
Set Smoothness Unlocks Clarke Hyper-stationarity in Bilevel Optimization

He Chen

SEEM

The Chinese University of Hong Kong
Shatin, Hong Kong
hchen@se.cuhk.edu.hk

Jiajin Li

Sauder School of Business
University of British Columbia
Vancouver, BC, Canada
jiajin.li@sauder.ubc.ca

Anthony Man-Cho So

SEEM

The Chinese University of Hong Kong
Shatin, Hong Kong
manchoso@se.cuhk.edu.hk

Abstract

Solving bilevel optimization (BLO) problems to global optimality is generally intractable. A common surrogate is to compute a hyper-stationary point—a stationary point of the hyper-objective function obtained by minimizing or maximizing the upper-level objective over the lower-level solution set. Existing methods, however, either provide weak notions of stationarity or require restrictive assumptions to guarantee the smoothness of hyper-objective functions. In this paper, we eliminate these impractical assumptions and show that strong (Clarke) hyper-stationarity remains computable even when the hyper-objective is nonsmooth. Our key ingredient is a new structural property, called *set smoothness*, which captures the variational dependence of the lower-level solution set on the upper-level variable. We prove that this property holds for a broad class of BLO problems and ensures weak convexity (resp. concavity) of pessimistic (resp. optimistic) hyper-objective functions. Building on this foundation, we show that a zeroth-order algorithm that computes approximate Clarke hyper-stationary points with non-asymptotic convergence guarantees. To the best of our knowledge, this is the first computational guarantee for Clarke-type stationarity in nonsmooth BLO. Beyond this specific application, the set smoothness property emerges as a structural concept of independent interest, with potential to inform the analysis of broader classes of optimization and variational problems.

1 Introduction

Bilevel optimization (BLO) models hierarchical decision-making with two agents acting sequentially [13, 14]. The follower responds to the leader’s decision by solving a lower-level optimization problem, while the leader seeks an optimal strategy to minimize its upper-level objective subject to this reaction. The follower’s attitude plays a central role: If the follower is favorable (resp. adverse) to the leader, the resulting BLO is termed optimistic (resp. pessimistic) [13, 51, 36]. Formally, the optimistic and pessimistic BLO take the following forms:

Optimistic BLO:

$$\begin{array}{ll} \min_{\mathbf{x} \in \mathbb{R}^m} \min_{\mathbf{y} \in \mathbb{R}^n} & F(\mathbf{x}, \mathbf{y}) \\ \text{subject to} & \mathbf{y} \in \arg \min_{\mathbf{y}' \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y}'), \end{array}$$

Pessimistic BLO:

$$\begin{array}{ll} \min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} & F(\mathbf{x}, \mathbf{y}) \\ \text{subject to} & \mathbf{y} \in \arg \min_{\mathbf{y}' \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y}'). \end{array}$$

These formulations appear in diverse domains such as Stackelberg games [14, 4, 50], hyperparameter optimization [17, 3, 6], reinforcement learning [29, 55, 21], and interdiction games [36, 5], among others. A standard approach to tackle such nested problems is to reformulate them into single-level problems via *hyper-objective functions*. Let $\mathcal{S}(\mathbf{x}) := \arg \min_{\mathbf{y}' \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y}')$ denote the follower's optimal response set. The *optimistic* and *pessimistic* hyper-objectives are then defined as

$$\varphi_o(\mathbf{x}) := \min_{\mathbf{y} \in \mathcal{S}(\mathbf{x})} F(\mathbf{x}, \mathbf{y}), \quad \varphi_p(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{S}(\mathbf{x})} F(\mathbf{x}, \mathbf{y}). \quad (1)$$

Solving an optimistic (resp. pessimistic) BLO is therefore equivalent to minimizing the corresponding hyper-objective φ_o (resp. φ_p).

Despite the single-level reformulation, the resulting hyper-objective functions are highly nonconvex [8, 31], which makes global optimization intractable. In practice, researchers therefore focus on finding stationary points rather than global minimizers, using algorithms such as implicit gradient descent [17, 18, 3] and fully first-order methods [32, 21]. These approaches assume that the lower-level problem is strongly convex, ensuring a unique solution, i.e., $\mathcal{S}(\mathbf{x}) := \{\mathbf{y}^*(\mathbf{x})\}$. Under this assumption, the hyper-objective reduces to a smooth function $\varphi(\mathbf{x}) := F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ [19]. One can then seek an ϵ -approximate hyper-stationary point satisfying $\|\nabla \varphi(\mathbf{x})\| \leq \epsilon$. Convergence is well understood under smoothness and uniqueness assumptions [19], but these conditions rarely hold in practice. With multiple lower-level solutions, the existing methods break down.

To move beyond the singleton lower-level solution set, Kwon et al. [32] introduced a penalty-based framework that allows multiple follower solutions. Building on this idea, Chen et al. [8] obtained a refined scheme with near-optimal convergence. However, ensuring smoothness of the induced hyper-objective still demands strong regularity: The penalized model function $h_\sigma(\mathbf{x}, \mathbf{y}) := \sigma F(\mathbf{x}, \mathbf{y}) + f(\mathbf{x}, \mathbf{y})$ must satisfy, uniformly in $\sigma \in [0, \bar{\sigma}]$, an error bound or a Polyak–Łojasiewicz (PL) condition in $\mathbf{y}$. Such requirements are often unrealistic in practice, as F and f typically have mismatched structures. More fundamentally, the Kurdyka–Łojasiewicz (KL) exponent is not preserved under summation [23], so smoothness of hyper-objective functions cannot be guaranteed.

Without relying on these stringent conditions, Chen et al. [7] and Khanduri et al. [28] proposed algorithms for nonsmooth hyper-objectives; however, by their zero-respecting nature (cf. [8, Thm. 3.2]), they cannot in general approximate hyper-stationary points and thus only guarantee convergence to (approximate) Goldstein stationary points [20]—a relatively weak notion. By contrast, a separate line of work studies alternative stationarity concepts via reformulations [35, 52, 38, 1, 39]; yet these notions (e.g., KKT stationarity [37, Sec. 2.1] and penalization stationarity [54, Sec. 4.2]) are posed jointly in $(\mathbf{x}, \mathbf{y})$ and do not ensure that, for a stationary pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, the lower-level solution $\bar{\mathbf{y}}$ actually minimizes or maximizes $F(\bar{\mathbf{x}}, \mathbf{y})$ over $\mathcal{S}(\bar{\mathbf{x}})$.

Given the above discussion, existing algorithms either fail to approximate a meaningful hyper-stationary point or rely on stringent assumptions to do so. This naturally leads to a fundamental question:

Can strong hyper-stationarity be computed in general settings where multiple lower-level solutions exist?

Addressing this question is challenging for a simple reason chain. When the lower level admits multiple solutions, the induced hyper-objective is typically nonsmooth and, under standard assumptions, no better than Lipschitz continuous [7, Corollary 6.1]. At precisely this level of regularity, computing (stronger) approximate Clarke stationary points is, in general, computationally intractable [30, 46]. Thus Lipschitz regularity alone is too weak for algorithmic purposes, motivating new, verifiable structural conditions that make meaningful hyper-stationarity attainable.

Our Contributions. In this paper, we address the above challenges and show that (strong) Clarke stationarity of hyper-objective functions is computable for a broad class of BLO problems. As our key contribution, we identify a hidden weak convexity/concavity structure of the hyper-objective

in nonconvex-PŁ BLO,¹ which places the analysis within the well-studied weakly convex/concave framework. Within this setting, approximate hyper-stationarity admits a natural Clarke-subdifferential characterization that we leverage to obtain computable guarantees.

The foundation of our analysis is a new concept, *set smoothness* (Definition 3), which extends classical smoothness to set-valued mappings and encompasses several variational regularity notions [40, 15, 7, 27]. Building on this notion, we prove two complementary statements. First, if the lower-level solution mapping is set smooth, then the optimistic (resp. pessimistic) hyper-objective is weakly concave (resp. weakly convex). Second, a broad and verifiable condition guarantees set smoothness: When the lower-level function satisfies an error bound condition—equivalently, the PŁ condition—the solution mapping is set smooth. Together, these statements provide checkable criteria under which the hyper-objective inherits a weak convexity/concavity structure.

Once the hidden weak convexity/concavity of the hyper-objective is in place, approximate Clarke hyper-stationary points can be computed by a simple inexact zeroth-order scheme. In the weakly convex case, results based on the Moreau envelope [12, 56, 41] provide convergence and complexity guarantees. For the weakly concave case, however, no existing algorithmic guarantee is known, and the absence of a Moreau-type smoothing technique makes the analysis significantly more challenging. We overcome this by developing a novel convergence proof based on a Brøndsted–Rockafellar-type approximation result [43, Theorem 2], and establish, to the best of our knowledge, the first general computational guarantee for finding approximate Clarke stationary points of nonsmooth hyper-objective functions.

Overall, these developments, particularly set smoothness, provide a principled foundation for the computability of hyper-stationarity in BLO and open new avenues for other structured nonsmooth optimization problems.

Organization. This paper is organized as follows. Sec. 2 collects assumptions and preliminaries. Sec. 3 introduces set smoothness and uses it to reveal a weak convexity/concavity structure of the hyper-objective. Sec. 4 presents an inexact zeroth-order scheme and establishes convergence guarantees for computing approximate Clarke hyper-stationary points. Sec. 5 concludes with final remarks.

Notation. The notation used in this paper is mostly standard. We use $\|\mathbf{x}\|$ to denote the Euclidean norm of a vector $\mathbf{x}$ and $\|\mathbf{A}\|$ to denote the l_2 norm of a matrix $\mathbf{A}$. We use $\mathbb{B}(\mathbf{z}, r)$ to denote the ball centering at $\mathbf{z}$ with radius r , i.e., $\{\mathbf{x} : \|\mathbf{x} - \mathbf{z}\| \leq r\}$. For a scalar $\alpha \in \mathbb{R}$ and a set $\mathcal{S} \subseteq \mathbb{R}^n$, we use $\alpha \cdot \mathcal{S}$ to denote their product $\{\alpha \mathbf{x} : \mathbf{x} \in \mathcal{S}\}$. We define the distance from a vector $\mathbf{x} \in \mathbb{R}^n$ to $\mathcal{S}$ by $\text{dist}(\mathbf{x}, \mathcal{S}) := \min_{\mathbf{z} \in \mathcal{S}} \|\mathbf{x} - \mathbf{z}\|$ and the projection of $\mathbf{x}$ onto $\mathcal{S}$ by $\Pi_{\mathcal{S}}(\mathbf{x}) := \arg \min_{\mathbf{z} \in \mathcal{S}} \|\mathbf{x} - \mathbf{z}\|$. We use $\text{Conv}(\mathcal{S})$ to denote the convex hull of $\mathcal{S}$. For two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^n$, define their Minkowski sum by $\mathcal{S}_1 + \mathcal{S}_2 := \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$, and define their Hausdorff distance (with respect to $\|\cdot\|$) by

$$d_{\text{H}}(\mathcal{S}_1, \mathcal{S}_2) := \max \left\{ \sup_{\mathbf{x}_1 \in \mathcal{S}_1} \text{dist}(\mathbf{x}_1, \mathcal{S}_2), \sup_{\mathbf{x}_2 \in \mathcal{S}_2} \text{dist}(\mathbf{x}_2, \mathcal{S}_1) \right\}.$$

For a differentiable function $g : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$, we use ∇g to denote its gradient w.r.t. the joint variables $(\mathbf{x}, \mathbf{y})$ and $\nabla_{\mathbf{x}} g$ (resp. $\nabla_{\mathbf{y}} g$) to denote its gradient w.r.t. $\mathbf{x}$ (resp. $\mathbf{y}$).

2 Preliminaries

In this paper, we focus on nonconvex-PŁ BLO problems and make the following assumptions:

Assumption 1 (Lower-level Functions).

(A1). The function f is L_f -smooth and twice differentiable. Moreover, $\nabla \nabla_{\mathbf{y}} f$ is H_f -Lipschitz continuous, i.e., for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$,

$$\|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)\| \leq H_f (\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|).$$

(A2). The solution set $\mathcal{S}(\mathbf{x}) = \arg \min_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$ is nonempty closed convex for all $\mathbf{x} \in \mathbb{R}^m$.

¹That is, BLO problems with a nonconvex upper-level objective and a lower-level function satisfying the PŁ condition.

- (A3). *There exists a scalar $\tau > 0$ such that for all $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$,*

$$\text{dist}(\mathbf{y}, \mathcal{S}(\mathbf{x})) \leq \tau \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|.$$

Assumption 2 (Upper-level and Hyper-objective Functions).

- (B1). *The function F is M_F -Lipschitz continuous and L_F -smooth.*
 (B2). *There exists $\mathbf{x}^* \in \mathbb{R}^m$ such that $\varphi_o(\mathbf{x}^*) > -\infty$ (resp. $\varphi_p(\mathbf{x}^*) < +\infty$) for the optimistic (resp. pessimistic) setting.*

Assumptions (A1), (A2), and (B1) are standard in BLO settings; see, e.g., [21, 8, 19, 52, 1] and the references therein. Assumption (B2) guarantees that the hyper-objective functions are well-defined and is imposed without loss of generality. Assumption (A3) imposes an error bound in the lower-level variable that holds uniformly over the upper-level parameter. This requirement is strictly weaker than the strong convexity-in- $\mathbf{y}$ conditions commonly used in prior work [18, 22, 21], as it allows the solution set $\arg \min_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ to be multi-valued. Under L_f -smoothness f in $\mathbf{y}$, (A3) implies the PL inequality

$$f(\mathbf{x}, \mathbf{y}) - \min_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y}) \leq \frac{\tau L_f^2}{2} \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n,$$

vice versus [33, Theorem 3.1]. Hence, our setting aligns with the widely adopted nonconvex-PL framework for BLO [45, 52, 35].

Under the standing assumptions, we begin with the solution mapping $\mathcal{S}$, which under (A3) admits the following equivalent characterization:

$$\mathcal{S}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \mathbf{0}\}. \quad (2)$$

Furthermore, under Assumption 1, the solution mapping $\mathcal{S}$ as well as the hyper-objective functions φ_o and φ_p are Lipschitz continuous.

