}}(x, y_t^*(x))}\right)\right\|$$

$$= \left\|\alpha^{-1} \frac{t C_t(x, y_t^*(x))^{-1} v_t^{\hat{j}}(x, y_t^*(x))}{t + \left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x)) v_t^{\hat{j}}(x, y_t^*(x))}\right\|$$

$$\overset{(ii)}{=} \left\|t^2 \alpha^{-2} \frac{C_t^{-1}(x, y_t^*(x)) \nabla_y h_{\hat{j}}(x, y_t^*(x))}{t + t^2 \alpha^{-2} \left(\nabla_y h_{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x)) \nabla_y h_{\hat{j}}(x, y_t^*(x))}\right\|$$

$$\leq \frac{\|C_t^{-1}(x, y_t^*(x)) \nabla_y h_{\hat{j}}(x, y_t^*(x))\|}{\left(\nabla_y h_{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x)) \nabla_y h_{\hat{j}}(x, y_t^*(x))},$$

(F.23)

where (i) is from the Sherman-Morrison formula, and in (ii) we plug in

$$v_t^{\hat{j}}(x, y_t^*(x)) = \frac{t}{-h_{\hat{j}}(x, y_t^*(x))} \nabla_y h_{\hat{j}}(x, y_t^*(x)) = t \alpha^{-1} \nabla_y h_{\hat{j}}(x, y_t^*(x))$$

78

Next we need to prove that $\left(\nabla_y h_{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x))\nabla_y h_{\hat{j}}(x, y_t^*(x))$ has a positive lower bound. Denote

$$\mathcal{I}^*(x) = \{i : h_i(x, y^*(x)) \text{ active}\} \tag{F.24}$$

$$V(x, y; z) = \text{span}\{\nabla_y h_i(x, y) : i \in \mathcal{I}^*(z), \ i \neq \hat{j}\}. \tag{F.25}$$

For further analysis, we decompose

$$C_t^{-1}(x, y_t^*(x))\nabla_y h_{\hat{j}}(x, y_t^*(x)) =: w_t(x) + w_t(x)^\perp,$$

where $w_t(x) \in V(x, y_t^*(x); x_0)$ and $w_t(x)^\perp \in (V(x, y_t^*(x); x_0))^\perp$ for some point x_0. Then we have

$$\left(\nabla_y h_{\hat{j}}(x, y_t^*(x))\right)^\top C_t^{-1}(x, y_t^*(x))\nabla_y h_{\hat{j}}(x, y_t^*(x))$$

$$= \left\langle \nabla_y h_{\hat{j}}(x, y_t^*(x)), w_t(x) + w_t(x)^\perp \right\rangle$$

$$= \left\langle \left(A_t(x, y_t^*(x)) + \sum_{i \neq \hat{j}, i=1}^k \frac{1}{t} v_t^{\hat{j}}\left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top\right)\left(w_t(x) + w_t(x)^\perp\right), w_t(x) + w_t(x)^\perp \right\rangle. \tag{F.26}$$

According to (F.13), we have

$$A_t(x, y_t^*(x)) + \sum_{i \neq \hat{j}, i=1}^k \frac{1}{t} v_t^{\hat{j}}\left(v_t^{\hat{j}}(x, y_t^*(x))\right)^\top \succeq \nabla_{yy}^2 g(x, y_t^*(x)) \succeq \mu_g I. \tag{F.27}$$

We aim to estimate the lower bound of $\|w_t(x) + w_t(x)^\perp\|$. It is difficult to obtain a global bound for $\|w_t(x) + w_t(x)^\perp\|$ in $\mathcal{X}_1$. This because that the active set $\mathcal{I}^*(x)$ changes as x_0 varies, and if we fix the point x_0, it's hard to analyze $C_t^{-1}(x, y_t^*(x))$. So we want to show that for any $x_0 \in \mathcal{X}_1$, there exists an open neighborhood W_{x_0} of x_0 such that $\|w_t(x) + w_t(x)^\perp\|$ is lower bounded on W_{x_0}.

To construct this open neighborhood, we embed $\mathcal{X}_1$ into $\{(x, y) : x \in \mathcal{X}, \ y \in \mathcal{Y}(x)\}$ by the following embedding map

$$\iota : \mathcal{X}_1 \to \{(x, y) : x \in \mathcal{X}, \ y \in \mathcal{Y}(x)\} \tag{F.28}$$

$$x \mapsto (x, y^*(x)).$$

Note that $\hat{j} \in \mathcal{I}^*(x)$ if $x \in \mathcal{X}_1$, by the LICQ assumption, we know the following holds for any $x \in \mathcal{X}_1$

$$\|\text{proj}_{V^\perp(x, y^*(x); x)}\nabla_y h_{\hat{j}}(x, y^*(x))\| > 0.$$

It is clear that $\|\text{proj}_{V^\perp(x, y^*(x); x)}\nabla_y h_{\hat{j}}(x, y^*(x))\|$ is lower semicontinuous in x. Therefore, by compactness of $\mathcal{X}_1$, there exists a constant $\delta > 0$ such that the following holds for any $x \in \mathcal{X}_1$

$$\|\text{proj}_{V^\perp(x, y^*(x); x)}\nabla_y h_{\hat{j}}(x, y^*(x))\| \geq \delta.$$

