On the hardness of deterministic second-order optimization of functions with Lipschitz gradients

Anonymous authors
Paper under double-blind review

Abstract

We show that no deterministic zero-respecting algorithm (resp., (general) deterministic algorithm) can compute Goldstein approximate second-order stationary points of functions with Lipschitz continuous gradients within a finite number of (resp., no more than $n-3$ with n being the input dimension) second-order oracle calls. This, among other consequences, shows that deterministic second-order weakly convex optimization is intractable.

1 Introduction

Differentiable functions with Lipschitz continuous gradients, commonly referred to as $C^{1,1}$-functions, are prevalent across multiple disciplines; e.g., machine learning and statistics. As a result, understanding the optimization properties of $C^{1,1}$-functions is of considerable interest. Owing to the substantial contributions of researchers over recent decades, the first-order analytical and algorithmic properties of $C^{1,1}$-optimization have been thoroughly understood; see, e.g., (Nesterov, 2018, Section 2), (Carmon et al., 2020), and the references therein. For instance, it is now folkloric that an ε-approximate first-order stationary point of a (lower-bounded) $C^{1,1}$-function can be found within at most $\mathcal{O}(\varepsilon^{-2})$ steps of gradient descent; see, e.g., (Nesterov, 2018, Section 1.2.3).

That said, as the gradients of $C^{1,1}$-functions can potentially be nonsmooth (e.g., the Huber loss (Huber, 1964), squared hinge loss (Hsieh et al., 2008; Chang et al., 2008), and exponential linear unit (Clevert et al., 2016)), the behavior of second-order methods for optimizing $C^{1,1}$-functions can be rather subtle. While substantial research progress has been made along this direction (see, e.g., (Rockafellar & Wets, 2009, Section 13), the recent book by Mordukhovich (2024), and the references therein), there are still several aspects that have received only limited attention; e.g., the (oracle) complexity of finding approximate second-order stationary points of $C^{1,1}$-functions. Conceivably, as the task of finding approximate first-order stationary points of Lipschitz functions is challenging (see, e.g., (Daniilidis & Drusvyatskiy, 2020; Zhang et al., 2020; Kornowski & Shamir, 2021; Tian & So, 2021; Kornowski & Shamir, 2022; Tian et al., 2022; Jordan et al., 2023; Tian & So, 2024) for a series of recent advances shedding light on this point, and (Tian & So, 2026) for an overview), so should the task of finding approximate second-order stationary points of $C^{1,1}$-functions.

In this paper, our goal is to rigorously justify this intuition. Towards that end, we first consider the Goldstein second-order subdifferential (Definition 2.5) and its associated stationarity concept (Definition 2.6), which arise naturally as second-order analogues of the first-order Goldstein framework (Goldstein, 1977), motivated by the goal of identifying an approximate stationarity concept for which complexity guarantees can be established, as in the first-order setting (Zhang et al., 2020). Then, we show that for any number of iterations $m \geq 1$, any dimension $n \geq 2$ (resp., $n \geq m+3$), and any second-order deterministic zero-respecting algorithm (resp., second-order (general) deterministic algorithm) $\boldsymbol{\mathcal{A}}$ that seeks solutions over $\mathbb{R}^n$, there exists some $C^{1,1}$-function $f : \mathbb{R}^n \to \mathbb{R}$ (with fixed Lipschitz constant and initial optimality gap) such that the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ generated by $\boldsymbol{\mathcal{A}}$ when applied to f are not Goldstein approximate second-order stationary up to some fixed tolerances (Theorem 3.1 and Corollary 3.2). Concretely, the class of second-order deterministic zero-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

respecting algorithms captures most of the standard nonsmooth Newton methods built on different second-order subdifferential constructions (Kummer, 1988; Pang, 1990; Qi, 1993; Qi & Sun, 1993; Hoheisel et al., 2012; Mordukhovich & Sarabi, 2021; Khanh et al., 2023), whereas the class of second-order (general) deterministic algorithms covers virtually all common deterministic second-order methods, including the nonconvex cutting-plane method (Fuduli et al., 2004) and grid search (Nesterov, 2018, Section 1.1.3). These hardness results well align with the fact that, without additional assumptions, no complexity upper bound is known for the *deterministic* computation of approximate second-order stationary points of $C^{1,1}$-functions. This should be contrasted with the *randomized* computation of such points, for which Lei et al. (2026) and Guan & So (2026) have recently developed a zeroth-order and a first-order algorithm, respectively, and established their complexity guarantees. In particular, a key takeaway is that finding approximate second-order stationary points of weakly convex functions in a deterministic manner is far from being tractable, as $C^{1,1}$-functions are automatically weakly convex; see, e.g., (Vial, 1983, Proposition 4.12).[1]