Lemma 1 (Lipschitz Continuity of $\mathcal{S}(\mathbf{x})$). *(cf. [7, Proposition 6.1]) Under Assumption 1, the lower-level solution set function is M_S -Lipschitz continuous with $M_S = L_f \tau$, i.e., for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$,*

$$d_H(\mathcal{S}(\mathbf{x}_1), \mathcal{S}(\mathbf{x}_2)) \leq M_S \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Lemma 2. *(cf. [7, Proposition 5.3]) Suppose that Assumption 1 and 2 hold. For the optimistic (resp. pessimistic) setting, φ_o (resp. φ_p) is M_φ -Lipschitz continuous with $M_\varphi = M_F(1 + L_f \tau)$.*

The Lipschitz continuity of the hyper-objective functions ensures that the Clarke and Goldstein subdifferentials are well defined.

Definition 1 (Clarke Subdifferential). *(cf. [9, Definition 1.1]) For a Lipschitz continuous function $g : \mathbb{R}^m \rightarrow \mathbb{R}$, the Clarke subdifferential of g at a point $\mathbf{x} \in \mathbb{R}^m$ is defined by*

$$\partial g(\mathbf{x}) := \text{Conv}(\{\mathbf{s} \in \mathbb{R}^m : \exists \mathbf{x}' \rightarrow \mathbf{x}, \nabla g(\mathbf{x}') \text{ exists, } \nabla g(\mathbf{x}') \rightarrow \mathbf{s}\}).$$

We say that $\mathbf{x}$ is an (ϵ, δ) -approximate Clarke stationary point of g if

$$\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \delta)} \partial g(\mathbf{z})\right) \leq \epsilon.$$

Remark 1. *The Clarke subdifferential ∂g reduces to the gradient ∇g when g is smooth. Moreover, if g is convex, then $\partial g(\mathbf{x})$ coincides with the vanilla subgradients defined by $\{\mathbf{s} : g(\mathbf{z}) \geq g(\mathbf{x}) + \mathbf{s}^T(\mathbf{z} - \mathbf{x}) \forall \mathbf{z} \in \mathbb{R}^m\}$.*

Then, the Goldstein δ -subdifferential at $\mathbf{x}$ can be constructed by the convex hull of the Clarke subdifferentials taken over a δ -neighborhood of $\mathbf{x}$. Here is the formal definition of Goldstein δ -subdifferential.

Definition 2 (Goldstein δ -Subdifferential). *(cf. [20, Definition 2.2]) For a Lipschitz continuous function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and a scalar $\delta \geq 0$, the Goldstein δ -subdifferential of g at a point $\mathbf{x} \in \mathbb{R}^m$ is defined by*

$$\partial_\delta g(\mathbf{x}) := \text{Conv}\left(\left\{\bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \delta)} \partial g(\mathbf{z})\right\}\right).$$

We say that $\mathbf{x}$ is an (ϵ, δ) -approximate Goldstein stationary point of g if $\text{dist}(\mathbf{0}, \partial_\delta g(\mathbf{x})) \leq \epsilon$.

Leveraging the Lipschitz continuity of the hyper-objective, recent work has established the computability of (ϵ, δ) -Goldstein hyper-stationary points [7]. However, Goldstein stationarity is strictly weaker and does not, in general, imply Clarke stationarity. Indeed, there exists a convex, 2-Lipschitz function $\tilde{g} : \mathbb{R}^2 \rightarrow \mathbb{R}$ and a point $\mathbf{x}$ such that $\mathbf{x}$ is $(0, \delta)$ -Goldstein stationary while $\min_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \delta)} \text{dist}(\mathbf{0}, \partial \tilde{g}(\mathbf{z})) \geq \frac{2}{\sqrt{5}}$; see [47, Proposition 2.7]. To obtain stronger algorithmic guarantees for hyper-objective minimization, we therefore focus on computing Clarke stationary points (and their (ϵ, δ) -approximate variants).

Despite the well-definedness, approximate Clarke stationarity is not achievable in finite time for general Lipschitz functions [30, 46]. A sufficient condition for its computability is the weak convexity of g [12]. To elaborate on this, we review some basic properties of weakly convex functions. Given a function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and a scalar $r > 0$, we say that g is r -weakly convex if the function $\mathbf{x} \mapsto g(\mathbf{x}) + \frac{r}{2} \|\mathbf{x}\|^2$ is convex. The following equivalent characterizations are useful for our analysis.

Lemma 3 (Equivalent Characterizations of Weak Convexity). *(cf. [11, Theorem 3.1] and [2, Proposition 2.2]) For a Lipschitz continuous function $g : \mathbb{R}^m \rightarrow \mathbb{R}$, the following statements are equivalent:*

- (i) g is r -weakly convex.
- (ii) For any $\theta \in [0, 1]$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, we have

$$g(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2) \leq \theta g(\mathbf{x}_1) + (1 - \theta) g(\mathbf{x}_2) + \frac{r}{2} \theta(1 - \theta) \|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

- (iii) For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$ with $\partial g(\mathbf{x}_1) \neq \emptyset$, and all subgradients $\mathbf{v} \in \partial g(\mathbf{x}_1)$, we have

$$\mathbf{v}^T (\mathbf{x}_2 - \mathbf{x}_1) \leq g(\mathbf{x}_2) - g(\mathbf{x}_1) + \frac{r}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|^2.$$

For an r -weakly convex function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ with $\gamma \in (0, \frac{1}{r})$, we define its Moreau envelope and the proximal mapping by

$$g_\gamma(\mathbf{x}) := \inf_{\mathbf{z} \in \mathbb{R}^n} \left\{ g(\mathbf{z}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{z}\|^2 \right\}, \quad \text{prox}_{\gamma, g}(\mathbf{x}) := \arg \min_{\mathbf{z} \in \mathbb{R}^n} \left\{ g(\mathbf{z}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{z}\|^2 \right\}.$$

Clearly, $\text{prox}_{\gamma, g}(\mathbf{x})$ is single-valued and well-defined, when g is r -weakly convex and $\gamma < \frac{1}{r}$. Next, we provide the standard result, which establishes a stationarity measure based on the gradient of the Moreau envelope.

Lemma 4 (Properties of Moreau Envelope). *(cf. [56, Proposition 2.1]) Suppose that $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a r -weakly convex function and $\gamma < \frac{1}{r}$. The following hold:*

- (i) $g_\gamma(\mathbf{x}) \leq g(\mathbf{x}) - \frac{1-\gamma r}{2\gamma} \|\mathbf{x} - \text{prox}_{\gamma, g}(\mathbf{x})\|^2$.
- (ii) $\gamma \text{dist}(\mathbf{0}, \partial g(\hat{\mathbf{x}})) \leq \|\mathbf{x} - \text{prox}_{\gamma, g}(\mathbf{x})\| \leq \frac{\gamma}{1-\gamma r} \text{dist}(\mathbf{0}, \partial g(\mathbf{x}))$.
- (iii) $\mathbf{x} = \text{prox}_{\gamma, g}(\mathbf{x})$ if and only if $\mathbf{0} \in \partial g(\mathbf{x})$.
- (iv) $\nabla g_\gamma(\mathbf{x}) = \frac{1}{\gamma} (\mathbf{x} - \text{prox}_{\gamma, g}(\mathbf{x}))$.

Lemma 4 (ii) and (iv) show that $\|\nabla g_\gamma(\mathbf{x})\|$ equals zero if and only if $\mathbf{x} = \text{prox}_{\gamma, g}(\mathbf{x})$ and $\mathbf{0} \in \partial g(\mathbf{x})$. Thus $\|\nabla g_\gamma(\mathbf{x})\|$ is a valid Clarke stationarity measure. Moreover, by Lemma 4 (iv), if $\|\nabla g_\gamma(\mathbf{x})\| \leq \epsilon$, then $\|\text{prox}_{\gamma, g}(\mathbf{x}) - \mathbf{x}\| = \gamma \|\nabla g_\gamma(\mathbf{x})\| \leq \gamma \epsilon$, hence $\text{dist}(\mathbf{0}, \partial g(\text{prox}_{\gamma, g}(\mathbf{x}))) \leq \epsilon$. Equivalently,

$$\|\nabla g_\gamma(\mathbf{x})\| \leq \epsilon \implies \text{dist} \left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \gamma \epsilon)} \partial g(\mathbf{z}) \right) \leq \epsilon. \quad (3)$$

Since Davis and Drusvyatskiy [12] establish non-asymptotic rates for finding $\mathbf{x}$ with $\|\nabla g_\gamma(\mathbf{x})\| \leq \epsilon$ when g is weakly convex, (3) implies that $(\epsilon, \gamma \epsilon)$ -approximate Clarke stationarity is computable in

this regime. This observation motivates us to establish weak-convexity-type structure for hyper-objectives; see Sec. 3.

Before leaving this section, we record *weak concavity*, a notion closely related to weak convexity. For a function $g : \mathbb{R}^n \rightarrow \mathbb{R}$, we say that g is r -weakly convex if $-g$ is r -weakly convex. We have the following facts.

Fact 1. *If a function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ is r -weakly concave, then for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, and $\mathbf{v} \in \partial g(\mathbf{x}_1)$, we have $g(\mathbf{x}_2) \leq g(\mathbf{x}_1) + \mathbf{v}^T(\mathbf{x}_2 - \mathbf{x}_1) + \frac{r}{2}\|\mathbf{x}_2 - \mathbf{x}_1\|^2$.*

Fact 2. *If a function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ is r -smooth, then g is r -weakly convex and r -weakly concave, and the following inequality holds:*

$$|\theta g(\mathbf{x}_1) + (1 - \theta)g(\mathbf{x}_2) - g(\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2)| \leq \frac{r}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

3 Unveiling Hidden Structural Properties

This section is devoted to unveiling the hidden structural properties of the hyper-objective functions, which is the key contribution of this paper. Recall that the hyper-objective functions in (1) are defined by minimizing/maximizing the upper-level function w.r.t. $\mathbf{y}$ over the parameterized set $\mathcal{S}(\mathbf{x})$. We are motivated to investigate the property of the set-valued function $\mathcal{S}$. Inspired by the smoothness of real-valued functions, we propose a novel concept of smoothness for set-valued functions, formalized in Definition 3. As we will show, the lower-level solution set function $\mathcal{S}$ satisfies this smoothness property, which in turn ensures the weak concavity (resp. convexity) of φ_o (resp. φ_p).

Definition 3 (Set Smoothness). *For a set-valued function $\mathcal{Y} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ with a convex domain $\text{dom}(\mathcal{Y}) \subseteq \mathbb{R}^m$, we say that it is L -smooth if for any $\mathbf{x}_1, \mathbf{x}_2 \in \text{dom}(\mathcal{Y})$, $\theta \in [0, 1]$, and all $\mathbf{y} \in \mathcal{Y}(\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2)$, there exist $\mathbf{y}_1 \in \mathcal{Y}(\mathbf{x}_1)$ and $\mathbf{y}_2 \in \mathcal{Y}(\mathbf{x}_2)$ such that*

$$\|\theta \mathbf{y}_1 + (1 - \theta)\mathbf{y}_2 - \mathbf{y}\| \leq \frac{L}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2; \quad (4)$$

$$\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \quad (5)$$

The condition (4) can be viewed as a natural extension of the gradient-Lipschitz smoothness condition for real-valued functions to the setting of set-valued mappings. It guarantees that a convex combination of $\mathbf{y}_1 \in \mathcal{Y}(\mathbf{x}_1)$ and $\mathbf{y}_2 \in \mathcal{Y}(\mathbf{x}_2)$ provides a close approximation to a point in $\mathcal{Y}(\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2)$, with an error that decays quadratically in $\|\mathbf{x}_1 - \mathbf{x}_2\|$. This yields the following set inclusion:

$$\mathcal{Y}(\theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2) \subseteq \theta \mathcal{Y}(\mathbf{x}_1) + (1 - \theta)\mathcal{Y}(\mathbf{x}_2) + \frac{L}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \cdot \mathbb{B}(\mathbf{0}, 1). \quad (6)$$

Intuitively, (5) enforces a consistent branch selection between $\mathcal{Y}(\mathbf{x}_1)$ and $\mathcal{Y}(\mathbf{x}_2)$: The chosen representatives $\mathbf{y}_1$ and $\mathbf{y}_2$ must remain aligned (Lipschitz-close) as the input varies, thereby excluding cross-branch pairings that could make the interpolation in (4) hold trivially while the underlying geometry is severely mismatched.

Example 1 (Why the condition (5) is needed: A trivialization for the condition (4)). Define the set-valued map $\mathcal{Y} : \mathbb{R} \rightrightarrows \mathbb{R}^2$ by $\mathcal{Y}(x) = \{(z, x) : z \in \mathbb{R}\}$. Pick $x_1 = a > 0$, $x_2 = -a$, and $\theta = \frac{1}{2}$; then $\theta x_1 + (1 - \theta)x_2 = 0$ and $\mathcal{Y}(0) = \{(z, 0) : z \in \mathbb{R}\}$. Choose $\mathbf{y} = \mathbf{0} \in \mathcal{Y}(0)$, $\mathbf{y}_1 = (K, a) \in \mathcal{Y}(a)$, and $\mathbf{y}_2 = (-K, -a) \in \mathcal{Y}(-a)$. We have

$$\frac{1}{2}\mathbf{y}_1 + \frac{1}{2}\mathbf{y}_2 = \mathbf{y},$$

so the condition (4) holds with zero error even though $\|\mathbf{y}_1 - \mathbf{y}_2\| = 2\sqrt{a^2 + K^2}$ can be made arbitrarily large as $K \rightarrow \infty$. Hence the condition (4) alone does not preclude severely mismatched pairings on a convex domain. In contrast, (5) enforces $\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|^2 = 4La^2$, which forces $K^2 \leq (L - 1)a^2$ and thereby rules out such cross-branch selections unless the Lipschitz modulus is correspondingly large. $\square$

With the notion of set smoothness in place, we now present our first main theoretical result. It shows that set smoothness serves as the key vehicle for establishing weak convexity/concavity of parametric optimization problems with coupled constraints: Under mild Lipschitz-type assumptions, the induced value function inherits weak convexity (or weak concavity). This is formalized in Theorem 1 below.