For any x_0 in $\mathcal{X}_1$, we will show that we can find a neighborhood U_{x_0} of $(x_0, y^*(x_0))$ in $\{(x, y) : x \in \mathcal{X}, \ y \in \mathcal{Y}(x)\}$ such that

1. For any $i \notin \mathcal{I}^*(x_0)$ and $(x, y) \in U_{x_0}$, we have $h_i(x, y) \leq -H_{x_0}$ for some constant $H_{x_0} > 0$;

2. The vectors $\{\nabla_y h_i(x, y) : i \in \mathcal{I}^*(x_0)\}$ are linear independent;

79

3. There exists a constant $\delta_{x_0} > 0$ such that $\|\text{proj}_{V^\perp(x,y;x_0)} \nabla_y h_{\hat{j}}(x,y)\| \geq \delta_{x_0}$ holds for any $(x,y) \in U_{x_0}$;

4. There exists a constant r_{x_0} such that for any $x \in \pi_x(U_{x_0})$, which is the projection of U_{x_0} onto the x component, we have $\{x\} \times \left(B_{y^*(x)}(r_{x_0}) \cap \mathcal{Y}(x) \right) \subset U_{x_0}$, where $B_{y^*(x)}(r_{x_0})$ is a ball in $\mathcal{Y}(x)$ centered at $y^*(x)$ with radius r_{x_0}.
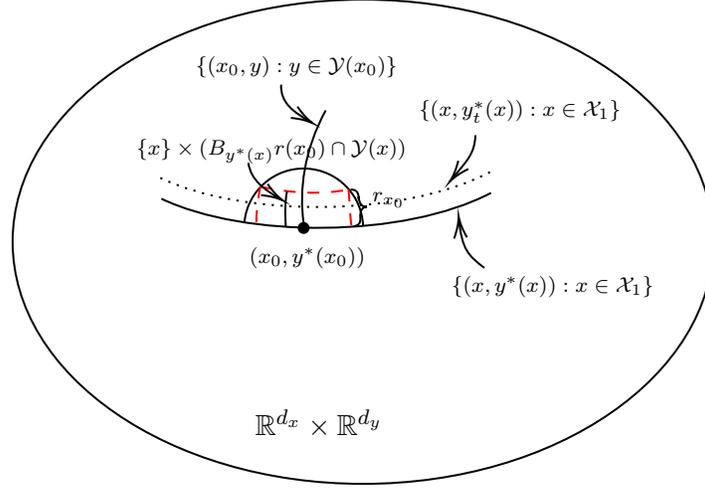
Figure 3: Find a spherical neighborhood that satisfies the first three properties, as shown by the solid hemisphere in the figure, then find a tubular neighborhood within it, as shown by the red dashed tube in the figure.

The first property can be directly obtained from the continuity of the constraint functions $h_i(x,y)$ and the fact that $h_i(x, y^*(x)) \neq 0$ for any $i \notin \mathcal{I}^*(x_0)$. The second property holds because $\{\nabla_y h_i(x_0, y^*(x_0)) : i \in \mathcal{I}^*(x_0)\}$ are linear independent by LICQ assumption. This property ensures that $V(x,y;x_0)$ and $V^\perp(x,y;x_0)$ are non-degenerate on U_{x_0}, so $\text{proj}_{V^\perp(x,y;x_0)} \nabla_y h_{\hat{j}}(x,y)$ is continuous with respect to (x,y). Note that

$$\|\text{proj}_{V^\perp(x_0,y^*(x_0);x_0)} \nabla_y h_{\hat{j}}(x_0, y^*(x_0))\| \geq \delta,$$

by continuity, the third property can also be satisfied. To show the fourth, note that we find a spherical neighborhood that satisfies the first three conditions. Therefore, within this spherical neighborhood, we can identify a tubular neighborhood, thus also satisfying the fourth property easily.

Denote $W_{x_0} = \pi_x(U_{x_0})$. The projection map is an open map, so W_{x_0} is open. we will prove that $\|w_t(x) + w_t(x)^\perp\|$ is lower bounded on W_{x_0}. We consider the case that $t \leq (r_{x_0}^2 \mu_g)/(2k)$. By Lemma B.1, we have $\|y^*(x) - y_t^*(x)\| \leq \sqrt{2kt/\mu_g}$, which means $\|y^*(x) - y_t^*(x)\| \leq r_{x_0}$ and thus $(x, y_t^*(x)) \in U_{x_0}$ by the fourth property of the open neighborhood. Then we obtain

$$\nabla_y h_{\hat{j}}(x, y_t^*(x))$$

$$= \left(A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right.$$

$$\left. + \sum_{i \in \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right) (w_t(x) + w_t(x)^\perp)$$

80

$$\overset{(i)}{=} \left(A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right) (w_t(x) + w_t(x)^\perp)$$

$$+ \sum_{i \in \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top w_t(x),$$

where (i) is because

$$\sum_{i \in \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top w_t(x)^\perp = 0.$$