In the process of establishing the said lower bounds, which delineate fundamental limits for subsequent algorithmic developments, we develop new machinery that contributes to second-order variational analysis. This includes scalarization formulae for the Clarke and Goldstein second-order subdifferentials (Proposition 2.4 and Corollary 2.7; see, e.g., (Mordukhovich, 2024, Proposition 1.50) for related results), and a second-order Goldstein subdifferential chain rule (Corollary 2.9).

The rest of the paper is organized as follows. We first lay out the notation and preliminaries in Section 2. Then, we move on to presenting the main results in Section 3, with proofs deferred to Section 4 and the appendices. Finally, we conclude this paper in Section 5.

2 Notation and preliminaries

2.1 Notation

The notation in this paper adheres primarily to standard conventions. For any $\boldsymbol{x} \in \mathbb{R}^n$, its support is defined by $\operatorname{supp}(\boldsymbol{x}) := \{i \in [n] : x_i \neq 0\}$, where $[n] := \{1, \dots, n\}$. Besides, for any $\mathbb{J} \subseteq [n]$, we use $\boldsymbol{x}_{\mathbb{J}} \in \mathbb{R}^{|\mathbb{J}|}$ to denote the subvector of $\boldsymbol{x}$ consisting of the entries indexed by $\mathbb{J}$. We use $\mathbb{B}^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$ and $\mathbb{S}^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$ to denote the unit ball and the unit sphere in $\mathbb{R}^n$, respectively. For any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we use $\|\boldsymbol{A}\|_\sigma$ to denote its spectral norm. For any $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^m$, we denote by $\operatorname{lip}\boldsymbol{F}(\boldsymbol{x})$ its Lipschitz modulus at $\boldsymbol{x} \in \mathbb{R}^n$; see, e.g., (Rockafellar & Wets, 2009, Definition 9.1(b)). For any $f : \mathbb{R}^n \to \mathbb{R}$, we say that it is C^k for some $k = 1, 2, \dots$ if it is k-times continuously differentiable. For any $\mathbb{X} \subseteq \mathbb{R}^n$, we denote by $\operatorname{conv}(\mathbb{X})$ its convex hull. For $i = 1, \dots, n$, we use $\boldsymbol{e}_i^n$ to denote the i-th standard basis vector in $\mathbb{R}^n$; the superscript n is often omitted when it is clear from the context. Unless stated otherwise, all limit superiors in this paper are taken in the sense of Painlevé–Kuratowski; see, e.g., (Mordukhovich, 2006, Equation 1.1) and (Rockafellar & Wets, 2009, Section 5.B). As an aside, in this paper, we do not distinguish between the maximum and supremum, nor between the minimum and infimum.

2.2 Generalized differentiation theory and stationarity concepts

To start, we introduce the (first-order) Clarke and Goldstein subdifferentials.

Definition 2.1 *Given a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $\boldsymbol{x} \in \mathbb{R}^n$:*

- *The Clarke subdifferential of f at $\boldsymbol{x}$ is defined as*

$$\partial_{\mathrm{C}} f(\boldsymbol{x}) := \operatorname{conv}\big(\limsup\nolimits_{\boldsymbol{y} \to \boldsymbol{x},\, \boldsymbol{y} \in \mathbb{D}} \{\nabla f(\boldsymbol{y})\}\big),$$

where $\mathbb{D}$ is the set on which f is differentiable; see, e.g., (Clarke, 1990, Theorem 2.5.1).

[1] It is noteworthy that weakly convex functions arise widely in machine learning and beyond; e.g., sparse dictionary learning (Davis & Drusvyatskiy, 2019, Example 2.4) and robust low-rank matrix recovery (Li et al., 2020b). For further examples, we refer interested readers to (Davis & Drusvyatskiy, 2019, Section 2.1), (Li et al., 2020a, Section IV), and the references therein.