Theorem 1 (Implication of Set Smoothness). *Consider a real-valued function $g : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ and a set-valued function $\mathcal{Y} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$. Let $\phi(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} g(\mathbf{x}, \mathbf{y})$ and $\mathcal{D} := \{\mathbf{x} : \phi(\mathbf{x}) > -\infty\}$. Assume that $\mathcal{D}$ is a nonempty closed convex set, $\mathcal{Y}$ is $L_{\mathcal{Y}}$ -smooth on $\mathcal{D}$, and g is M_g -Lipschitz continuous w.r.t. $\mathbf{y}$, L_g -smooth on $\mathcal{D} \times \text{Conv}(\bigcup_{\mathbf{x} \in \mathcal{D}} \mathcal{Y}(\mathbf{x}))$. Then, the function ϕ is ρ -weakly convex with $\rho = M_g L_{\mathcal{Y}} + L_g(1 + L_{\mathcal{Y}})$.*

We now instantiate the framework in the bilevel setting. Our first step is to certify *set smoothness* for the lower-level solution map. Under the error-bound (EB) condition in Assumption 1 (A3), the mapping $\mathcal{S} : \mathbf{x} \mapsto \arg \min_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$ is $L_{\mathcal{S}}$ -smooth (Theorem 2). Combining this with Theorem 1 shows that the pessimistic hyper-objective φ_p inherits weak convexity (resp. the optimistic φ_o inherits weak concavity).

Theorem 2 (EB Implies Set Smoothness). *If a function $f : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies Assumption 1, then its associated solution set function $\mathcal{S} : \mathbf{x} \mapsto \arg \min_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y})$ is $L_{\mathcal{S}}$ -smooth with $L_{\mathcal{S}} = \max\{2H_f\tau(1 + 9L_f^2\tau^2), 4L_f^2\tau^2\}$.*

Proof idea (why residual backfilling is essential) Fix $\mathbf{x}_1, \mathbf{x}_2$ and $\theta \in (0, 1)$, and set $\mathbf{x}^\theta := \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$. Given any $\mathbf{y} \in \mathcal{S}(\mathbf{x}^\theta)$, our goal is to select $\mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)$ and $\mathbf{y}_2 \in \mathcal{S}(\mathbf{x}_2)$ so that (4) and (5) hold. A natural choice is to project $\mathbf{y}$ onto the endpoint fibers, yielding $\bar{\mathbf{y}}_i := \Pi_{\mathcal{S}(\mathbf{x}_i)}(\mathbf{y})$ for $i = 1, 2$. Using Lemma 1, it is easy to see that this naive selection satisfies (5).

However, even when each fiber $\mathcal{S}(\mathbf{x})$ is convex, the midpoint $\bar{\mathbf{y}}^\theta := \theta\bar{\mathbf{y}}_1 + (1 - \theta)\bar{\mathbf{y}}_2$ may correspond, at $\mathbf{x}^\theta$, to a *different local selection* of the set-valued map $\mathcal{S}(\cdot)$ than the given $\mathbf{y} \in \mathcal{S}(\mathbf{x}^\theta)$. Consequently, in general multi-solution settings the naive midpoint error can be *first-order*,

$$\|\bar{\mathbf{y}}^\theta - \mathbf{y}\| = \Theta(\|\mathbf{x}_1 - \mathbf{x}_2\|),$$

which motivates an additional correction to *synchronize the selections*.

We therefore *align the midpoint and backfill the residual*: First project the naive midpoint to the middle fiber, $\hat{\mathbf{y}} := \Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\bar{\mathbf{y}}^\theta)$, and then use the residual $\mathbf{y} - \hat{\mathbf{y}}$ to refine the endpoint representatives:

$$\mathbf{y}_i := \Pi_{\mathcal{S}(\mathbf{x}_i)}(\bar{\mathbf{y}}_i + (\mathbf{y} - \hat{\mathbf{y}})), \quad i = 1, 2.$$

This construction cancels the first-order branch mismatch in the convex combination and leaves only a quadratic remainder. Consequently, (4) holds while (5) remains valid.

Remark 2. *Think of $\mathcal{S}(\mathbf{x})$ as a family of convex “fibers”. The direct projections $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2$ may live on selections that are not synchronized with the selection containing $\mathbf{y}$, so their convex combination carries a first-order drift. Projecting $\bar{\mathbf{y}}^\theta$ to $\mathcal{S}(\mathbf{x}^\theta)$ identifies the correct selection at the midpoint; adding the same residual $\mathbf{y} - \hat{\mathbf{y}}$ to both endpoints moves them to the same selection as $\mathbf{y}$, making the first-order terms cancel in the average and exposing the desired $\mathcal{O}(\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$ behavior.*

Theorem 2 is not limited to the lower-level problem of BLO but applies to general parametric optimization problems. The established set smoothness property offers new insights into the structure of the solution mapping, which goes beyond the variational conditions considered in the literature [40, 27, 15, 7, 53].

Remark 3 (Local Conditions are Sufficient). *Suppose the solution mapping $\mathcal{S}$ is defined on a bounded convex domain $\mathcal{D} \subseteq \mathbb{R}^m$. To ensure Theorem 2, it suffices that Assumption 1 holds on the set $\mathcal{D} \times \mathcal{Y}$, where*

$$\mathcal{Y} = \text{Conv}\left(\bigcup_{\mathbf{x} \in \mathcal{D}} \mathcal{S}(\mathbf{x})\right) + \frac{1}{2} M_{\mathcal{S}} \text{diam}(\mathcal{D}) \mathbb{B}(\mathbf{0}, 1).$$

The following simple example shows that the set smoothness of $\mathcal{S}$ does not, in general, require Assumption 1. This suggests that alternative sufficient conditions may guarantee set smoothness; identifying such conditions is an interesting direction for future work.

Example 2. Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by $f(x, y) = g(\sin x + y)$, where $g : \mathbb{R} \rightarrow \mathbb{R}$ has a nonempty set of minimizers $\mathcal{V} = \arg \min_{z \in \mathbb{R}} g(z)$. Then the solution set admits a closed form:

$$\mathcal{S}(x) = \arg \min_{y \in \mathbb{R}} f(x, y) = \mathcal{V} - \sin x := \{v - \sin x : v \in \mathcal{V}\}.$$

In particular, S is 1-smooth in the sense of Definition 3 (since it is a translation of the fixed set $\mathcal{V}$ by the scalar $-\sin x$), even though f need not satisfy Assumption 1. $\square$

With Theorems 1 and 2 in place, we now state our main result on the weak convexity/concavity of the hyper-objective φ_o (resp. φ_p).

Theorem 3 (Weak convexity/concavity of the hyper-objectives). *Assume Assumptions 1 and 2. Let L_S be the set-smoothness modulus of S from Theorem 2, and define $\rho := M_F L_S + L_F (1 + L_S)$. Then the following hold:*

- (i) *The optimistic hyper-objective φ_o is ρ -weakly concave.*
- (ii) *The pessimistic hyper-objective φ_p is ρ -weakly convex.*

Proof of Theorem 3. Theorem 2 guarantees that the set-valued function S is L_S -smooth. Then, the result (ii) directly follows from Theorem 1. Hence, we only need to prove (i). Note that

$$-\varphi_o(\mathbf{x}) = -\min_{\mathbf{y} \in S(\mathbf{x})} F(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y} \in S(\mathbf{x})} -F(\mathbf{x}, \mathbf{y}).$$

We see that $-\varphi_o$ is ρ -weakly convex by Theorem 1. It follows that φ_o is ρ -weakly concave. $\square$

Theorem 3 establishes the weak concavity/convexity of the hyper-objectives in nonconvex-PŁ bilevel optimization (BLO). This stands in contrast to classical results (e.g., [19, Lemma 2.2]), which impose strong convexity of the lower level to obtain smooth hyper-objectives. Our result is significant because it places the minimization of these generally *nonsmooth* hyper-objectives within the framework of weakly concave/convex optimization. As a consequence, computing approximate Clarke hyperstationary points becomes tractable—an avenue we pursue in the next section. Crucially, all of these developments hinge on the *set smoothness* property (Definition 3), highlighting the utility of this notion.

Remark 4 (Lower-level Constraints Matter). *Under Assumptions 1 and 2, imposing an upper-level constraint $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^m$ with $\mathcal{X}$ nonempty, closed, and convex preserves the conclusions of Theorem 3: The functions $\varphi_o(\mathbf{x}) + \iota_{\mathcal{X}}(\mathbf{x})$ and $\varphi_p(\mathbf{x}) + \iota_{\mathcal{X}}(\mathbf{x})$ remain weakly concave and weakly convex, respectively, where $\iota_{\mathcal{X}}$ denotes the indicator of $\mathcal{X}$. In contrast, adding a lower-level constraint $\mathbf{y} \in \mathcal{Y}$ can destroy the weak concavity/convexity of the hyper-objectives, because the set smoothness of S may fail in this case; see Example 3. Developing structural conditions that recover such properties for lower-level constrained BLO is an interesting direction for future work.*

Example 3. Let $\mathcal{Y} = [0, 1] \times [0, 1]$. Consider the pessimistic bilevel problem with a lower-level constraint:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}} \max_{\mathbf{y} \in \mathbb{R}^2} \quad & -\mathbf{1}^\top \mathbf{y} \\ \text{s.t.} \quad & \mathbf{y} \in \arg \min_{\mathbf{y}' \in \mathcal{Y}} \|\mathbf{y}' - (x, 2)\|^2. \end{aligned} \tag{7}$$

Assumptions 1 and 2 are directly satisfied for (7), except for the unconstrained lower level; the only deviation here is the added constraint $\mathbf{y} \in \mathcal{Y}$.

The lower-level solution set is the projection of $(x, 2)$ onto the box $\mathcal{Y}$, hence

$$S(x) = \begin{cases} \{(0, 1)\}, & x \leq 0, \\ \{(x, 1)\}, & 0 \leq x \leq 1, \\ \{(1, 1)\}, & x \geq 1. \end{cases}$$

Therefore the pessimistic hyper-objective is

$$\varphi_p(x) = \begin{cases} -1, & x \leq 0, \\ -x - 1, & 0 \leq x \leq 1, \\ -2, & x \geq 1. \end{cases}$$

This function is *not* weakly convex. Indeed, for any $\rho \geq 0$ consider $h_\rho(x) := \varphi_p(x) + \frac{\rho}{2}x^2$. Then h_ρ has left and right derivatives at $x = 0$ given by $h'_\rho(0^-) = 0$ and $h'_\rho(0^+) = -1$ (the quadratic term has zero slope at 0), which violates the monotonicity of one-sided derivatives required by convexity. Hence no ρ makes h_ρ convex, i.e., φ_p is not weakly convex. $\square$

4 Computing Approximate Clarke Hyper-stationarity

Equipped with the weak convexity/concavity of the hyper-objectives, our next goal is to establish the computability of Clarke stationary points. First-order methods are impractical here because subgradients of the hyper-objectives are typically unavailable. In contrast, under mild conditions—e.g., $F(\mathbf{x}, \cdot)$ is concave (resp. convex) so that the inner maximization (resp. minimization) is tractable, the *function values* of the hyper-objectives can be (approximately) evaluated at a given $\mathbf{x}$ [16, 44]. This motivates the use of *zeroth-order* methods for minimizing hyper-objectives [7, 34]. In particular, we adopt the inexact zeroth-order method (IZOM) in Algorithm 1, which employs a deterministic subroutine $\mathcal{A}$ to approximately evaluate $\varphi_\beta(\mathbf{x})$ (with additive accuracy w) by solving the inner problem in (1); see [7, 24–26] for practical implementations of $\mathcal{A}$.

Algorithm 1 Inexact Zeroth-order Method (cf. [7, Algorithm 2])

Input: Radius $\varepsilon > 0$, iteration number $T \in \mathbb{N}$, stepsize η , initial point $\mathbf{x}_0 \in \mathbb{R}^m$, inexact error $w > 0$, and mode parameter $\beta \in \{1, 0\}$
for $t = 0, 1, \dots, T - 1$ **do**
 Sample $\mathbf{u}_t$ from the uniform distribution on the unit sphere in $\mathbb{R}^m$
 Compute $\mathcal{A}_w^\beta(\mathbf{x}_t + \varepsilon \mathbf{u}_t)$ and $\mathcal{A}_w^\beta(\mathbf{x}_t - \varepsilon \mathbf{u}_t)$ by subroutine $\mathcal{A}$
 Set $\tilde{G}(\mathbf{x}_t) = \frac{m}{2\varepsilon}(\mathcal{A}_w^\beta(\mathbf{x}_t + \varepsilon \mathbf{u}_t) - \mathcal{A}_w^\beta(\mathbf{x}_t - \varepsilon \mathbf{u}_t))\mathbf{u}_t$
 $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \tilde{G}(\mathbf{x}_t)$
end for
Output: $\bar{\mathbf{x}}$ uniformly chosen from $\{\mathbf{x}_t\}_{t=0}^{T-1}$

Algorithm 2 Deterministic Subroutine $\mathcal{A}$

Input: Accuracy $w > 0$, iterate point $\mathbf{x} \in \mathbb{R}^m$, and mode $\beta \in \{1, 0\}$
if $\beta = 1$ **then**
 Compute a value $\tilde{\varphi}(\mathbf{x})$ satisfying $|\tilde{\varphi}(\mathbf{x}) - \varphi_o(\mathbf{x})| \leq w$
else
 Compute a value $\tilde{\varphi}(\mathbf{x})$ satisfying $|\tilde{\varphi}(\mathbf{x}) - \varphi_p(\mathbf{x})| \leq w$
end if
Output: $\mathcal{A}_w^\beta(\mathbf{x}) = \tilde{\varphi}(\mathbf{x})$

Let $\varphi_{p,\gamma}$ denote the Moreau envelope of φ_p with parameter γ . We quantify hyper-stationarity as follows: In the optimistic case we use the approximate Clarke stationarity measure, i.e., Definition 1, while in the pessimistic case we use the gradient norm of the envelope, i.e., $\|\nabla \varphi_{p,\gamma}(\mathbf{x})\|$. These two criteria can be unified in principle via (3); in either form they are strictly stronger than the Goldstein stationarity measure; see Sec. 2.