On the other hand, we have

$$\nabla_y h_{\hat{j}}(x, y_t^*(x)) = \mathsf{proj}_{V(x, y_t^*(x); x_0)} \nabla_y h_{\hat{j}}(x, y_t^*(x)) + \mathsf{proj}_{V^\perp(x, y_t^*(x); x_0)} \nabla_y h_{\hat{j}}(x, y_t^*(x)).$$

Note that $\sum_{i \in \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top w_t(x) \in V(x, y_t^*(x); x_0)$, we have

$$\mathsf{proj}_{V^\perp(x, y_t^*(x); x_0)} \nabla_y h_{\hat{j}}(x, y_t^*(x))$$

$$= \mathsf{proj}_{V^\perp(x, y_t^*(x); x_0)} \left[\left(A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right) (w_t(x) + w_t(x)^\perp) \right],$$

which means

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right) (w_t(x) + w_t(x)^\perp) \right\|$$

$$\geq \left\| \mathsf{proj}_{V^\perp(x, y_t^*(x); x_0)} \left[\left(A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}} \right)^\top \right) (w_t(x) + w_t(x)^\perp) \right] \right\|$$

$$= \left\| \mathsf{proj}_{V^\perp(x, y_t^*(x); x_0)} \nabla_y h_{\hat{j}}(x, y_t^*(x)) \right\|$$

$$\overset{(i)}{\geq} \delta_{x_0}. \tag{F.29}$$

where in (i) we use the third property of the open neighborhood. By (F.18), $t/(-h_i(x, y_t^*(x)))$ is uniformly bounded. Combining Assumption 4.4(4)(7), it is not hard to see that the maximum eigenvalue of $A_t(x, y_t^*(x)) = \nabla_{yy}^2 g(x, y_t^*(x)) + \sum_{i=1}^k \frac{t \nabla_{yy}^2 h_i(x, y_t^*(x))}{-h_i(x, y_t^*(x))}$ has upper bounded denoted as λ_{max}. According to the first property that for any $i \notin \mathcal{I}(x_0)$ and $(x, y) \in U_{x_0}$, $h_i(x, y) \leq -H_{x_0}$ holds for some constant $H_{x_0} > 0$, we obtain that

$$\left\| A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(v_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right\|$$

$$= \left\| A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{1}{t} \frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \left(\frac{t}{-h_i(x, y_t^*(x))} \nabla_y h_i(x, y_t^*(x)) \right)^\top \right\|$$

81

$$= \left\| A_t(x, y_t^*(x)) + \sum_{i \notin \mathcal{I}^*(x_0), i \neq \hat{j}} \frac{t}{(-h_i(x, y_t^*(x)))^2} \nabla_y h_i(x, y_t^*(x)) \left(\nabla_y h_i(x, y_t^*(x)) \right)^\top \right\|$$

$$\leq \lambda_{max} + \frac{kL_h^2}{H_{x_0}^2}.$$

Combing (F.29), we obtain $\|w_t(x) + w_t(x)^\perp\| \geq \delta_{x_0}/(\lambda_{max} + \frac{kL_h^2}{H_{x_0}^2})$. Now we proved the lower boundedness of $\|w_t(x) + w_t(x)^\perp\|$ on W_{x_0}. By (F.26) and (F.27), we obtain the lower boundedness of

$$\left(\nabla_y h_{\hat{j}}(x, y_t^*(x)) \right)^\top C_t^{-1}(x, y_t^*(x)) \nabla_y h_{\hat{j}}(x, y_t^*(x)).$$

According to (F.23), this means the following term is upper bounded on W_{x_0}

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^k \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x)) \right\|. \tag{F.30}$$

Recalling (F.22), this further implies the upper boundedness of the following term on W_{x_0}

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^k \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(u_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right\|. \tag{F.31}$$

Note that $\{W_{x_0}\}_{x_0 \in \mathcal{X}_1}$ forms an open cover of $\mathcal{X}_1$ in $\mathcal{X}$. Since $\mathcal{X}_1$ is compact, we can choose a finite subcover, denoted as $\{W_i\}_{i=1}^p$ where $W_i = W_{x_0}$ for some x_0 in $\mathcal{X}_1$. Denote $T_i = (r_i^2 \mu_g)/(2k)$, $r_i = r_{x_0}$ from the fourth property and G_i be the upper bound of norm of (F.17) when $0 < t \leq T_i$. Take $T^* = \min\{T_i\}$, $G^* = \max\{G_i\}$. It is not hard to see that the norm of (F.17) is upper bounded by G^* for any $0 < t \leq T^*$ and x in $\bigcup_{i=1}^p W_i$.