- *Given a constant $\delta \geq 0$, the Goldstein δ-subdifferential of f at $\boldsymbol{x}$ is defined as*

$$\partial_\delta f(\boldsymbol{x}) := \operatorname{conv}\Big(\bigcup\nolimits_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n} \partial_{\mathrm{C}} f(\boldsymbol{y})\Big);$$

see (Goldstein, 1977, Section 2).

For a Lipschitz function f and a point $\boldsymbol{x}$, the subdifferentials $\partial_{\mathrm{C}} f(\boldsymbol{x})$ and $\partial_\delta f(\boldsymbol{x})$ are always nonempty, convex, and compact; see, e.g., (Clarke, 1990, Proposition 2.1.2(a)) and (Goldstein, 1977, Proposition 2.3). With the first-order constructions in place, we next introduce tools from second-order variational analysis. We begin with the Clarke generalized Hessian, defined analogously to $\partial_{\mathrm{C}} f(\boldsymbol{x})$.

Definition 2.2 *Given a $C^{1,1}$-function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $\boldsymbol{x} \in \mathbb{R}^n$:*

- *The Clarke generalized Hessian of f at $\boldsymbol{x}$ is defined as*

$$\partial^2_{\mathrm{C}} f(\boldsymbol{x}) := \operatorname{conv}\big(\limsup\nolimits_{\boldsymbol{y}\to\boldsymbol{x},\, \boldsymbol{y}\in\mathbb{D}} \{\nabla^2 f(\boldsymbol{y})\}\big) \subseteq \mathbb{R}^{n\times n}_{\mathrm{sym}},$$

where $\mathbb{D}$ is the set on which f is twice differentiable; see, e.g., (Clarke, 1990, Definition 2.6.1) and (Rockafellar & Wets, 2009, Theorem 13.52).

We note that similar to the first-order constructions, for a $C^{1,1}$-function f and a point $\boldsymbol{x}$, the generalized Hessian $\partial^2_{\mathrm{C}} f(\boldsymbol{x})$ is also nonempty, convex, and compact; see, e.g., (Clarke, 1990, Proposition 2.6.2(a)). Besides, in alignment with existing second-order subdifferential constructions (see, e.g., (Mordukhovich, 2024, Section 1.2)), in what follows, we shall identify the set $\partial^2_{\mathrm{C}} f(\boldsymbol{x})$ with the set-valued mapping $\big(\boldsymbol{w} \mapsto \{\boldsymbol{A}\boldsymbol{w} : \boldsymbol{A} \in \partial^2_{\mathrm{C}} f(\boldsymbol{x})\}\big) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

Very naturally, the construction of $\partial^2_{\mathrm{C}} f(\boldsymbol{x})$ gives rise to a second-order optimality condition paralleling the smooth case; i.e., vanishing gradient and positive semidefinite (PSD) Hessian.

Lemma 2.3 ((Hiriart-Urruty et al., 1984, Theorem 3.1)) *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$, and let $\boldsymbol{x} \in \mathbb{R}^n$ be a local minimizer of f. Then, we have*

$$\nabla f(\boldsymbol{x}) = \boldsymbol{0} \quad \text{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{z}\in\partial^2_{\mathrm{C}} f(\boldsymbol{x})(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{w}\rangle \geq 0. \tag{1}$$

As an aside, the following connection between $\partial_{\mathrm{C}} f(\boldsymbol{x})$ and $\partial^2_{\mathrm{C}} f(\boldsymbol{x})$ is worth noting.

Proposition 2.4 *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$. Then, we have*

$$\partial^2_{\mathrm{C}} f(\boldsymbol{x})(\boldsymbol{w}) = \partial_{\mathrm{C}} \langle \boldsymbol{w}, \nabla f\rangle(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n.$$

We note that this is nothing but a scalarization formula for $\partial^2_{\mathrm{C}} f(\boldsymbol{x})(\boldsymbol{w})$ paralleling, e.g., (Mordukhovich, 2024, Proposition 1.50) and (Rockafellar & Wets, 2009, Proposition 9.24(b)).