We then present the main theorem of this section.

Theorem 4. *Suppose that Assumptions 1 and 2 hold. Given an iteration number $T \in \mathbb{N}$, set $\eta = \Theta(m^{-\frac{1}{2}}T^{-\frac{1}{2}})$, $\varepsilon = \mathcal{O}(T^{-\frac{1}{2}})$, $w = \mathcal{O}(m^{-\frac{3}{4}}T^{-\frac{3}{4}})$ for Algorithm 1. Then, the following hold:*

- (i) *Let $\Delta_o := \varphi_o(\mathbf{x}_0) - \min_{\mathbf{x}} \varphi_o(\mathbf{x}) + 2M_\varphi \varepsilon$ with M_φ given in Lemma 2. For optimistic BLO, we have*

$$\mathbb{E} \left[\text{dist} \left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\bar{\mathbf{x}}, \delta)} \partial \varphi_o(\mathbf{z}) \right)^2 \right] = \mathcal{O} \left(\frac{\sqrt{m}(\Delta_o + 1)}{\sqrt{T}} \right) \text{ with } \delta = \mathcal{O} \left(T^{-\frac{1}{4}} \right).$$

- (ii) *Let $\gamma \in (0, \frac{1}{\rho+1})$ with $\rho > 0$ given in Theorem 3, and $\Delta_p := \varphi_{p,\gamma}(\mathbf{x}_0) - \min_{\mathbf{x}} \varphi_{p,\gamma}(\mathbf{x})$. For pessimistic BLO, we have*

$$\mathbb{E} [\|\nabla \varphi_{p,\gamma}(\bar{\mathbf{x}})\|^2] = \mathcal{O} \left(\frac{\sqrt{m}(\Delta_p + 1)}{\sqrt{T}} \right).$$

Theorem 4 demonstrates, for the first time, that approximate Clarke hyper-stationarity is computable for nonconvex-PŁ BLO in both optimistic and pessimistic settings. This result significantly improves the existing computational guarantees for nonsmooth hyper-objective functions, which are mainly based on the Goldstein stationarity [7, 28]. The proof of the optimistic case relies on a Brøndsted-Rockafellar-like relation, details of which can be found in Appendix D.1.

5 Conclusion and Discussion

In this paper, we established the first theoretical guarantee for computing approximate Clarke hyper-stationarity in nonconvex-PŁ BLO. The key step is unveiling the hidden structural properties of hyper-objective functions via the newly introduced smoothness concept for set-valued functions. Specifically, we proved that (i) the smoothness of the set-valued function $\mathcal{Y}$ ensures the weak convexity of the function $\mathbf{x} \mapsto \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \phi(\mathbf{x}, \mathbf{y})$; and (ii) the lower-level solution set function of BLO satisfies set smoothness. Consequently, we obtained the weak convexity/concavity of hyper-objective functions. With these properties in hand, we showed that an inexact zeroth-order method can compute approximate Clarke stationary points of hyper-objective functions.

We believe that our developments contribute to a deeper understanding of the computability properties of BLO and open up several directions for future research. First, with the established structural properties, our work calls for designing faster algorithms for computing Clarke hyper-stationarity. Second, it would be valuable to generalize our methodology to establish adapted properties for BLO in other settings (e.g., structured lower-level constrained BLO [28]). Furthermore, our set smoothness property, along with Theorem 1, may find applications in other fields such as coupled minmax optimization [48] and set-valued optimization [27], where set-valued functions play a central role.

Acknowledgements

Jiajin Li was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2025-05817. Anthony Man-Cho So was supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) project CUHK 14204823.

References

- [1] N. Abolfazli, S. Sharifi, M. Fazlyab, and E. Yazdandoost Hamedani. Perturbed gradient descent via convex quadratic approximation for nonconvex bilevel optimization. *arXiv preprint arXiv:2504.17215*, 2025.
- [2] F. Atenas, C. Sagastizábal, P. J. Silva, and M. Solodov. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1):89–115, 2023.
- [3] Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pages 810–821. PMLR, 2020.
- [4] M. Brückner and T. Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–555, 2011.
- [5] A. Caprara, M. Carvalho, A. Lodi, and G. J. Woeginger. Bilevel knapsack with interdiction constraints. *INFORMS Journal on Computing*, 28(2):319–333, 2016.
- [6] H. Chen, H. Xu, R. Jiang, and A. M.-C. So. Lower-level duality based reformulation and majorization minimization algorithm for hyperparameter optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR, 2024.
- [7] L. Chen, J. Xu, and J. Zhang. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023.
- [8] L. Chen, J. Xu, and J. Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.
- [9] F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.
- [10] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.
- [11] A. Daniilidis and J. Malick. Filling the gap between lower-C1 and lower-C2 functions. *Journal of Convex Analysis*, 12(2):315–329, 2005.

- [12] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
- [13] S. Dempe. *Foundations of Bilevel Programming*. Springer Science & Business Media, 2002.
- [14] S. Dempe and A. Zemkoho. Bilevel optimization. In *Springer Optimization and Its Applications*, volume 161. Springer, 2020.
- [15] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings*, volume 543. Springer, 2009.
- [16] J. Dutta and T. Pandit. Algorithms for simple bilevel programming. *Bilevel Optimization: Advances and Next Challenges*, pages 253–291, 2020.
- [17] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.
- [18] 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.
- [19] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- [20] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.
- [21] 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.
- [22] 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.
- [23] R. Jiang and X. Li. Hölderian error bounds and Kurdyka-Łojasiewicz inequality for the trust region subproblem. *Mathematics of Operations Research*, 47(4):3025–3050, 2022.
- [24] R. Jiang, N. Abolfazli, A. Mokhtari, and E. Y. Hamedani. Generalized Frank-Wolfe algorithm for bilevel optimization. *arXiv preprint arXiv:2206.08868*, 2022.
- [25] R. Jiang, X. Shi, and J. Wang. A near-optimal algorithm for convex simple bilevel optimization under weak assumptions. *arXiv preprint arXiv:2409.08948*, 2024.
- [26] H. D. Kaushik and F. Yousefian. A method with convergence rates for optimization problems with variational inequality constraints. *SIAM Journal on Optimization*, 31(3):2171–2198, 2021.
- [27] A. A. Khan, C. Tammer, and C. Zalinescu. *Set-valued Optimization*. Springer, 2016.
- [28] P. Khanduri, I. Tsaknakis, Y. Zhang, S. Liu, and M. Hong. A doubly stochastically perturbed algorithm for linearly constrained bilevel optimization. *arXiv preprint arXiv:2504.04545*, 2025.
- [29] V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.
- [30] G. Kornowski and O. Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 34:324–334, 2021.
- [31] J. Kwon, D. Kwon, S. Wright, and R. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023.
- [32] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

- [33] F.-Y. Liao, L. Ding, and Y. Zheng. Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *6th Annual Learning for Dynamics & Control Conference*, pages 993–1005. PMLR, 2024.
- [34] T. Lin, Z. Zheng, and M. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35: 26160–26175, 2022.
- [35] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! Bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.
- [36] 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.
- [37] R. Liu, Y. Liu, W. Yao, S. Zeng, and J. Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *International Conference on Machine Learning*, pages 21839–21866. PMLR, 2023.
- [38] S. Lu. Slm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. *Advances in Neural Information Processing Systems*, 36:80414–80454, 2023.
- [39] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.
- [40] B. S. Mordukhovich. *Second-order Variational Analysis in Optimization, Variational Stability, and Control: Theory, Algorithms, Applications*. Springer Nature, 2024.
- [41] P. Nazari, D. A. Tarzanagh, and G. Michailidis. Adaptive first-and zeroth-order methods for weakly convex stochastic optimization problems. *arXiv preprint arXiv:2005.09261*, 2020.
- [42] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.
- [43] S. M. Robinson. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. *Mathematical Programming*, 86:41–50, 1999.
- [44] Y. Shehu, P. T. Vuong, and A. Zemkoho. An inertial extrapolation method for convex simple bilevel optimization. *Optimization Methods and Software*, 36(1):1–19, 2021.
- [45] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.
- [46] L. Tian and A. M.-C. So. On the hardness of computing near-approximate stationary points of clarke regular nonsmooth nonconvex problems and certain DC programs. In *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.
- [47] L. Tian, K. Zhou, and A. M.-C. So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.
- [48] 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, 2023. doi: 10.1137/21M1462428.
- [49] J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.
- [50] J. Wang, H. Chen, R. Jiang, X. Li, and Z. Li. Fast algorithms for stackelberg prediction game with least squares loss. In *International Conference on Machine Learning*, pages 10708–10716. PMLR, 2021.
- [51] W. Wiesemann, A. Tsoukalas, P.-M. Kleniati, and B. Rustem. Pessimistic bilevel optimization. *SIAM Journal on Optimization*, 23(1):353–380, 2013.

- [52] Q. Xiao, S. Lu, and T. Chen. An generalized alternating method for bilevel optimization under the Polyak-Łojasiewicz condition. *Advances in Neural Information Processing Systems*, 36: 63847–63873, 2023.
- [53] W. Yao and X. Yang. Relative Lipschitz-like property of parametric systems via projectional coderivatives. *SIAM Journal on Optimization*, 33(3):2021–2040, 2023.
- [54] W. Yao, H. Yin, S. Zeng, and J. Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. *arXiv preprint arXiv:2406.01992*, 2024.
- [55] S. Zeng, T. T. Doan, and J. Romberg. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *SIAM Journal on Optimization*, 34(1): 946–976, 2024.
- [56] L. Zhao, D. Chen, D. Zhu, and X. Li. Randomized coordinate subgradient method for nonsmooth optimization. *arXiv preprint arXiv:2206.14981*, 2022.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The developments established in Sec. 3, 4 support the claims made in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: We used Example 3 to show that our work does not apply to the lower-level constrained case. In Sec. 5, we gave some remarks that reflect the limitations of our work and call for future research.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

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

Answer: [Yes]

Justification: We provided the full set of assumptions for our theoretical results, as detailed in Sec. 3 and 4. Complete proofs can be found in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

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

Answer: [Yes]

Justification: The authors are sure that the research conducted in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

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

Answer: [NA]

Justification: This paper considers the theoretical computability of BLO. We do not see any potential negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

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

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper focuses on the theoretical computability of hyper-stationarity for BLO and does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Missing Proofs for Sec. 2

A.1 Proof of Lemma 1

Proof. Thanks to Assumption 1 (A3), we have

$$\text{dist}(\mathbf{y}_1, \mathcal{S}(\mathbf{x}_2)) \leq \tau \|\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_1)\|,$$

for any $\mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)$. Moreover, we have

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_1)\| = \|\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1)\| \leq L_f \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where the equality follows from $\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) = \mathbf{0}$, and the inequality is due to the L_f -smoothness of f . Putting them together yields $\text{dist}(\mathbf{y}_1, \mathcal{S}(\mathbf{x}_2)) \leq L_f \tau \|\mathbf{x}_1 - \mathbf{x}_2\|$ for all $\mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)$.

By the same argument with $\mathbf{x}_1$ and $\mathbf{x}_2$ interchanged, we have $\text{dist}(\mathbf{y}_2, \mathcal{S}(\mathbf{x}_1)) \leq L_f \tau \|\mathbf{x}_1 - \mathbf{x}_2\|$ for all $\mathbf{y}_2 \in \mathcal{S}(\mathbf{x}_2)$. By the definition of $d_H(\mathcal{S}(\mathbf{x}_1), \mathcal{S}(\mathbf{x}_2))$, we conclude

$$\begin{aligned} d_H(\mathcal{S}(\mathbf{x}_1), \mathcal{S}(\mathbf{x}_2)) &= \max \left\{ \sup_{\mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)} \text{dist}(\mathbf{y}_1, \mathcal{S}(\mathbf{x}_2)), \sup_{\mathbf{y}_2 \in \mathcal{S}(\mathbf{x}_2)} \text{dist}(\mathbf{y}_2, \mathcal{S}(\mathbf{x}_1)) \right\} \\ &\leq L_f \tau \|\mathbf{x}_1 - \mathbf{x}_2\|. \end{aligned}$$

This completes the proof. $\square$

A.2 Proof of Lemma 2

Proof. We prove the M_φ -Lipschitz continuity for φ_p only, as the argument for φ_o is entirely analogous.

Then, it suffices to show that for any $\mathbf{x}_1, \mathbf{x}_2 \in \text{dom}(\varphi_p)$,

$$|\varphi_p(\mathbf{x}_1) - \varphi_p(\mathbf{x}_2)| \leq M_\varphi \|\mathbf{x}_2 - \mathbf{x}_1\|.$$

Note that $\mathcal{S}(\mathbf{x})$ is closed but not necessarily compact. We can only find a sequence $\{\mathbf{y}_1^k\}_{k \in \mathbb{N}} \subseteq \mathcal{S}(\mathbf{x}_1)$ such that $F(\mathbf{x}_1, \mathbf{y}_1^k) \rightarrow \sup_{\mathbf{y} \in \mathcal{S}(\mathbf{x}_1)} F(\mathbf{x}_1, \mathbf{y}) = \varphi_p(\mathbf{x}_1)$. Let $\mathbf{y}_2^k := \Pi_{\mathcal{S}(\mathbf{x}_2)}(\mathbf{y}_1^k)$. Then, by $\mathbf{y}_2^k \in \mathcal{S}(\mathbf{x}_2)$, we have $\varphi_p(\mathbf{x}_2) = \sup_{\mathbf{y} \in \mathcal{S}(\mathbf{x}_2)} F(\mathbf{x}_2, \mathbf{y}) \geq F(\mathbf{x}_2, \mathbf{y}_2^k)$. This observation, combined with the M_F -Lipschitz continuity of F , yields

$$F(\mathbf{x}_1, \mathbf{y}_1^k) - \varphi_p(\mathbf{x}_2) \leq F(\mathbf{x}_1, \mathbf{y}_1^k) - F(\mathbf{x}_2, \mathbf{y}_2^k) \leq M_F (\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1^k - \mathbf{y}_2^k\|),$$

where the last inequality follows from Assumption 2 (B1).