Step 2: Consider $\mathcal{X}_2 := \mathcal{X} \setminus \bigcup_{i=1}^p W_i$ which is compact in $\mathcal{X}$. Since $\bigcup_{i=1}^p W_i$ has covered $\mathcal{X}_1$, it is easy to see that $h_{\hat{j}}(x, y^*(x))$ has a negative upper bound in $\mathcal{X}_2$, denoted as $-Q$. We assume that $t \leq (\mu_g Q^2)/(4kL_h^2)$. By Lemma B.1, we have $\|y_t^*(x) - y^*(x)\| \leq \sqrt{2kt/\mu_g} \leq Q/(2L_h)$. Thus $|h_{\hat{j}}(x, y_t^*(x)) - h_{\hat{j}}(x, y^*(x))| \leq Q/2$, which means $h_{\hat{j}}(x, y_t^*(x)) \leq -Q/2$. Then the following term is uniformly bounded for any $x \in \mathcal{X}_2$ and $0 < t \leq (\mu_g Q^2)/(4kL_h^2)$:

$$\left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^k \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \frac{1}{t} v_t^{\hat{j}}(x, y_t^*(x)) \left(u_t^{\hat{j}}(x, y_t^*(x)) \right)^\top \right\|$$

$$= \left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^k \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \frac{t}{h_{\hat{j}}^2(x, y_t^*(x))} \nabla_y h_{\hat{j}}(x, y_t^*(x)) \left(\nabla_x h_{\hat{j}}(x, y_t^*(x)) \right)^\top \right\|$$

$$\overset{(i)}{\leq} \frac{1}{\mu_g} \frac{t}{h_{\hat{j}}^2(x, y_t^*(x))} \left\| \nabla_y h_{\hat{j}}(x, y_t^*(x)) \right\| \cdot \left\| \nabla_x h_{\hat{j}}(x, y_t^*(x)) \right\|$$

$$\overset{(ii)}{\leq} \frac{4L_h^2}{\mu_g Q^2}. \tag{F.32}$$

(i) is because (F.19) and $v_t^{\hat{j}}(x, y_t^*(x)) = \frac{t}{-h_{\hat{j}}(x, y_t^*(x))} \nabla_y h_{\hat{j}}(x, y_t^*(x))$. In (ii), we utilize $t \leq 1$, $h_{\hat{j}}(x, y_t^*(x)) \leq -Q/2$ and $\|\nabla_y h_{\hat{j}}(x, y_t^*(x))\| \leq L_h$. Furthermore, we obtain (F.17) has bounded norm.

82

To conclude, we have proved that $\|\nabla_x y_t^*(x)\|$ is upper bounded for any x in $\mathcal{X}$ and $0 < t \leq \widetilde{T} :=$ $\min\{T^*, (\mu_g Q^2)/(4k L_h^2)\}$. If $\widetilde{T}$ is less than 1, see $\|\nabla_x y_t^*(x)\|$ as a continuous function defined of the compact set $\mathcal{X} \times [\widetilde{T}, 1]$, so it has an upper bound. Therefore, we proved that $\|\nabla_x y_t^*(x)\|$ is upper bounded for any $x \in \mathcal{X}$ and $0 < t \leq 1$.

F.12 Proof of Lemma C.2

Let $-h_{\hat{j}}(x, y_t^*(x)) = \alpha$ and use the notations (F.13)-(F.16), we obtain

$$
\left\| (\nabla_{yy}\widetilde{g}_t(x))^{-1} \frac{t\nabla_y h_{\hat{j}}(x, y_t^*(x))}{(-h_{\hat{j}}(x, y_t^*(x)))^2} \right\|
$$
$$
= \left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \alpha^{-1} v_t^{\hat{j}}(x, y_t^*(x)) \right\|. \tag{F.33}
$$

We have proved this term is uniformly bounded on $\bigcup_{i=1}^{p} W_i$, as detailed from (F.30) to the end of Step 1 in Section F.11. For $\mathcal{X} \setminus \bigcup_{i=1}^{p} W_i$, we can assume $h_{\hat{j}}(x, y^*(x)) \leq -Q$. Then

$$
\left\| (\nabla_{yy}\widetilde{g}_t(x))^{-1} \frac{t\nabla_y h_{\hat{j}}(x, y_t^*(x))}{(-h_{\hat{j}}(x, y_t^*(x)))^2} \right\|
$$
$$
= \left\| \left(A_t(x, y_t^*(x)) + \sum_{i=1}^{k} \frac{1}{t} v_t^i(x, y_t^*(x)) \left(v_t^i(x, y_t^*(x)) \right)^\top \right)^{-1} \frac{t}{h_{\hat{j}}^2(x, y_t^*(x))} \nabla_y h_{\hat{j}}(x, y_t^*(x)) \right\|
$$

is bounded by the same process of (F.32).

F.13 Proof of Lemma C.3

Following the argument at the end of Step 1 in proof of Lemma C.1, we select an open neighborhood $\{W_i\}_{i=1}^{p}$ of $\mathcal{X}_1$.

If $x \in \bigcup_{i=1}^{p} W_i$, denote $\beta_1 = \min_i\{r_i\}$. According to the fourth property of the open neighborhood, $\|y - y^*(x)\| \leq r_i$ for any i means y is in the W_i for any i. By replicating the proof in Step 1 and simply substituting $y_t^*(x)$ with y, we can obtain $\|(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}\nabla_{yx}\widetilde{g}_t(x, y)\| \leq G^*$ for any $y \in \mathrm{int}\mathcal{Y}(x)$ such that $\|y - y^*(x)\| \leq \beta_1$.