Proof. By (Rockafellar & Wets, 2009, Theorem 13.52) and the fact that linear maps commute with convex hulls (as can be seen directly from (Rockafellar, 1997, Theorem 3.4)),

$$\partial^2_{\mathrm{C}} f(\boldsymbol{x})(\boldsymbol{w}) = \operatorname{conv}\big(\partial^2 f(\boldsymbol{x})(\boldsymbol{w})\big) \quad \text{for all } \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n,$$

where ∂^2 denotes the Mordukhovich (or limiting) second-order subdifferential (Mordukhovich, 2024, Definition 1.46). As $\partial^2 f(\boldsymbol{x})(\boldsymbol{w}) = \partial\langle \boldsymbol{w}, \nabla f\rangle(\boldsymbol{x})$ (Mordukhovich, 2024, Proposition 1.50), where ∂ denotes the Mordukhovich (or limiting) subdifferential (Rockafellar & Wets, 2009, Definition 8.3(b)), and $\langle \boldsymbol{w}, \nabla f\rangle$ is locally Lipschitz (Rockafellar & Wets, 2009, Exercise 9.9), the desired result immediately follows from (Cui & Pang, 2021, Proposition 4.3.2(b)). □

We now turn to the central object: The Goldstein second-order δ-subdifferential.

Definition 2.5 (Goldstein second-order δ-subdifferential) *Given a constant $\delta \geq 0$, a $C^{1,1}$-function $f : \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \mathbb{R}^n$, the Goldstein second-order δ-subdifferential of f at $\boldsymbol{x}$ is the set-valued mapping $\partial^2_\delta f(\boldsymbol{x}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as*

$$\partial^2_\delta f(\boldsymbol{x})(\boldsymbol{w}) := \operatorname{conv}\Big(\bigcup\nolimits_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n} \partial^2_{\mathrm{C}} f(\boldsymbol{y})(\boldsymbol{w})\Big) \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^n.$$

In other words, $\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w})$ is the smallest convex set that contains all possible $\partial_{\mathrm{C}}^2 f(\boldsymbol{y})(\boldsymbol{w})$ from points $\boldsymbol{y}$ with distance at most δ from $\boldsymbol{x}$, in direct analogy with $\partial_\delta f(\boldsymbol{x})$; this explains why $\partial_\delta^2 f(\boldsymbol{x})$ is referred to as "Goldstein". Note that $\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w})$ can potentially encompass a broader set of second-order generalized derivatives than $\partial_{\mathrm{C}}^2 f(\boldsymbol{x})(\boldsymbol{w})$. Thus, the stationarity concept that $\partial_\delta^2 f(\boldsymbol{x})$ induces, which we introduce in Definition 2.6, is much weaker and thus potentially easier to compute than (1). Nonetheless, our main results (Theorem 3.1 and Corollary 3.2) show that it remains deterministically intractable in general. We also note that $\partial_\delta^2 f(\boldsymbol{x})$ is closely related to, yet distinct from, the second-order ε-jet introduced by Gebken; see (Gebken, 2022, Definition 1).

Definition 2.6 (Goldstein approximate second-order stationarity) *Given constants $\varepsilon, \eta, \delta \geq 0$, a $C^{1,1}$-function $f : \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \mathbb{R}^n$, we say that $\boldsymbol{x}$ is an $(\varepsilon, \eta, \delta)$-Goldstein approximate second-order stationary point of f if*

$$\|\nabla f(\boldsymbol{x})\| \leq \varepsilon \quad \text{and} \quad \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{z} \in \partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{w} \rangle \geq -\eta.$$

We remark that, while formulated differently, the Goldstein generalized Hessian of Lei et al. (2026, Definition 3.1) and of Guan & So (2026, Definition 2.4) induce exactly the same stationarity notion as Definition 2.6, as linear operators commute with convex hulls. Nevertheless, we retain our directional formulation of the Goldstein second-order subdifferential, as it facilitates the development of the second-order chain rule in Corollary 2.9.

As expected, a scalarization formula for $\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w})$ can be established.

Corollary 2.7 *Suppose that $\delta \geq 0$ and that $f : \mathbb{R}^n \to \mathbb{R}$ is $C^{1,1}$. Then, we have*

$$\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w}) = \partial_\delta \langle \boldsymbol{w}, \nabla f \rangle(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n.$$

In particular, Corollary 2.7 implies that $\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w})$ is also nonempty, convex, and compact.

Proof. Indeed, it directly follows by combining Definition 2.5 with Proposition 2.4 and Definition 2.1 that $\partial_\delta^2 f(\boldsymbol{x})(\boldsymbol{w}) = \operatorname{conv}\Big(\bigcup_{\boldsymbol{y} \in \boldsymbol{x} + \delta \mathbb{B}^n} \partial_{\mathrm{C}} \langle \boldsymbol{w}, \nabla f \rangle(\boldsymbol{y})\Big) = \partial_\delta \langle \boldsymbol{w}, \nabla f \rangle(\boldsymbol{x})$. □

In later developments, we will need a second-order Goldstein subdifferential chain rule. In order to derive this, let us first make explicit a (first-order) Goldstein subdifferential chain rule hidden in the proof of (Tian & So, 2024, Theorem 1). As the proof is a bit involved, it is deferred to Appendix A.