Moreover, we have

$$\|\mathbf{y}_1^k - \mathbf{y}_2^k\| = \|\mathbf{y}_1^k - \Pi_{\mathcal{S}(\mathbf{x}_2)}(\mathbf{y}_1^k)\| = \text{dist}(\mathbf{y}_1^k, \mathcal{S}(\mathbf{x}_2)) \leq d_H(\mathcal{S}(\mathbf{x}_1), \mathcal{S}(\mathbf{x}_2)) \leq M_S \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where the last inequality is due to Lemma 1. Then, we can see that for all $k \in \mathbb{N}$,

$$F(\mathbf{x}_1, \mathbf{y}_1^k) - \varphi_p(\mathbf{x}_2) \leq M_F(1 + M_S) \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Letting $k \rightarrow \infty$ and recalling $F(\mathbf{x}_1, \mathbf{y}_1^k) \leftarrow \varphi_p(\mathbf{x}_1)$, we obtain

$$\varphi_p(\mathbf{x}_1) - \varphi_p(\mathbf{x}_2) \leq M_F(1 + M_S) \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

which completes the proof. $\square$

A.3 Proof of Fact 1

Proof. By [49, Proposition 4.4], the weakly convex function $-g$ is locally Lipschitz (hence so is g), so the Clarke subdifferential ∂g is well defined. Moreover, the Clarke subdifferential satisfies $\partial(-g) = -\partial g$ [10, Prop. 2.3.1]. Applying Lemma 3 (iii) to $-g$ then yields the claim. $\square$

B Proof of Theorem 1

Proof. To establish the weak convexity of ϕ , we verify the condition in Lemma 3 (ii). Specifically, we show that there exists a constant $\rho > 0$ such that, for all $\theta \in [0, 1]$ and all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$,

$$\phi(\mathbf{x}^\theta) \leq \theta\phi(\mathbf{x}_1) + (1 - \theta)\phi(\mathbf{x}_2) + \frac{\rho}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2, \quad (8)$$

where we let $\mathbf{x}^\theta := \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$ for notation convenience.

By definition, $\phi(\mathbf{x}^\theta) = \max_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x}^\theta)} g(\mathbf{x}^\theta, \mathbf{y}')$. Hence (8) will follow if we show that, for all $\mathbf{y} \in \mathcal{Y}(\mathbf{x}^\theta)$,

$$g(\mathbf{x}^\theta, \mathbf{y}) - \theta\phi(\mathbf{x}_1) - (1 - \theta)\phi(\mathbf{x}_2) \leq \frac{\rho}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \quad (9)$$

Taking the maximization over $\mathbf{y} \in \mathcal{Y}(\mathbf{x}^\theta)$ then yields (8).

We proceed via the $L_{\mathcal{Y}}$ -smoothness of $\mathcal{Y}$, which guarantees the existence of $\mathbf{y}_1 \in \mathcal{Y}(\mathbf{x}_1)$ and $\mathbf{y}_2 \in \mathcal{Y}(\mathbf{x}_2)$ such that

$$\|\theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2 - \mathbf{y}\| \leq \frac{L_{\mathcal{Y}}}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2; \quad (10)$$

$$\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \leq L_{\mathcal{Y}}\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \quad (11)$$

Then, using the fact that $\phi(\mathbf{x}_i) = \max_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x}_i)} g(\mathbf{x}_i, \mathbf{y}') \geq g(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, 2$, we have

$$\begin{aligned} & g(\mathbf{x}^\theta, \mathbf{y}) - \theta\phi(\mathbf{x}_1) - (1 - \theta)\phi(\mathbf{x}_2) \\ & \leq g(\mathbf{x}^\theta, \mathbf{y}) - \theta g(\mathbf{x}_1, \mathbf{y}_1) - (1 - \theta)g(\mathbf{x}_2, \mathbf{y}_2) \\ & \leq g(\mathbf{x}^\theta, \mathbf{y}) - g(\mathbf{x}^\theta, \theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2) + \frac{L_g}{2}\theta(1 - \theta)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2\right), \end{aligned}$$

where the last inequality is due to the L_g -smoothness of g on $\mathcal{D} \times \text{Conv}(\bigcup_{\mathbf{x} \in \mathcal{D}} \mathcal{Y}(\mathbf{x}))$ and Fact 2.

This, together with the M_g -Lipschitz continuity of g w.r.t. $\mathbf{y}$ and the set-smoothness inequalities (10) and (11), yields

$$\begin{aligned} & g(\mathbf{x}^\theta, \mathbf{y}) - \theta\phi(\mathbf{x}_1) - (1 - \theta)\phi(\mathbf{x}_2) \\ & \leq M_g\|\theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2 - \mathbf{y}\| + \frac{L_g}{2}\theta(1 - \theta)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2\right) \\ & \leq \frac{M_g L_{\mathcal{Y}}}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \frac{L_g}{2}\theta(1 - \theta)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + L_{\mathcal{Y}}\|\mathbf{x}_1 - \mathbf{x}_2\|^2\right) \\ & = \frac{M_g L_{\mathcal{Y}} + L_g(1 + L_{\mathcal{Y}})}{2}\theta(1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|^2, \end{aligned}$$

for all $\mathbf{y} \in \mathcal{Y}(\mathbf{x}^\theta)$. We prove the desired inequality (9) with $\rho = M_g L_{\mathcal{Y}} + L_g(1 + L_{\mathcal{Y}})$. □

C Proof of Theorem 2

We start by stating a lemma that will be used in the sequel.

Lemma 5. Under Assumption 1 (A1), for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$, we have

$$\begin{aligned} \text{(i)} \quad & \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2)\| \\ & \leq \frac{H_f}{2}\left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2\right); \\ \text{(ii)} \quad & \|\theta \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) + (1 - \theta) \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)\| \\ & \leq \frac{H_f}{2}\theta(1 - \theta)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2\right), \quad \forall \theta \in [0, 1], \end{aligned}$$

where $\mathbf{x}^\theta := \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$ and $\mathbf{y}^\theta := \theta\mathbf{y}_1 + (1 - \theta)\mathbf{y}_2$.

Proof of Lemma 5. (i) The argument directly follows from [42, Lemma 1]:

$$\begin{aligned}
& \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2)\| \\
&= \left\| \int_0^1 \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{y}_1 + t(\mathbf{y}_2 - \mathbf{y}_1))(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2) dt \right. \\
&\quad \left. - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2) \right\| \\
&= \left\| \int_0^1 (\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{y}_1 + t(\mathbf{y}_2 - \mathbf{y}_1)) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2))(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2) dt \right\| \\
&\leq \int_0^1 \|(\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{y}_1 + t(\mathbf{y}_2 - \mathbf{y}_1)) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2))(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2)\| dt \\
&\leq \int_0^1 \|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1), \mathbf{y}_1 + t(\mathbf{y}_2 - \mathbf{y}_1)) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)\| \|\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2\| dt \\
&\leq \int_0^1 H_f t \|\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2\| \cdot \|\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2\| dt \\
&= \frac{H_f}{2} (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2),
\end{aligned}$$

where the first inequality is due to Jensen's inequality; the second inequality is due to the Cauchy inequality; and the third inequality uses the H_f -Lipschitz continuity of $\nabla \nabla_{\mathbf{y}} f$.

(ii) Using the result of (i), we have

$$\begin{aligned}
& \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)(\mathbf{x}_1 - \mathbf{x}^\theta, \mathbf{y}_1 - \mathbf{y}^\theta)\| \\
&\leq \frac{H_f}{2} (\|\mathbf{x}_1 - \mathbf{x}^\theta\|^2 + \|\mathbf{y}_1 - \mathbf{y}^\theta\|^2).
\end{aligned}$$

It follows from $\mathbf{x}^\theta = \theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2$ and $\mathbf{y}^\theta = \theta \mathbf{y}_1 + (1 - \theta) \mathbf{y}_2$ that

$$\begin{aligned}
& \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta) - (1 - \theta) \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2)\| \\
&\leq \frac{H_f}{2} (1 - \theta)^2 (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2).
\end{aligned} \tag{12}$$

Using the same arguments with $(\mathbf{x}_1, \mathbf{y}_1)$ replaced by $(\mathbf{x}_2, \mathbf{y}_2)$, we have

$$\begin{aligned}
& \|\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta) - \theta \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)(\mathbf{x}_2 - \mathbf{x}_1, \mathbf{y}_2 - \mathbf{y}_1)\| \\
&\leq \frac{H_f}{2} \theta^2 (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2).
\end{aligned} \tag{13}$$

Then, the desired inequality follows from the weighted sum $\theta \times (12) + (1 - \theta) \times (13)$, and the triangle inequality that

$$\begin{aligned}
& \theta \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta) - (1 - \theta) \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{y}_1 - \mathbf{y}_2)\| \\
&\quad + (1 - \theta) \|\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta) - \theta \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)(\mathbf{x}_2 - \mathbf{x}_1, \mathbf{y}_2 - \mathbf{y}_1)\| \\
&\geq \|\theta \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) + (1 - \theta) \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}^\theta)\|.
\end{aligned}$$

□

We are now ready to prove Proposition 2. Fix $\theta \in [0, 1]$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, and let $\mathbf{y} \in \mathcal{S}(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2)$. Our goal is to construct $\mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)$ and $\mathbf{y}_2 \in \mathcal{S}(\mathbf{x}_2)$ such that (4) and (5) hold.

C.1 Step 1: Choose projection points as the candidate approximation points.

Let $\bar{\mathbf{y}}_1 := \Pi_{\mathcal{S}(\mathbf{x}_1)}(\mathbf{y})$ and $\bar{\mathbf{y}}_2 := \Pi_{\mathcal{S}(\mathbf{x}_2)}(\mathbf{y})$. For simplicity, we write $\mathbf{x}^\theta = \theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2$ and $\bar{\mathbf{y}}^\theta = \theta \bar{\mathbf{y}}_1 + (1 - \theta) \bar{\mathbf{y}}_2$. We start with giving basic estimates on $\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|$ and $\text{dist}(\bar{\mathbf{y}}^\theta, \mathcal{S}(\mathbf{x}^\theta))$, in Claim 1.

Claim 1. Let $L_0 := H_f \tau(1 + M_S^2)/2$. The following hold:

$$\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| \leq M_S \|\mathbf{x}_1 - \mathbf{x}_2\| \tag{14}$$

$$\text{dist}(\bar{\mathbf{y}}^\theta, \mathcal{S}(\mathbf{x}^\theta)) \leq L_0 \theta(1 - \theta) \|\mathbf{x}_1 - \mathbf{x}_2\|^2. \tag{15}$$

Proof of Claim 1. We have

$$\begin{aligned}
\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| &\leq \|\bar{\mathbf{y}}_1 - \mathbf{y}\| + \|\mathbf{y} - \bar{\mathbf{y}}_2\| \\
&= \text{dist}(\mathbf{y}, \mathcal{S}(\mathbf{x}_1)) + \text{dist}(\mathbf{y}, \mathcal{S}(\mathbf{x}_2)) \\
&\leq d_H(\mathcal{S}(\mathbf{x}^\theta), \mathcal{S}(\mathbf{x}_1)) + d_H(\mathcal{S}(\mathbf{x}^\theta), \mathcal{S}(\mathbf{x}_2)) \\
&\leq M_S \|\mathbf{x}_1 - (\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2)\| + M_S \|\mathbf{x}_2 - (\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2)\| \\
&= M_S \|\mathbf{x}_1 - \mathbf{x}_2\|,
\end{aligned}$$

where the second inequality follows from $\mathbf{y} \in \mathcal{S}(\mathbf{x}^\theta)$ and the definition of the Hausdorff distance; the third inequality is due to Lemma 1 and $\mathbf{x}^\theta = \theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2$. This proves (14).

We then prove (15). First, by the H_f -Lipschitz continuity of $\nabla \nabla_{\mathbf{y}} f$ and Lemma 5 (ii), we have

$$\begin{aligned}
&\|\theta \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) + (1 - \theta) \nabla_{\mathbf{y}} f(\mathbf{x}_2, \bar{\mathbf{y}}_2) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \bar{\mathbf{y}}^\theta)\| \\
&\leq \frac{H_f}{2} \theta(1 - \theta) (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|^2) \\
&\leq \frac{H_f}{2} \theta(1 - \theta) (1 + M_S^2) \|\mathbf{x}_1 - \mathbf{x}_2\|^2,
\end{aligned} \tag{16}$$

where the second inequality is due to (14).

Since $\bar{\mathbf{y}}_1 \in \mathcal{S}(\mathbf{x}_1)$ and $\bar{\mathbf{y}}_2 \in \mathcal{S}(\mathbf{x}_2)$, we have the first-order optimality

$$\nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) = 0, \quad \nabla_{\mathbf{y}} f(\mathbf{x}_2, \bar{\mathbf{y}}_2) = 0.$$

Substituting these identities into (16) yields

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \bar{\mathbf{y}}^\theta)\| \leq \frac{H_f}{2} \theta(1 - \theta) (1 + M_S^2) \|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Combining the error bound from Assumption 1, $\text{dist}(\bar{\mathbf{y}}^\theta, \mathcal{S}(\mathbf{x}^\theta)) \leq \tau \|\nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \bar{\mathbf{y}}^\theta)\|$, with the preceding inequality, we obtain

$$\text{dist}(\bar{\mathbf{y}}^\theta, \mathcal{S}(\mathbf{x}^\theta)) \leq \frac{H_f \tau}{2} \theta(1 - \theta) (1 + M_S^2) \|\mathbf{x}_1 - \mathbf{x}_2\|^2,$$

which accords with (15) with $L_0 = H_f \tau (1 + M_S^2) / 2$. Claim 1 is proved. $\square$

Note that (15) controls only the distance from $\bar{\mathbf{y}}^\theta$ to the set $\mathcal{S}(\mathbf{x}^\theta)$; it does not guarantee that the specific point $\theta \bar{\mathbf{y}}_1 + (1 - \theta) \bar{\mathbf{y}}_2$ lies on the same branch as the metric projection $\Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\bar{\mathbf{y}}^\theta)$. Hence one cannot conclude that $\|\theta \bar{\mathbf{y}}_1 + (1 - \theta) \bar{\mathbf{y}}_2 - \mathbf{y}\| = \mathcal{O}(\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$ in general. Therefore we cannot simply take $\mathbf{y}_1 = \bar{\mathbf{y}}_1$ and $\mathbf{y}_2 = \bar{\mathbf{y}}_2$, which motivates the rectification in Step 2.