If $x \in \mathcal{X} \setminus \bigcup_{i=1}^{p} W_i$, denote $\beta_2 = Q/(2L_h)$. Similar to Step 2 in proof of Lemma C.1, we can also find that $\|(\nabla_{yy}\widetilde{g}_t(x, y))^{-1}\nabla_{yx}\widetilde{g}_t(x, y)\|$ is upper bounded for any $y \in \mathrm{int}\mathcal{Y}(x)$ such that $\|y - y^*(x)\| \leq \beta_2$. Define $\beta = \min\{\beta_1, \beta_2\}$, we complete the proof.

F.14 Proof of Lemma C.4

The proof is similar to that of Lemma 4.2 and Lemma 4.3. First we aim to prove the local Lipschitz smoothness of $\widetilde{\phi}_t(x)$. Following (F.6), for any x_1, x_2 satisfies $\|x_1 - x\|$, $\|x_2 - x\| \leq \frac{d}{2L_h}$, we need to estimate

$$
\|\nabla_x \widetilde{\phi}_t(x_1) - \nabla_x \widetilde{\phi}_t(x_2)\|
$$
$$
\leq \|(\nabla_x f(x_1, y_t^*(x_1)) - \nabla_x f(x_2, y_t^*(x_2)))\| \tag{F.34}
$$
$$
+ \left\| \nabla_{xy}^2 \widetilde{g}_t(x_2, y_t^*(x_2))(\nabla_{yy}^2 \widetilde{g}_t(x_2, y_t^*(x_2)))^{-1}(\nabla_y f(x_2, y_t^*(x_2)) - \nabla_y f(x_1, y_t^*(x_1))) \right\| \tag{F.35}
$$
$$
+ \left\| \nabla_{xy}^2 \widetilde{g}_t(x_2, y_t^*(x_2))((\nabla_{yy}^2 \widetilde{g}_t(x_2, y_t^*(x_2)))^{-1} - (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1})\nabla_y f(x_1, y_t^*(x_1)) \right\| \tag{F.36}
$$

83

$$+ \left\| (\nabla^2_{xy} \widetilde{g}_t(x_2, y^*_t(x_2)) - \nabla^2_{xy} \widetilde{g}_t(x_1, y^*_t(x_1)))(\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} \nabla_y f(x_1, y^*_t(x_1)) \right\|. \qquad \text{(F.37)}$$

For (F.34), by Assumption 4.4(2), we have

$$\|(\nabla_x f(x_1, y^*_t(x_1)) - \nabla_x f(x_2, y^*_t(x_2)))\| \leq \overline{L}_f \|(x_1, y^*_t(x_1)) - (x_2, y^*_t(x_2))\|.$$

We evaluate (F.35), (F.36) and (F.37) separately.

The term (F.35): According to Lemma C.1, $\|\nabla^2_{xy} \widetilde{g}_t(x, y^*_t(x))(\nabla^2_{yy} \widetilde{g}_t(x, y^*_t(x)))^{-1}\| \leq J_1$. Combining the Lipschitz smoothness of $f(x, y)$ from Assumption4.4(2). Therefore

$$\left\| \nabla^2_{xy} \widetilde{g}_t(x_2, y^*_t(x_2))(\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)))^{-1} (\nabla_y f(x_2, y^*_t(x_2)) - \nabla_y f(x_1, y^*_t(x_1))) \right\|$$
$$\leq J_1 \overline{L}_f \|(x_1, y^*_t(x_1)) - (x_2, y^*_t(x_2))\|.$$

The term (F.36): By Assumption 4.4(1), $\|\nabla_y f(x, y)\| \leq L_f$. To estimate

$$\left\| \nabla^2_{xy} \widetilde{g}_t(x_2, y^*_t(x_2))((\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)))^{-1} - (\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1}) \nabla_y f(x_1, y^*_t(x_1)) \right\|,$$

It is sufficient to evaluate the following term

$$\left\| \nabla^2_{xy} \widetilde{g}_t(x_2, y^*_t(x_2))((\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)))^{-1} - (\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1}) \right\|$$
$$\stackrel{(i)}{=} \left\| \nabla^2_{xy} \widetilde{g}_t(x_2, y^*_t(x_2))(\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)))^{-1} (\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)) - \nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))(\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} \right\|$$
$$\stackrel{(ii)}{\leq} J_1 \|(\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} (\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)) - \nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))\|, \qquad \text{(F.38)}$$

where (i) is due to $A^{-1} - B^{-1} = -A^{-1}(A - B)B^{-1}$, and (ii) is from Lemma C.1, which says $\|\nabla^2_{xy} \widetilde{g}_t(x, y^*_t(x))(\nabla^2_{yy} \widetilde{g}_t(x, y^*_t(x)))^{-1}\| \leq J_1$. To estimate (F.38), we directly compute as following