Lemma 2.8 *Let $\delta \geq 0$ be a constant, $f : \mathbb{R}^m \to \mathbb{R}$ be Lipschitz, and $\boldsymbol{U} \in \mathbb{R}^{n \times m}$ be such that $\boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} = \boldsymbol{I}$. Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ satisfies $g(\boldsymbol{x}) = f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Then,*

$$\partial_\delta g(\boldsymbol{x}) = \boldsymbol{U} \partial_\delta f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n.$$

We remark that Lemma 2.8 is slightly stronger than the chain rule implicit in the proof of (Tian & So, 2024, Theorem 1), as here $\boldsymbol{U}$ is only required to have orthonormal columns, rather than to be orthogonal. With Lemma 2.8 in place, we are now ready to derive the second-order chain rule.

Corollary 2.9 *Let $\delta \geq 0$ be a constant, $f : \mathbb{R}^m \to \mathbb{R}$ be $C^{1,1}$, and $\boldsymbol{U} \in \mathbb{R}^{n \times m}$ be such that $\boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} = \boldsymbol{I}$. Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ satisfies $g(\boldsymbol{x}) = f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Then,*

$$\partial_\delta^2 g(\boldsymbol{x})(\boldsymbol{w}) = \boldsymbol{U} \partial_\delta^2 f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x})(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n.$$

Proof. To start, we observe through the smooth chain rule that for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$,

$$\langle \boldsymbol{w}, \nabla g \rangle(\boldsymbol{x}) = \langle \boldsymbol{w}, \nabla g(\boldsymbol{x}) \rangle = \langle \boldsymbol{w}, \boldsymbol{U} \nabla f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}) \rangle = \langle \boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}, \nabla f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}) \rangle = \langle \boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}, \nabla f \rangle(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}).$$

Besides, as f is $C^{1,1}$ as assumed, we know that $\langle \boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}, \nabla f \rangle$ is Lipschitz continuous and g is $C^{1,1}$. It then follows by combining the above facts with Corollary 2.7 and Lemma 2.8 that

$$\partial_\delta^2 g(\boldsymbol{x})(\boldsymbol{w}) = \partial_\delta \langle \boldsymbol{w}, \nabla g \rangle(\boldsymbol{x}) = \boldsymbol{U} \partial_\delta \langle \boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}, \nabla f \rangle(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}) = \boldsymbol{U} \partial_\delta^2 f(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{x})(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{w}),$$

as desired. This completes the proof. □

We note in closing this subsection that second-order variational analysis is a rich and thriving subject that encompasses concepts, theories, and applications far beyond the preceding discussions. We refer interested readers to the recent book by Mordukhovich (2024).

2.3 Oracles, algorithm classes, and function classes

2.3.1 Oracles

In this paper, unless stated otherwise, we restrict attention to second-order oracles. We say that $\boldsymbol{\mathcal{O}} : \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^n \times 2^{\mathbb{R}^{n\times n}_{\text{sym}}}$ is the second-order oracle that corresponds to a $C^{1,1}$-function $f : \mathbb{R}^n \to \mathbb{R}$ if $\boldsymbol{\mathcal{O}}(\boldsymbol{x}) = \big(f(\boldsymbol{x}), \nabla f(\boldsymbol{x}), \partial^2_{\text{C}} f(\boldsymbol{x})\big)$ for all $\boldsymbol{x} \in \mathbb{R}^n$. With a slight abuse of notation, we write $\boldsymbol{\mathcal{O}}(\boldsymbol{x}) = \big(f(\boldsymbol{x}), \nabla f(\boldsymbol{x}), \nabla^2 f(\boldsymbol{x})\big)$ when f is C^2 near $\boldsymbol{x}$ (even if f may fail to be (globally) $C^{1,1}$), identifying the singleton set $\partial^2_{\text{C}} f(\boldsymbol{x})$ with its unique element $\nabla^2 f(\boldsymbol{x})$.