C.2 Step 2: Translate the candidate approximation points

Let $\hat{\mathbf{y}} := \Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\bar{\mathbf{y}}^\theta)$, $\hat{\mathbf{y}}_1 := \bar{\mathbf{y}}_1 + (\mathbf{y} - \hat{\mathbf{y}})$, and $\hat{\mathbf{y}}_2 := \bar{\mathbf{y}}_2 + (\mathbf{y} - \hat{\mathbf{y}})$. We will bound the three quantities $\|\hat{\mathbf{y}} - \mathbf{y}\|$, $\text{dist}(\hat{\mathbf{y}}_1, \mathcal{S}(\mathbf{x}_1))$, and $\text{dist}(\hat{\mathbf{y}}_2, \mathcal{S}(\mathbf{x}_2))$.

Claim 2. Let $L_1 := H_f \tau (1 + 17 M_S^2) / 2$. The following hold:

$$\|\hat{\mathbf{y}} - \mathbf{y}\| \leq 2\theta(1 - \theta) M_S \|\mathbf{x}_1 - \mathbf{x}_2\|; \tag{17}$$

$$\text{dist}(\hat{\mathbf{y}}_1, \mathcal{S}(\mathbf{x}_1)) \leq L_1 (1 - \theta)^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2; \quad \text{dist}(\hat{\mathbf{y}}_2, \mathcal{S}(\mathbf{x}_2)) \leq L_1 \theta^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2. \tag{18}$$

Proof of Claim 2. We begin with the proof of (17). The non-expansiveness of the projection operator $\Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\cdot)$ yields

$$\|\hat{\mathbf{y}} - \mathbf{y}\| = \|\Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\bar{\mathbf{y}}^\theta) - \Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\mathbf{y})\| \leq \|\bar{\mathbf{y}}^\theta - \mathbf{y}\| \leq \theta \|\bar{\mathbf{y}}_1 - \mathbf{y}\| + (1 - \theta) \|\bar{\mathbf{y}}_2 - \mathbf{y}\|. \tag{19}$$

Moreover, by $\bar{\mathbf{y}}_i = \Pi_{\mathcal{S}(\mathbf{x}_i)}(\mathbf{y})$ for $i = 1, 2$, $\mathbf{y} \in \mathcal{S}(\mathbf{x}^\theta)$, and the M_S -Lipschitz continuity of $\mathcal{S}$ from Lemma 1, we have

$$\|\bar{\mathbf{y}}_i - \mathbf{y}\| = \text{dist}(\mathbf{y}, \mathcal{S}(\mathbf{x}_i)) \leq d_H(\mathcal{S}(\mathbf{x}^\theta), \mathcal{S}(\mathbf{x}_i)) \leq M_S \|\mathbf{x}_i - \mathbf{x}^\theta\|, \quad i = 1, 2. \tag{20}$$

Combining (19) and (20) leads to

$$\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \theta M_S \|\mathbf{x}_1 - \mathbf{x}^\theta\| + (1 - \theta) M_S \|\mathbf{x}_2 - \mathbf{x}^\theta\| = 2\theta(1 - \theta) M_S \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

We then continue to prove (18). It suffices to upper bound $\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1)\|$, due to Assumption 1 (A3). By Lemma 5, we have $\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1)$:

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1)(\mathbf{0}, \hat{\mathbf{y}}_1 - \bar{\mathbf{y}}_1)\| \leq \frac{H_f}{2} \|\hat{\mathbf{y}}_1 - \bar{\mathbf{y}}_1\|^2.$$

Noticing $\bar{\mathbf{y}}_1 \in \mathcal{S}(\mathbf{x}_1)$, we have $\nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) = \mathbf{0}$, and thus

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1)(\mathbf{0}, \hat{\mathbf{y}}_1 - \bar{\mathbf{y}}_1)\| \leq \frac{H_f}{2} \|\hat{\mathbf{y}}_1 - \bar{\mathbf{y}}_1\|^2.$$

Using the triangle inequality and the identity $\hat{\mathbf{y}}_1 - \bar{\mathbf{y}}_1 = \mathbf{y} - \hat{\mathbf{y}}$ (by the definition of $\hat{\mathbf{y}}_1$), we obtain

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1)\| \leq \|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1)(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\| + \frac{H_f}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2. \quad (21)$$

We proceed to control $\|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1)(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\|$. To do so, we first estimate a closely related norm $\|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y})(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\|$. We apply Lemma 5 again to obtain

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \hat{\mathbf{y}}) - \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y})(\mathbf{0}, \hat{\mathbf{y}} - \mathbf{y})\| \leq \frac{H_f}{2} \|\hat{\mathbf{y}} - \mathbf{y}\|^2.$$

Note that both $\mathbf{y}$ and $\hat{\mathbf{y}}$ belong to the set $\mathcal{S}(\mathbf{x}^\theta)$, which leads to

$$\nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \hat{\mathbf{y}}) = \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}) = \mathbf{0}.$$

It follows that

$$\|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y})(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\| \leq \frac{H_f}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2. \quad (22)$$

Putting everything together yields

$$\begin{aligned} & \|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1)\| \\ & \leq \|(\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y}))(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\| + \|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y})(\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})\| \\ & \quad + \frac{H_f}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ & \leq \|\nabla \nabla_{\mathbf{y}} f(\mathbf{x}_1, \bar{\mathbf{y}}_1) - \nabla \nabla_{\mathbf{y}} f(\mathbf{x}^\theta, \mathbf{y})\| \|\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}}\| + \frac{H_f}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \frac{H_f}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ & \leq H_f (\|\mathbf{x}_1 - \mathbf{x}^\theta\| + \|\bar{\mathbf{y}}_1 - \mathbf{y}\|) \cdot \|\mathbf{y} - \hat{\mathbf{y}}\| + H_f \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ & \leq \frac{H_f}{2} \|\mathbf{x}_1 - \mathbf{x}^\theta\|^2 + \frac{H_f}{2} \|\bar{\mathbf{y}}_1 - \mathbf{y}\|^2 + H_f \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + H_f \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ & = \frac{H_f}{2} \|\mathbf{x}_1 - \mathbf{x}^\theta\|^2 + \frac{H_f}{2} \|\bar{\mathbf{y}}_1 - \mathbf{y}\|^2 + 2H_f \|\mathbf{y} - \hat{\mathbf{y}}\|^2, \end{aligned}$$

where the second inequality uses the definition of matrix's l_2 norm and (22); the third inequality is due to the H_f -Lipschitz continuity of $\nabla \nabla_{\mathbf{y}} f$ and the triangle inequality; the forth inequality is due to the Cauchy inequality.

Combining the above estimate with $\|\mathbf{x}_1 - \mathbf{x}^\theta\| = (1 - \theta)\|\mathbf{x}_1 - \mathbf{x}_2\|$, (20) (with $i = 1$), and the bound from (17), namely $\|\mathbf{y} - \hat{\mathbf{y}}\| \leq 2(1 - \theta)M_S\|\mathbf{x}_1 - \mathbf{x}_2\|$, we obtain

$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \hat{\mathbf{y}}_1)\| \leq \frac{H_f (1 + 17M_S^2)}{2} (1 - \theta)^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2. \quad (23)$$

Finally, armed with the error bound condition in Assumption 1 (A3), we have

$$\text{dist}(\hat{\mathbf{y}}_1, \mathcal{S}(\mathbf{x}_1)) \leq \frac{H_f \tau (1 + 17M_S^2)}{2} (1 - \theta)^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

By the symmetric arguments, we can also have $\text{dist}(\hat{\mathbf{y}}_2, \mathcal{S}(\mathbf{x}_2)) \leq H_f \tau (1 + 17M_S^2) \theta^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2 / 2$. We finished our proof for Claim 2. $\square$

C.3 Step 3: Define approximation points

With the preparations in Steps 1–2, we now define the approximation points

$$\mathbf{y}_1 := \Pi_{\mathcal{S}(\mathbf{x}_1)}(\hat{\mathbf{y}}_1), \quad \mathbf{y}_2 := \Pi_{\mathcal{S}(\mathbf{x}_2)}(\hat{\mathbf{y}}_2).$$

To establish the smoothness of $\mathcal{S}$, it remains to show that there exists a constant $L_{\mathcal{S}} > 0$ such that

$$\|\theta \mathbf{y}_1 + (1 - \theta) \mathbf{y}_2 - \mathbf{y}\| \leq \frac{L_{\mathcal{S}}}{2} \theta(1 - \theta) \|\mathbf{x}_1 - \mathbf{x}_2\|^2; \quad (24)$$

$$\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \leq L_{\mathcal{S}} \|\mathbf{x}_1 - \mathbf{x}_2\|^2. \quad (25)$$

We first prove the inequality (24). We have

$$\begin{aligned} & \|\theta \mathbf{y}_1 + (1 - \theta) \mathbf{y}_2 - \mathbf{y}\| \\ &= \|\theta(\mathbf{y}_1 - \hat{\mathbf{y}}_1) + (1 - \theta)(\mathbf{y}_2 - \hat{\mathbf{y}}_2) + (\theta \hat{\mathbf{y}}_1 + (1 - \theta) \hat{\mathbf{y}}_2 - \mathbf{y})\| \\ &\leq \theta \|\mathbf{y}_1 - \hat{\mathbf{y}}_1\| + (1 - \theta) \|\mathbf{y}_2 - \hat{\mathbf{y}}_2\| + \|(\theta \hat{\mathbf{y}}_1 + (1 - \theta) \hat{\mathbf{y}}_2 - \mathbf{y})\| \\ &= \theta \text{dist}(\hat{\mathbf{y}}_1, \mathcal{S}(\mathbf{x}_1)) + (1 - \theta) \text{dist}(\hat{\mathbf{y}}_2, \mathcal{S}(\mathbf{x}_2)) + \|\bar{\mathbf{y}}^\theta - \hat{\mathbf{y}}\| \\ &= \theta \text{dist}(\hat{\mathbf{y}}_1, \mathcal{S}(\mathbf{x}_1)) + (1 - \theta) \text{dist}(\hat{\mathbf{y}}_2, \mathcal{S}(\mathbf{x}_2)) + \text{dist}(\bar{\mathbf{y}}^\theta, \mathcal{S}(\mathbf{x}^\theta)), \end{aligned}$$

where the second equality follows from $\mathbf{y}_i = \Pi_{\mathcal{S}(\mathbf{x}_i)}(\hat{\mathbf{y}}_i)$ and $\hat{\mathbf{y}}_i = \bar{\mathbf{y}}_i + \mathbf{y} - \hat{\mathbf{y}}$ for $i = 1, 2$; the final one is due to $\hat{\mathbf{y}} = \Pi_{\mathcal{S}(\mathbf{x}^\theta)}(\bar{\mathbf{y}}^\theta)$.

This, together with (15) and (18), implies

$$\|\theta \mathbf{y}_1 + (1 - \theta) \mathbf{y}_2 - \mathbf{y}\| \leq (L_0 + L_1) \theta(1 - \theta) \|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Hence, (24) holds if $L_{\mathcal{S}} \geq 2(L_0 + L_1)$. It is left to show (25).

To prove (25), we estimate the distance $\|\mathbf{y}_1 - \mathbf{y}_2\|$. To begin, we apply the triangle inequality to obtain

$$\|\mathbf{y}_1 - \mathbf{y}_2\| \leq \|\mathbf{y}_1 - \bar{\mathbf{y}}_1\| + \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| + \|\bar{\mathbf{y}}_2 - \mathbf{y}_2\|.$$

Then, use the definition $\mathbf{y}_i = \Pi_{\mathcal{S}(\mathbf{x}_i)}(\hat{\mathbf{y}}_i) = \Pi_{\mathcal{S}(\mathbf{x}_i)}(\bar{\mathbf{y}}_i + \mathbf{y} - \hat{\mathbf{y}})$ and notice $\bar{\mathbf{y}}_i \in \mathcal{S}(\mathbf{x}_i)$ for $i = 1, 2$. We have

$$\begin{aligned} \|\mathbf{y}_1 - \mathbf{y}_2\| &\leq \|\Pi_{\mathcal{S}(\mathbf{x}_1)}(\bar{\mathbf{y}}_1 + (\mathbf{y} - \hat{\mathbf{y}})) - \Pi_{\mathcal{S}(\mathbf{x}_1)}(\bar{\mathbf{y}}_1)\| + \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| \\ &\quad + \|\Pi_{\mathcal{S}(\mathbf{x}_2)}(\bar{\mathbf{y}}_2) - \Pi_{\mathcal{S}(\mathbf{x}_2)}(\bar{\mathbf{y}}_2 + (\mathbf{y} - \hat{\mathbf{y}}))\| \\ &\leq \|\hat{\mathbf{y}} - \mathbf{y}\| + \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| + \|\hat{\mathbf{y}} - \mathbf{y}\| \\ &= 2 \|\hat{\mathbf{y}} - \mathbf{y}\| + \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|, \end{aligned} \quad (26)$$

where the second inequality follows from the non-expansiveness of the projectors $\Pi_{\mathcal{S}(\mathbf{x}_1)}(\cdot)$ and $\Pi_{\mathcal{S}(\mathbf{x}_2)}(\cdot)$

Recall from (14) that $\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\| \leq M_{\mathcal{S}} \|\mathbf{x}_1 - \mathbf{x}_2\|$, and note that (17) further implies

$$\|\hat{\mathbf{y}} - \mathbf{y}\| \leq 2\theta(1 - \theta) M_{\mathcal{S}} \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \frac{1}{2} M_{\mathcal{S}} \|\mathbf{x}_1 - \mathbf{x}_2\|. \quad (27)$$

Combining (26) and (27) yields

$$\|\mathbf{y}_1 - \mathbf{y}_2\| \leq 2M_{\mathcal{S}} \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

which proves (25) with $L_{\mathcal{S}} \geq 4M_{\mathcal{S}}^2$.