$$\|(\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} (\nabla^2_{yy} \widetilde{g}_t(x_2, y^*_t(x_2)) - \nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))\|$$
$$= \left\| (\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} \left[\nabla^2_{yy} g(x_1, y^*_t(x_1)) - \nabla^2_{yy} g(x_2, y^*_t(x_2)) \right. \right. \qquad \text{(F.39)}$$
$$+ t \sum_{i=1}^{k} \frac{h_i(x_1, y^*_t(x_1)) \nabla^2_{yy} h_i(x_2, y^*_t(x_2)) - h_i(x_2, y^*_t(x_2)) \nabla^2_{yy} h_i(x_1, y^*_t(x_1))}{h_i(x_1, y^*_t(x_1)) h_i(x_2, y^*_t(x_2))} \qquad \text{(F.40)}$$
$$\left. \left. + t \sum_{i=1}^{k} \frac{h_i^2(x_2, y^*_t(x_2)) \nabla_y h_i(x_1, y^*_t(x_1)) \nabla_y h_i(x_1, y^*_t(x_1))^\top - h_i^2(x_1, y^*_t(x_1)) \nabla_y h_i(x_2, y^*_t(x_2)) \nabla_y h_i(x_2, y^*_t(x_2))^\top}{h_i^2(x_1, y^*_t(x_1)) h_i^2(x_2, y^*_t(x_2))} \right] \right\|.$$
$$\text{(F.41)}$$

We respectively evaluate (F.39), (F.40), and (F.41).

For (F.39) and (F.40), we use Lipschitz continuity of $\nabla^2_{yy} g(x, y)$ from Assumption 4.4(5), the result that $h_i(x_2, y^*_t(x_2)) \leq -m^{loc}$ from Lemma 4.2, and the same process in (F.3). Note that $\nabla^2_{yy} \widetilde{g}(x, y) \succeq \mu_g I$, we get

$$\left\| (\nabla^2_{yy} \widetilde{g}_t(x_1, y^*_t(x_1)))^{-1} \left[\nabla^2_{yy} g(x_1, y^*_t(x_1)) - \nabla^2_{yy} g(x_2, y^*_t(x_2)) \right. \right.$$
$$\left. \left. + t \sum_{i=1}^{k} \frac{h_i(x_1, y^*_t(x_1)) \nabla^2_{yy} h_i(x_2, y^*_t(x_2)) - h_i(x_2, y^*_t(x_2)) \nabla^2_{yy} h_i(x_1, y^*_t(x_1))}{h_i(x_1, y^*_t(x_1)) h_i(x_2, y^*_t(x_2))} \right] \right\|$$
$$\leq \frac{1}{\mu_g} \left\| \nabla^2_{yy} g(x_1, y^*_t(x_1)) - \nabla^2_{yy} g(x_2, y^*_t(x_2)) \right\|$$

$$+ \frac{1}{\mu_g} \left\| t \sum_{i=1}^{k} \frac{h_i(x_1, y_t^*(x_1)) \nabla_{yy}^2 h_i(x_2, y_t^*(x_2)) - h_i(x_2, y_t^*(x_2)) \nabla_{yy}^2 h_i(x_1, y_t^*(x_1))}{h_i(x_1, y_t^*(x_1)) h_i(x_2, y_t^*(x_2))} \right\|$$

$$\leq \frac{1}{\mu_g} \left(\overline{\overline{L}}_g + tk \left(\frac{\overline{\overline{L}}_h}{m^{loc}} + \frac{\overline{L}_h L_h}{(m^{loc})^2} \right) \right) \| (x_1, y_t^*(x_1)) - (x_2, y_t^*(x_2)) \|$$

For (F.41), Using the same method of adding and subtracting like terms as in (F.4), and applying the triangle inequality, we obtain

$$\left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} t \sum_{i=1}^{k} \left(\frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right.$$
$$\left. \left. - \frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\|$$

$$\leq t \sum_{i=1}^{k} \left[\left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} \left(\frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right. \right.$$
$$\left. \left. - \frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\|$$
$$+ \left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} \left(\frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right.$$
$$\left. \left. - \frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\|$$
$$+ \left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} \left(\frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right.$$
$$\left. \left. \left. - \frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\| \right]$$

$$\overset{(i)}{\leq} \frac{tk L_h \overline{L}_h}{\mu_g (m^{loc})^2} \| (x_1, y_t^*(x_1)) - (x_2, y_t^*(x_2)) \|.$$
$$+ t \sum_{i=1}^{k} \left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} \left(\frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right.$$
$$\left. \left. - \frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\|$$
$$+ \frac{tk L_h \overline{L}_h}{\mu_g (m^{loc})^2} \| (x_1, y_t^*(x_1)) - (x_2, y_t^*(x_2)) \|.$$

$$= t \sum_{i=1}^{k} \left\| (\nabla_{yy}^2 \widetilde{g}_t(x_1, y_t^*(x_1)))^{-1} \left(\frac{h_i^2(x_2, y_t^*(x_2)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right. \right.$$
$$\left. \left. - \frac{h_i^2(x_1, y_t^*(x_1)) \nabla_y h_i(x_1, y_t^*(x_1)) \nabla_y h_i(x_2, y_t^*(x_2))^\top}{h_i^2(x_1, y_t^*(x_1)) h_i^2(x_2, y_t^*(x_2))} \right) \right\| \tag{F.42}$$
$$+ 2 \frac{tk L_h \overline{L}_h}{\mu_g (m^{loc})^2} \| (x_1, y_t^*(x_1)) - (x_2, y_t^*(x_2)) \|.$$