2.3.2 Algorithm classes

In this paper, an algorithm that seeks solutions over $\mathbb{R}^n$ (for a general oracle $\boldsymbol{\mathcal{O}} : \mathbb{R}^n \to \mathbb{O}$ where $\mathbb{O}$ is an arbitrary set) consists of a (possibly randomized) initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$ and a sequence of (possibly randomized) mappings $\boldsymbol{A}_k : (\mathbb{R}^n \times \mathbb{O})^{k+1} \to \mathbb{R}^n$ for $k = 0, 1, \ldots$. The sequence of iterates generated by the algorithm is defined recursively by

$$\boldsymbol{x}_{k+1} := \boldsymbol{A}_k\big(\boldsymbol{x}_0, \boldsymbol{\mathcal{O}}(\boldsymbol{x}_0), \ldots, \boldsymbol{x}_k, \boldsymbol{\mathcal{O}}(\boldsymbol{x}_k)\big) \quad \text{for } k = 0, 1, \ldots.$$

With a slight abuse of terminology, we will occasionally call $\boldsymbol{x}_0$ an iterate as well; i.e., $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ are collectively referred to as the iterates generated by the algorithm. We say that an algorithm is deterministic if its constituents $\boldsymbol{x}_0$ and $\boldsymbol{A}_0, \boldsymbol{A}_1, \ldots$ are so.

As we shall only consider functions that are C^2 near the queried points (see Section 2.3.3), in what follows, it suffices to consider algorithms that access Hessians rather than Clarke generalized Hessians. For this purpose, let us denote by

$$\mathbb{A}^n_{\text{det}}(m) := \big\{(\boldsymbol{x}_0, \boldsymbol{A}_0, \ldots, \boldsymbol{A}_{m-1}) : \boldsymbol{x}_0 \in \mathbb{R}^n \text{ and } \boldsymbol{A}_k : (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{n\times n}_{\text{sym}})^{k+1} \to \mathbb{R}^n\big\}$$

the set of all second-order deterministic algorithms that seek solutions over $\mathbb{R}^n$ with $m \geq 1$ iterations.

With the above preparations in place, we turn to recall the notion of second-order zero-respecting algorithms (Carmon et al., 2020, Equation 5), adapted to our setting. In brief, these algorithms can only update coordinates that have been revealed as relevant through prior oracle calls.

Definition 2.10 *We say that $\boldsymbol{\mathcal{A}} \in \mathbb{A}^n_{\text{det}}(m)$ is second-order zero-respecting if for every $f : \mathbb{R}^n \to \mathbb{R}$ that is C^2 near the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{m-1}$ generated by $\boldsymbol{\mathcal{A}}$ when applied to f,*

$$\operatorname{supp}(\boldsymbol{x}_k) \subseteq \Big(\bigcup\nolimits_{0\leq \ell < k} \operatorname{supp}\big(\nabla f(\boldsymbol{x}_\ell)\big)\Big) \cup \Big(\bigcup\nolimits_{0\leq \ell < k} \big\{i \in [n] : (\nabla^2 f(\boldsymbol{x}_\ell))_i \neq \boldsymbol{0}\big\}\Big) \tag{2}$$

for all $k = 0, \ldots, m$. The set of all such algorithms is denoted by $\mathbb{A}^n_{\text{det-zr}}(m)$.

By letting $k = 0$ in (2), we obtain $\boldsymbol{x}_0 = \boldsymbol{0}$; in other words, second-order deterministic zero-respecting algorithms are necessarily initialized at the origin.

We close this part with a discussion of the scope of the algorithm class $\mathbb{A}^n_{\text{det-zr}}(m) \subseteq \mathbb{A}^n_{\text{det}}(m)$. As the functions we consider are indistinguishable from C^2-functions near the queried points, different second-order subdifferential constructions all collapse to the Hessian at the queried points (see, e.g., (Mordukhovich, 2024, Proposition 1.74)), and methods for smooth optimization are also applicable. As a consequence, the class of second-order deterministic zero-respecting algorithms (i.e., $\mathbb{A}^n_{\text{det-zr}}(m)$) covers a wide array of existing smooth and nonsmooth methods. Concretely, this includes, e.g., the gradient method (Nesterov, 2018, Section 1.2.3), Nesterov's accelerated gradient method (Nesterov, 1983; Su et al., 2016)