Finally, to ensure (24) it suffices to choose

$$L_{\mathcal{S}} \geq \max\{2(L_0 + L_1), 4M_{\mathcal{S}}^2\}.$$

Recalling $M_{\mathcal{S}} = L_f \tau$, $L_0 = \frac{H_f \tau}{2}(1 + M_{\mathcal{S}}^2)$, and $L_1 = \frac{H_f \tau}{2}(1 + 17M_{\mathcal{S}}^2)$, we obtain

$$L_{\mathcal{S}} = \max\left\{2H_f \tau(1 + 9L_f^2 \tau^2), 4L_f^2 \tau^2\right\}.$$

This completes the proof of Theorem 2.

C.4 Proof of Remark 3

In the proof of Theorem 2, Assumption 1 is only used to guarantee Lemma 1 and the estimations on the involved points. Notice that to ensure Lemma 1, it suffices for Assumption 1 to hold on the set $\mathcal{D} \times \text{Conv}(\bigcup_{\mathbf{x} \in \mathcal{D}} \mathcal{S}(\mathbf{x})) \subseteq \mathcal{D} \times \mathcal{Y}$. We only need to check that the involved points belong to the set $\mathcal{D} \times \mathcal{Y}$ to prove Remark 3.

As the domain of $\mathcal{S}$ is $\mathcal{D}$, to verify the set smoothness of $\mathcal{S}$, we can choose $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$. Then, we have $\mathbf{x}^\theta \in \mathcal{D}$ due to the convexity of $\mathcal{D}$. Hence, it suffices to check that $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \bar{\mathbf{y}}^\theta, \mathbf{y}_1, \mathbf{y}_2, \hat{\mathbf{y}}, \hat{\mathbf{y}}_1$, and $\hat{\mathbf{y}}_2$ belong to the convex set $\mathcal{Y}$.

First, recall that $\bar{\mathbf{y}}_1, \mathbf{y}_1 \in \mathcal{S}(\mathbf{x}_1)$, $\bar{\mathbf{y}}_2, \mathbf{y}_2 \in \mathcal{S}(\mathbf{x}_2)$, $\hat{\mathbf{y}} \in \mathcal{S}(\mathbf{x}^\theta)$, and $\bar{\mathbf{y}}^\theta \in \text{Conv}(\mathcal{S}(\mathbf{x}_1) \cup \mathcal{S}(\mathbf{x}_2)) \subseteq \text{Conv}(\bigcup_{\mathbf{x} \in \mathcal{D}} \mathcal{S}(\mathbf{x}))$ due to their definitions. We see that they belong to the set $\mathcal{Y}$. We then focus on $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$. Recall that $\hat{\mathbf{y}}_i = \bar{\mathbf{y}}_i + (\mathbf{y} - \hat{\mathbf{y}})$ for $i = 1, 2$. We see that for $i = 1, 2$,

$$\text{dist}(\hat{\mathbf{y}}_i, \mathcal{S}(\mathbf{x}_i)) \leq \|\hat{\mathbf{y}}_i - \bar{\mathbf{y}}_i\| = \|\mathbf{y} - \hat{\mathbf{y}}\|.$$

By (27) and $\|\mathbf{x}_1 - \mathbf{x}_2\| \leq \text{diam}(\mathcal{D})$, for $i = 1, 2$, we further have

$$\text{dist}(\hat{\mathbf{y}}_i, \mathcal{S}(\mathbf{x}_i)) \leq \frac{1}{2} M_S \cdot \text{diam}(\mathcal{D}).$$

This implies that $\hat{\mathbf{y}}_i \in \mathcal{S}(\mathbf{x}_i) + \frac{1}{2} M_S \cdot \text{diam}(\mathcal{D}) \cdot \mathbb{B}(\mathbf{0}, 1) \subseteq \mathcal{Y}$ for $i = 1, 2$. We complete the proof.

D Proof of Theorem 4

We first introduce some background on zeroth-order methods before the formal proof. Let $\mathbb{P}$ denote the uniform distribution on the unit sphere in $\mathbb{R}^m$. Given a function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and a radius $\varepsilon > 0$, we define the randomized smooth approximation g by $g^\varepsilon(\mathbf{x}) := \mathbb{E}_{\mathbf{u} \sim \mathbb{P}}[g(\mathbf{x} + \varepsilon \mathbf{u})]$. We have the following properties for g^ε .

Lemma 6 (Basic Properties of Randomized Smoothing). *The following hold:*

- (i) *If g is M_g -Lipschitz continuous, then g^ε is differentiable, M_g -Lipschitz continuous, and satisfies*

$$|g(\mathbf{x}) - g^\varepsilon(\mathbf{x})| \leq \varepsilon M_g. \quad (28)$$

- (ii) *If g is r -weakly convex (resp. concave), then g^ε is r -weakly convex (resp. concave).*

Proof. (i) See [34, Proposition 2.3].

(ii) If g is r -weakly convex, then by the same arguments of Nazari et al. [41, Lemma 16], we have the r -weak convexity of g^ε . When g is r -weakly concave, we note that $-g$ is r -weakly convex, and hence $(-g)^\varepsilon$ is r -weakly convex. Due to the simple fact $(-g)^\varepsilon = -g^\varepsilon$, we have the r -weak convexity of $-g^\varepsilon$, i.e., the r -weak concavity of g^ε . $\square$

As the approximation for a subdifferential ∂g via randomized smoothing can be inexact, we need the following ε -subdifferential.

Definition 4. *Consider a convex function $g : \mathbb{R}^m \rightarrow \mathbb{R}$ and a scalar $\nu \geq 0$. We define the ν -subdifferential of g at $\mathbf{x} \in \mathbb{R}^m$ by*

$$\partial_\nu g(\mathbf{x}) = \{ \mathbf{s} \in \mathbb{R}^m : g(\mathbf{z}) \geq g(\mathbf{x}) + \mathbf{s}^T(\mathbf{z} - \mathbf{x}) - \nu, \forall \mathbf{z} \in \mathbb{R}^m \}.$$

We then develop the convergence rates of IZOM for optimistic and pessimistic BLO, respectively. We remark that our analysis remains valid when the hyper-objective functions φ_o (resp. φ_p) are replaced with general M -Lipschitz continuous, ρ -weakly concave (resp. convex) functions.

D.1 Optimistic Case

To begin, we record a celebrated proposition on subdifferential transportation.

Proposition 1. (cf. [2, Theorem 5.5] and [43, Theorem 2]) *Let $g : \mathbb{R}^m \rightarrow \mathbb{R}$ be a proper lower semicontinuous convex function. Suppose that $\nu \geq 0$ and $G \in \partial_\nu g(\mathbf{x})$. Then, for each $r > 0$, there is a unique vector $\mathbf{v} \in \mathbb{R}^m$ such that*

$$G - \frac{1}{r} \mathbf{v} \in \partial g(\mathbf{x} + r\mathbf{v}), \quad \|\mathbf{v}\| \leq \sqrt{\nu}.$$

Proposition 1 plays an important role in relating an ϵ -subdifferential to the Clarke subdifferential at a near point, leading to the following lemma.

Lemma 7. *Let $g : \mathbb{R}^m \rightarrow \mathbb{R}$ be an M_g -Lipschitz continuous and ρ -weakly concave function. Let g^ϵ be the randomized approximation of g with radius $\epsilon > 0$ and $\nu = 2\epsilon M_g$. Then, for all $\mathbf{x} \in \mathbb{R}^m$, we have*

$$\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \sqrt{\nu})} \partial g(\mathbf{z})\right) \leq \|\nabla g^\epsilon(\mathbf{x})\| + (\rho + 1)\sqrt{\nu}. \quad (29)$$

Proof of Lemma 7. By Lemma 6 (ii), g^ϵ is ρ -weakly concave, i.e., the function $\mathbf{x} \mapsto \frac{\rho}{2}\|\mathbf{x}\|^2 - g^\epsilon(\mathbf{x})$ is convex. Then, we know that for all $\mathbf{z}, \mathbf{x} \in \mathbb{R}^m$,

$$\frac{\rho}{2}\|\mathbf{z}\|^2 - g^\epsilon(\mathbf{z}) \geq \frac{\rho}{2}\|\mathbf{x}\|^2 - g^\epsilon(\mathbf{x}) + (\rho\mathbf{x} - \nabla g^\epsilon(\mathbf{x}))^T(\mathbf{z} - \mathbf{x}).$$

This, together with Lemma 6 (i), implies

$$\frac{\rho}{2}\|\mathbf{z}\|^2 - g(\mathbf{z}) \geq \frac{\rho}{2}\|\mathbf{x}\|^2 - g(\mathbf{x}) + (\rho\mathbf{x} - \nabla g^\epsilon(\mathbf{x}))^T(\mathbf{z} - \mathbf{x}) - 2\epsilon M_g,$$

which is equivalent to

$$\rho\mathbf{x} - \nabla g^\epsilon(\mathbf{x}) \in \partial_\nu\left(\frac{\rho}{2}\|\mathbf{x}\|^2 - g(\mathbf{x})\right) \quad \text{with} \quad \nu = 2\epsilon M_g.$$

For simplicity, we define $\bar{g} : \mathbf{x} \mapsto \frac{\rho}{2}\|\mathbf{x}\|^2 - g(\mathbf{x})$. Clearly, $\bar{g}$ is convex due to the ρ -weak concavity of g . Applying Proposition 1, we see that there exists $\mathbf{v} \in \mathbb{R}^m$ with $\|\mathbf{v}\| \leq \sqrt{\nu}$ such that

$$\rho\mathbf{x} - \nabla g^\epsilon(\mathbf{x}) - \mathbf{v} \in \partial \bar{g}(\mathbf{x} + \mathbf{v}). \quad (30)$$

Recall that $-g$ is ρ -weakly convex, and thus is regular according to [49, Proposition 4.5]. Then, we have $\partial \bar{g}(\mathbf{z}) = \rho\mathbf{z} + \partial(-g)(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^m$ by [10, Corollary 3 of Proposition 2.3.3]. On the other hand, we have $\partial(-g) = -\partial g$ by [10, Proposition 2.3.1] and Lipschitz continuity of g . Hence, we see that

$$\partial \bar{g}(\mathbf{z}) = \rho\mathbf{z} - \partial g(\mathbf{z}), \quad \forall \mathbf{z} \in \mathbb{R}^m.$$

In particular, we have

$$\partial \bar{g}(\mathbf{x} + \mathbf{v}) = \rho(\mathbf{x} + \mathbf{v}) - \partial g(\mathbf{x} + \mathbf{v}).$$

This, together with (30), implies

$$\nabla g^\epsilon(\mathbf{x}) + (\rho + 1)\mathbf{v} \in \partial g(\mathbf{x} + \mathbf{v}).$$

It follows that

$$\text{dist}(\mathbf{0}, \partial g(\mathbf{x} + \mathbf{v})) \leq \|\nabla g^\epsilon(\mathbf{x}) + (\rho + 1)\mathbf{v}\| \leq \|\nabla g^\epsilon(\mathbf{x})\| + (\rho + 1)\|\mathbf{v}\|.$$

Observe that $\|\mathbf{v}\| \leq \sqrt{\nu}$ and

$$\text{dist}(\mathbf{0}, \partial g(\mathbf{x} + \mathbf{v})) \geq \text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \|\mathbf{v}\|)} \partial g(\mathbf{z})\right) \geq \text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \sqrt{\nu})} \partial g(\mathbf{z})\right).$$

We obtain the desired inequality

$$\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}, \sqrt{\nu})} \partial g(\mathbf{z})\right) \leq \|\nabla g^\epsilon(\mathbf{x})\| + (\rho + 1)\sqrt{\nu}.$$

□

Now, we are ready to prove the convergence rate for IZOM. To begin, we define the subdifferential approximation function G by

$$G(\mathbf{x}_t) = \frac{m}{2\epsilon}(\varphi_o(\mathbf{x}_t + \epsilon \mathbf{u}_t) - \varphi_o(\mathbf{x}_t - \epsilon \mathbf{u}_t))\mathbf{u}_t.$$

By [34, Lemma D.1], it holds that

$$\mathbb{E}[G(\mathbf{x}_t)|\mathbf{x}_t] = \nabla \varphi_o(\mathbf{x}_t); \quad \mathbb{E}[\|G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq 16\sqrt{2\pi}mM_\varphi^2. \quad (31)$$

Since $|\tilde{\varphi}(\mathbf{x}_t) - \varphi_o(\mathbf{x}_t)| \leq w$ by the subroutine $\mathcal{A}$, we have $\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\| \leq \frac{mw}{\epsilon}$. This, together with the simple fact $\|\tilde{G}(\mathbf{x}_t)\|^2 \leq 2\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2 + 2\|G(\mathbf{x}_t)\|^2$ due to the Cauchy inequality, implies

$$\mathbb{E}[\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq \left(\frac{mw}{\epsilon}\right)^2; \quad \mathbb{E}[\|\tilde{G}(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq 32\sqrt{2\pi}mM_\varphi^2 + 2\left(\frac{mw}{\epsilon}\right)^2. \quad (32)$$

Next, we combine the update of IZOM and ρ -weak concavity of φ_o^ε to develop a sufficient decrease property for the t -th iteration. Using Fact 1 and the update $\mathbf{x}_{t+1} - \mathbf{x}_t = -\eta\tilde{G}(\mathbf{x}_t)$ of Algorithm 1, we obtain the following estimate:

$$\begin{aligned}
& \varphi_o^\varepsilon(\mathbf{x}_{t+1}) \\
& \leq \varphi_o^\varepsilon(\mathbf{x}_t) + \nabla\varphi_o^\varepsilon(\mathbf{x}_t)^T(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{\rho}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
& = \varphi_o^\varepsilon(\mathbf{x}_t) - \eta\nabla\varphi_o^\varepsilon(\mathbf{x}_t)^T\tilde{G}(\mathbf{x}_t) + \frac{\rho}{2}\eta^2\|\tilde{G}(\mathbf{x}_t)\|^2 \\
& = \varphi_o^\varepsilon(\mathbf{x}_t) - \eta\nabla\varphi_o^\varepsilon(\mathbf{x}_t)^TG(\mathbf{x}_t) - \eta\nabla\varphi_o^\varepsilon(\mathbf{x}_t)^T(\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)) + \frac{\rho}{2}\eta^2\|\tilde{G}(\mathbf{x}_t)\|^2 \\
& \leq \varphi_o^\varepsilon(\mathbf{x}_t) - \eta\nabla\varphi_o^\varepsilon(\mathbf{x}_t)^TG(\mathbf{x}_t) + \frac{\eta}{2}\|\nabla\varphi_o^\varepsilon(\mathbf{x}_t)\|^2 + \frac{\eta}{2}\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2 + \frac{\rho}{2}\eta^2\|\tilde{G}(\mathbf{x}_t)\|^2.
\end{aligned}$$

where the last inequality is due to the Cauchy inequality.