In (i) we used the result that $h_i(x_1, y_t^*(x_1)),\ h_i(x_2, y_t^*(x_2)) \leq -m^{loc}$ from Lemma 4.2, Lips-

chitz smoothness of $h_i(x,y)$ from Assumption 4.4(7), boundedness of $\nabla_y h_i(x,y)$ from 4.4(6), and $\nabla^2_{yy}\widetilde{g}_t(x,y_t^*(x)) \succeq \mu_g I$. Finally, we need to evaluate (F.42), i.e.

$$t\sum_{i=1}^k \left\|(\nabla_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\left(\frac{(-h_i(x_2,y_t^*(x_2)))^2\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{(h_i(x_1,y_t^*(x_1)h_i(x_2,y_t^*(x_2)))^2}\right.\right.$$
$$\left.\left.-\frac{(-h_i(x_1,y_t^*(x_1)))^2\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{(h_i(x_1,y_t^*(x_1)h_i(x_2,y_t^*(x_2)))^2}\right)\right\|$$
$$=t\sum_{i=1}^k \left\|(\nabla_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\left(\frac{(h_i(x_1,y_t^*(x_1))-h_i(x_2,y_t^*(x_2)))\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{h_i^2(x_1,y_t^*(x_1)h_i(x_2,y_t^*(x_2)))}\right.\right.$$
$$\left.\left.+\frac{(h_i(x_1,y_t^*(x_1))-h_i(x_2,y_t^*(x_2)))\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{h_i(x_1,y_t^*(x_1)h_i^2(x_2,y_t^*(x_2)))}\right)\right\|. \tag{F.43}$$

From Lemma C.3, we have

$$t\sum_{i=1}^k \left\|(\nabla_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}\left(\frac{(h_i(x_1,y_t^*(x_1))-h_i(x_2,y_t^*(x_2)))\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{h_i^2(x_1,y_t^*(x_1)h_i(x_2,y_t^*(x_2)))}\right)\right\|$$
$$\leq J_2\frac{kL_h^2}{m^{loc}}\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|.$$

For the second term in (F.43), we observe that

$$\sum_{i=1}^k \left\|(\nabla_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}t\left(\frac{(h_i(x_1,y_t^*(x_1))-h_i(x_2,y_t^*(x_2)))\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{h_i(x_1,y_t^*(x_1)h_i^2(x_2,y_t^*(x_2)))}\right)\right\|$$
$$=\sum_{i=1}^k \left\|(\nabla_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1}t\left(\frac{(h_i(x_1,y_t^*(x_1))-h_i(x_2,y_t^*(x_2)))\nabla_y h_i(x_1,y_t^*(x_1))(\nabla_y h_i(x_2,y_t^*(x_2)))^\top}{h_i^2(x_1,y_t^*(x_1)h_i(x_2,y_t^*(x_2)))}\right.\right.$$
$$\left.\left.\times\frac{h_i(x_1,y_t^*(x_1))}{h_i(x_2,y_t^*(x_2))}\right)\right\|$$
$$\leq J_2\frac{kL_h^2}{m^{loc}}\left|\frac{-h_i(x_1,y_t^*(x_1))}{-h_i(x_2,y_t^*(x_2))}\right|\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|$$
$$\overset{(i)}{\leq} J_2\frac{kL_h^2}{m^{loc}}\left|\frac{-h_i(x,y_t^*(x))+\frac{m}{2}}{-h_i(x,y_t^*(x))-\frac{m}{2}}\right|\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|$$
$$= J_2\frac{kL_h^2}{m^{loc}}\left|1+\frac{m}{-h_i(x,y_t^*(x))-\frac{m}{2}}\right|\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|$$
$$\leq J_2\frac{kL_h^2}{m^{loc}}\left(1+\left|\frac{m}{m-\frac{m}{2}}\right|\right)\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|$$
$$\leq J_2\frac{3kL_h^2}{m^{loc}}\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|,$$

where (i) is because $h_i(x_1,y_t^*(x_1))$, $h_i(x_2,y_t^*(x_2))$, $h_i(x,y_t^*(x)) \leq 0$, $h_i(x,y_t^*(x))$ is $L_h(1+J_1)$ Lipschitz continuous, and $\|x-x_1\|$, $\|x-x_2\| \leq m/(2L_h(1+J_1))$. To summarize, we get

$$\left\|\nabla^2_{xy}\widetilde{g}_t(x_2,y_t^*(x_2))((\nabla^2_{yy}\widetilde{g}_t(x_2,y_t^*(x_2)))^{-1}-(\nabla^2_{yy}\widetilde{g}_t(x_1,y_t^*(x_1)))^{-1})\nabla_y f(x_1,y_t^*(x_1))\right\|$$
$$\leq L_f J_1\left(\frac{\overline{\overline{L}}_g}{\mu_g}+\frac{tk\overline{\overline{L}}_h}{\mu_g m^{loc}}+\frac{tk\overline{L_h}L_h}{\mu_g (m^{loc})^2}+2\frac{tkL_h\overline{L}_h}{\mu_g (m^{loc})^2}+J_2\frac{kL_h^2}{m^{loc}}+J_2\frac{3kL_h^2}{m^{loc}}\right)\|(x_1,y_t^*(x_1))-(x_2,y_t^*(x_2))\|.$$

The term (F.37): This term can be evaluated directly by Lemma 4.1 and Lemma C.1 as follows:

$$\left\|(\nabla^2_{xy}\widetilde{g}_t(x_2,y^*_t(x_2)) - \nabla^2_{xy}\widetilde{g}_t(x_1,y^*_t(x_1)))(\nabla^2_{yy}\widetilde{g}_t(x_1,y^*_t(x_1)))^{-1}\nabla_y f(x_1,y^*_t(x_1))\right\|$$
$$\leq \overline{\overline{L}}_{\widetilde{g}_t,m^{loc}}J_1\|(x_1,y^*_t(x_1)) - (x_2,y^*_t(x_2))\|.$$

Note that $\|(x_1,y^*_t(x_1)) - (x_2,y^*_t(x_2))\| \leq \|\nabla_x y^*_t(x)\|\|x_1 - x_2\| \leq J_1\|x_1 - x_2\|$ from Lemma C.1. Setting

$$\overline{L}^{loc}_{\widetilde{\phi}_t} = J_1\left(\overline{L}_f + \overline{L}_f J_1 + L_f J_1\left(\frac{\overline{\overline{L}}_g}{\mu_g} + \frac{tk\overline{\overline{L}}_h}{\mu_g m^{loc}} + \frac{tk\overline{L_h}L_h}{\mu_g(m^{loc})^2}\right.\right.$$
$$\left.\left. + 2\frac{tkL_h\overline{L}_h}{\mu_g(m^{loc})^2} + J_2\frac{kL^2_h}{m^{loc}} + J_2\frac{3kL^2_h}{m^{loc}}\right) + \overline{\overline{L}}_{\widetilde{g}_t,m^{loc}}J_1\right),$$

we conclude

$$\|\nabla_x\widetilde{\phi}_t(x_1) - \nabla_x\widetilde{\phi}_t(x_2)\| \leq \overline{L}^{loc}_{\widetilde{\phi}_t}\|x_1 - x_2\|.$$

Following the proof of Lemma 4.3 around (F.7), m^{loc} has a lower bound $M^* = m(D/4) = \mathcal{O}(t)$. By design of $\overline{\overline{L}}_{\widetilde{g}_t,m^{loc}}$ in Lemma 4.4, we know that $\overline{\overline{L}}_{\widetilde{g}_t,m^{loc}} \leq \overline{\overline{L}}_{\widetilde{g}_t,M^*} = \mathcal{O}(1/t)$, so $\overline{L}^{loc}_{\widetilde{\phi}_t}$ has an upper bound

$$\overline{L}_{\widetilde{\phi}_t} = J_1\left(\overline{L}_f + \overline{L}_f J_1 + L_f J_1\left(\frac{\overline{\overline{L}}_g}{\mu_g} + \frac{tk\overline{\overline{L}}_h}{\mu_g M^*} + \frac{tk\overline{L_h}L_h}{\mu_g(M^*)^2}\right.\right.$$
$$\left.\left. + 2\frac{tkL_h\overline{L}_h}{\mu_g(M^*)^2} + J_2\frac{kL^2_h}{M^*} + J_2\frac{3kL^2_h}{M^*}\right) + \overline{\overline{L}}_{\widetilde{g}_t,M^*}J_1\right)$$
$$= \mathcal{O}(\frac{1}{t}).$$

F.15 Proof of Theorem C.1

The proof is almost identical to the proof of Theorem 4.3. The only noteworthy point is that $\hat{\nabla}_x\widetilde{\phi}_t(x_s)$ now is upper bounded. This can be proved as follows: due to Lemma B.1, $\|y^*(x_s) - y^*_t(x_s)\| \leq \beta/2$. The lower-level algorithm will guarantee (by Theorem 4.2) $\|\hat{y}_s - y^*_t(x_s)\| \leq \beta/2$, which means $\|y^*(x_s) - \hat{y}_s\| \leq \beta$. By Lemma C.3, $\|(\nabla_{yy}\widetilde{g}_t(x_s,\hat{y}_s))^{-1}\nabla_{yx}\widetilde{g}_t(x_s,\hat{y}_s)\| \leq J_3$, so

$$\|\hat{\nabla}_x\widetilde{\phi}_t(x_s)\| = \|\nabla_x f(x_s,\tilde{y}_s) - \nabla^2_{xy}\widetilde{g}(x_s,\tilde{y}_s)(\nabla^2_{yy}\widetilde{g}(x_s,\tilde{y}_s))^{-1}\nabla_y f(x_s,\tilde{y}_s)\|$$
$$\leq \|\nabla_x f(x_s,\tilde{y}_s)\| + \|\nabla^2_{xy}\widetilde{g}(x_s,\tilde{y}_s)(\nabla^2_{yy}\widetilde{g}(x_s,\tilde{y}_s))^{-1}\|\|\nabla_y f(x_s,\tilde{y}_s)\|$$
$$\leq L_f(1 + J_3).$$

The remaining steps to complete the proof are analogous to those in Theorem 4.3.