We take expectation conditioning on $\mathbf{x}_t$ for the above inequality. Recall that $\mathbb{E}[G(\mathbf{x}_t)|\mathbf{x}_t] = \nabla\varphi_o^\varepsilon(\mathbf{x}_t)$ by (31). We see that

$$\begin{aligned}
\mathbb{E}[\varphi_o^\varepsilon(\mathbf{x}_{t+1})|\mathbf{x}_t] & \leq \varphi_o^\varepsilon(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla\varphi_o^\varepsilon(\mathbf{x}_t)\|^2 + \frac{\eta}{2}\mathbb{E}[\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \\
& \quad + \frac{\rho\eta^2}{2}\mathbb{E}[\|\tilde{G}(\mathbf{x}_t)\|^2|\mathbf{x}_t].
\end{aligned}$$

Apply Lemma 7 to φ_o and use the Cauchy inequality. We obtain

$$\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}_t, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2 \leq 2\|\nabla\varphi_o^\varepsilon(\mathbf{x}_t)\|^2 + 2(\rho+1)^2\varepsilon.$$

Combining the above two inequalities with (32), we have

$$\begin{aligned}
\mathbb{E}[\varphi_o^\varepsilon(\mathbf{x}_{t+1})|\mathbf{x}_t] & \leq \varphi_o^\varepsilon(\mathbf{x}_t) - \frac{\eta}{4}\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}_t, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2 + \frac{\eta}{2}(\rho+1)^2\varepsilon \\
& \quad + \frac{\eta}{2}\left(\frac{mw}{\varepsilon}\right)^2 + \frac{\rho}{2}\eta^2\left(32\sqrt{2\pi}mM_\varphi^2 + 2\left(\frac{mw}{\varepsilon}\right)^2\right).
\end{aligned}$$

Recall that $\nu = 2\varepsilon M_\varphi$, $\eta = \Theta(\frac{1}{\sqrt{mT}})$, $\varepsilon = \mathcal{O}(\frac{1}{\sqrt{T}})$, and $w = \mathcal{O}(\frac{1}{m^{\frac{3}{4}}T^{\frac{3}{4}}})$. Ignoring some scalars, we further have

$$\mathbb{E}[\varphi_o^\varepsilon(\mathbf{x}_{t+1})|\mathbf{x}_t] \leq \varphi_o^\varepsilon(\mathbf{x}_t) - \frac{\eta}{4}\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}_t, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2 + \mathcal{O}\left(\frac{1}{T}\right). \quad (33)$$

Summing (33) over $t = 0, 1, \dots, T-1$ and taking full expectation, we obtain

$$\mathbb{E}[\varphi_o^\varepsilon(\mathbf{x}_T)] \leq \varphi_o^\varepsilon(\mathbf{x}_0) - \frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\left[\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}_t, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2\right] + \mathcal{O}(1).$$

Note that the definition of $\bar{\mathbf{x}}$ yields

$$\mathbb{E}\left[\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\bar{\mathbf{x}}, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\mathbf{x}_t, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2\right].$$

On the other hand, by Lemma 6 (i) and M_φ -Lipschitz continuity of φ_o ,

$$\varphi_o^\varepsilon(\mathbf{x}_0) - \mathbb{E}[\varphi_o^\varepsilon(\mathbf{x}_T)] \leq 2\varepsilon M_\varphi + \varphi_o(\mathbf{x}_0) - \min_{\mathbf{x}} \varphi_o(\mathbf{x}) = \Delta_o.$$

Putting all the things together, we have

$$\mathbb{E}\left[\text{dist}\left(\mathbf{0}, \bigcup_{\mathbf{z} \in \mathbb{B}(\bar{\mathbf{x}}, \sqrt{\nu})} \partial\varphi_o(\mathbf{z})\right)^2\right] = \mathcal{O}\left(\frac{\Delta_o + 1}{\eta T}\right) = \mathcal{O}\left(\frac{\sqrt{m}(\Delta_o + 1)}{\sqrt{T}}\right),$$

where $\sqrt{\nu} = \sqrt{2\varepsilon M_\varphi} = \mathcal{O}(T^{-\frac{1}{4}})$.

D.2 Pessimistic Case

In this section, we let $G(\mathbf{x}_t) = \frac{m}{2\varepsilon}(\varphi_p(\mathbf{x}_t + \varepsilon \mathbf{u}_t) - \varphi_p(\mathbf{x}_t - \varepsilon \mathbf{u}_t))\mathbf{u}_t$ and use $\hat{\mathbf{x}}$ to denote $\text{prox}_{\gamma, \varphi_p}(\mathbf{x})$ for simplicity. Similar to the arguments on (31), (32), we have the following due to [34, Lemma D.1] and $|\hat{\varphi}(\mathbf{x}_t) - \varphi_p(\mathbf{x}_t)| \leq w$ given by Algorithm 1:

$$\mathbb{E}[G(\mathbf{x}_t)|\mathbf{x}_t] = \nabla \varphi_p^\varepsilon(\mathbf{x}_t); \quad \mathbb{E}[\|G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq 16\sqrt{2\pi}mM_\varphi^2. \quad (34)$$

$$\mathbb{E}[\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq \left(\frac{mw}{\varepsilon}\right)^2; \quad \mathbb{E}[\|\tilde{G}(\mathbf{x}_t)\|^2|\mathbf{x}_t] \leq 32\sqrt{2\pi}mM_\varphi^2 + 2\left(\frac{mw}{\varepsilon}\right)^2. \quad (35)$$

Invoking the methodology of Davis and Drusvyatskiy [12], we first estimate $\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|$ for $t = 0, 1, \dots, T-1$. To begin, the update of Algorithm 1 and direct computation give the following estimate:

$$\begin{aligned} & \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 \\ &= \|\hat{\mathbf{x}}_t - \mathbf{x}_t + \eta \tilde{G}(\mathbf{x}_t)\|^2 \\ &= \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \eta^2 \|\tilde{G}(\mathbf{x}_t)\|^2 + 2\eta \tilde{G}(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\ &= \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \eta^2 \|\tilde{G}(\mathbf{x}_t)\|^2 + 2\eta G(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) + 2\eta (\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t))^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\ &\leq \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \eta^2 \|\tilde{G}(\mathbf{x}_t)\|^2 + 2\eta G(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) + \eta \|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2 + \eta \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2. \end{aligned} \quad (36)$$

Taking expectation in (36) and using (34), (35), we obtain the following inequality:

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2|\mathbf{x}_t] \\ &\leq (1+\eta)\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\eta \nabla \varphi_p^\varepsilon(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) + \eta \mathbb{E}[\|\tilde{G}(\mathbf{x}_t) - G(\mathbf{x}_t)\|^2|\mathbf{x}_t] \\ &\quad + \eta^2 \mathbb{E}[\|\tilde{G}(\mathbf{x}_t)\|^2|\mathbf{x}_t] \\ &\leq (1+\eta)\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\eta \nabla \varphi_p^\varepsilon(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) + \eta \left(\frac{mw}{\varepsilon}\right)^2 + 32\sqrt{2\pi}mM_\varphi^2\eta^2 + 2\left(\frac{mw}{\varepsilon}\right)^2\eta^2. \end{aligned} \quad (37)$$

We then turn to estimate $\varphi_{p,\gamma}(\mathbf{x}_{t+1})$. By the definition of $\varphi_{p,\gamma}$, we have

$$\varphi_{p,\gamma}(\mathbf{x}_{t+1}) \leq \varphi_p(\hat{\mathbf{x}}_t) + \frac{1}{2\gamma} \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2.$$

This, together with (37), implies

$$\begin{aligned} & \mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_{t+1})|\mathbf{x}_t] - \varphi_p(\hat{\mathbf{x}}_t) \\ &\leq \frac{1+\eta}{2\gamma} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \frac{\eta}{\gamma} \nabla \varphi_p^\varepsilon(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) + \frac{\eta}{2\gamma} \left(\frac{mw}{\varepsilon}\right)^2 + 16\sqrt{2\pi}mM_\varphi^2\frac{\eta^2}{\gamma} + \left(\frac{mw}{\varepsilon}\right)^2\frac{\eta^2}{\gamma}. \end{aligned}$$

Recall that φ_p^ε is ρ -weakly convex by Theorem 3 and Lemma 6 (ii). Then, using Lemma 3 (iii), we have

$$\begin{aligned} \nabla \varphi_p^\varepsilon(\mathbf{x}_t)^T (\hat{\mathbf{x}}_t - \mathbf{x}_t) &\leq \varphi_p^\varepsilon(\hat{\mathbf{x}}_t) - \varphi_p^\varepsilon(\mathbf{x}_t) + \frac{\rho}{2} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\ &\leq \varphi_p(\hat{\mathbf{x}}_t) - \varphi_p(\mathbf{x}_t) + \frac{\rho}{2} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\varepsilon M_\varphi, \end{aligned}$$

where the second inequality is due to Lemma 6 (i).

Combining the above two estimates gives

$$\begin{aligned} & \mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_{t+1})|\mathbf{x}_t] \\ &\leq \varphi_p(\hat{\mathbf{x}}_t) + \frac{1}{2\gamma} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \frac{\eta}{\gamma} \left(\varphi_p(\hat{\mathbf{x}}_t) - \varphi_p(\mathbf{x}_t) + \frac{\rho+1}{2} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \right) + \frac{2\eta}{\gamma} \varepsilon M_\varphi \\ &\quad + \frac{\eta}{2\gamma} \left(\frac{mw}{\varepsilon}\right)^2 + 16\sqrt{2\pi}mM_\varphi^2\frac{\eta^2}{\gamma} + \left(\frac{mw}{\varepsilon}\right)^2\frac{\eta^2}{\gamma} \\ &= \varphi_{p,\gamma}(\mathbf{x}_t) + \frac{\eta}{\gamma} \left(\varphi_{p,\gamma}(\mathbf{x}_t) - \varphi_p(\mathbf{x}_t) + \frac{(\rho+1)\gamma-1}{2\gamma} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \right) + \frac{2\eta}{\gamma} \varepsilon M_\varphi \\ &\quad + \frac{\eta}{2\gamma} \left(\frac{mw}{\varepsilon}\right)^2 + 16\sqrt{2\pi}mM_\varphi^2\frac{\eta^2}{\gamma} + \left(\frac{mw}{\varepsilon}\right)^2\frac{\eta^2}{\gamma}, \end{aligned}$$

where the equation uses $\varphi_{p,\gamma}(\mathbf{x}_t) = \varphi_p(\hat{\mathbf{x}}_t) + \frac{1}{2\gamma}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2$. Recall that by Lemma 4, it holds that $\mathbf{x}_t - \hat{\mathbf{x}}_t = \gamma \nabla \varphi_{p,\gamma}(\mathbf{x}_t)$ and

$$\varphi_{p,\gamma}(\mathbf{x}_t) - \varphi_p(\mathbf{x}_t) \leq -\frac{1-\gamma\rho}{2\gamma}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \leq -\frac{1-\gamma(\rho+1)}{2\gamma}\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2.$$

We further have

$$\begin{aligned} \mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq \varphi_{p,\gamma}(\mathbf{x}_t) - \eta(1-\gamma(\rho+1))\|\nabla \varphi_{p,\gamma}(\mathbf{x}_t)\|^2 + \frac{2\eta}{\gamma}\varepsilon M_\varphi \\ &\quad + \frac{\eta}{2\gamma}\left(\frac{mw}{\varepsilon}\right)^2 + 16\sqrt{2\pi}mM_\varphi^2\frac{\eta^2}{\gamma} + \left(\frac{mw}{\varepsilon}\right)^2\frac{\eta^2}{\gamma}. \end{aligned}$$

Recall that $\eta = \Theta(\frac{1}{\sqrt{mT}})$, $\varepsilon = \mathcal{O}(\frac{1}{\sqrt{T}})$, $w = \mathcal{O}(\frac{1}{m^{\frac{3}{4}}T^{\frac{3}{4}}})$, and $\gamma \in (0, 1/(\rho+1))$. Neglecting some scalars, it follows that

$$\mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_{t+1})|\mathbf{x}_t] \leq \varphi_{p,\gamma}(\mathbf{x}_t) - \eta(1-\gamma(\rho+1))\|\nabla \varphi_{p,\gamma}(\mathbf{x}_t)\|^2 + \mathcal{O}\left(\frac{1}{T}\right). \quad (38)$$

Summing (38) over $t = 0, 1, \dots, T-1$ and taking full expectation, we have

$$\mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_T)] \leq \varphi_{p,\gamma}(\mathbf{x}_0) - \eta(1-\gamma(\rho+1))\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla \varphi_{p,\gamma}(\mathbf{x}_t)\|^2] + \mathcal{O}(1).$$

Note that the definition of $\bar{\mathbf{x}}$ gives $\sum_{t=0}^{T-1}\|\nabla \varphi_{p,\gamma}(\mathbf{x}_t)\|^2 = T\mathbb{E}[\|\nabla \varphi_{p,\gamma}(\bar{\mathbf{x}})\|^2]$. Also, notice that

$$\varphi_{p,\gamma}(\mathbf{x}_0) - \mathbb{E}[\varphi_{p,\gamma}(\mathbf{x}_T)] \leq \varphi_{p,\gamma}(\mathbf{x}_0) - \min_{\mathbf{x}} \varphi_{p,\gamma}(\mathbf{x}) = \Delta_p.$$

Putting all the things together, we obtain

$$\mathbb{E}[\|\nabla \varphi_{p,\gamma}(\bar{\mathbf{x}})\|^2] = \mathcal{O}\left(\frac{\Delta_p + 1}{\eta T}\right) = \mathcal{O}\left(\frac{\sqrt{m}(\Delta_p + 1)}{\sqrt{T}}\right).$$

This completes the proof.