(as long as the extrapolated points are also treated as part of the iterates), the conjugate gradient method (Nocedal & Wright, 2006, Algorithm 5.4) (as long as the step sizes are, e.g., determined *a priori*), Newton's method (Nesterov, 2018, Section 1.2.4) (as long as the Newton direction is taken as $-\nabla^2 f(\boldsymbol{x})^\dagger \nabla f(\boldsymbol{x})$ when $\nabla^2 f(\boldsymbol{x})$ is singular[2]), nonsmooth Newton methods based on Bouligand generalized Hessians (Qi, 1993), directional derivatives (Pang, 1990), graphical derivatives (Hoheisel et al., 2012; Mordukhovich & Sarabi, 2021), Mordukhovich coderivatives (Mordukhovich & Sarabi, 2021; Khanh et al., 2023), and Clarke generalized Hessians (Kummer, 1988; Qi & Sun, 1993) (as long as the same convention for choosing the Newton direction is adopted), the BFGS method (Nocedal & Wright, 2006, Algorithm 6.1) (as long as the initial inverse Hessian approximation is, e.g., diagonal, and the step sizes are, e.g., determined *a priori*), the cubic Newton method (Nesterov & Polyak, 2006) (as long as the (cubic) regularization weights are, e.g., determined *a priori*), the trust-region method (Nocedal & Wright, 2006, Algorithm 4.1) (as long as the trust-region step is taken as $-\nabla^2 f(\boldsymbol{x})^\dagger \nabla f(\boldsymbol{x})$ when the trust-region constraint is redundant, and the trial points are also treated as part of the iterates), and any deterministic second-order method that satisfies the so-called linear span assumption (Nesterov, 2018, Assumption 2.1.4), assuming that they are all initialized at the origin. In particular, this implies that the hardness result in Theorem 3.1 applies to all of these methods.

2.3.3 Function classes

In this paper, we focus exclusively on subclasses of $C^{1,1}$-functions that are bounded from below. To ease the exposition, we introduce for all $\boldsymbol{\mathcal{A}} \in \mathbb{A}^n_{\text{det}}(m)$ the shorthand

$$\mathbb{F}^n_{C^1}(L, G; \boldsymbol{\mathcal{A}}) := \left\{ f : \mathbb{R}^n \to \mathbb{R} : \begin{array}{c} f \text{ is } C^1 \text{ with } \nabla f \text{ being } L\text{-Lipschitz continuous} \\ f(\boldsymbol{x}_0) - \min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\} \leq G \\ f \text{ is } C^2 \text{ around the iterates } \boldsymbol{x}_0, \dots, \boldsymbol{x}_m \text{ generated by } \boldsymbol{\mathcal{A}} \end{array} \right\}.$$

We remark that imposing the C^2-smoothness of f around $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ is solely to expand the range of applicable algorithms; see the discussion at the end of Section 2.3.2. Besides, as this additional restriction only shrinks the function class, the lower bounds in Theorem 3.1 and Corollary 3.2 carry over *a fortiori* to the larger class of all lower-bounded $C^{1,1}$-functions.

3 Main results

3.1 Lower bound for second-order deterministic zero-respecting algorithms

Our first lower bound, the impossibility of computing Goldstein approximate second-order stationary points of $C^{1,1}$-functions in finite time by algorithms in $\mathbb{A}^n_{\text{det-zr}}(m)$, is as follows.

Theorem 3.1 *We have*

$$\max_{\substack{m=1,2,\dots \\ n=2,3,\dots \\ \boldsymbol{\mathcal{A}} \in \mathbb{A}^n_{\text{det-zr}}(m)}} \min_{f \in \mathbb{F}^n_{C^1}(4.086, 17/256; \boldsymbol{\mathcal{A}})} \max_{k=0,\dots,m} \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{z} \in \partial^2_{1/8} f(\boldsymbol{x}_k)(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{w} \rangle \leq -0.0092714,$$

where $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ are the iterates generated by $\boldsymbol{\mathcal{A}}$ when applied to f; see Section 2.3.2.

In particular, Theorem 3.1 asserts that for any $m \geq 1$, $n \geq 2$, and $\boldsymbol{\mathcal{A}} \in \mathbb{A}^n_{\text{det-zr}}(m)$, there exists an $f \in \mathbb{F}^n_{C^1}(4.086, 17/256; \boldsymbol{\mathcal{A}})$, such that the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ generated by $\boldsymbol{\mathcal{A}}$ satisfy

$$\min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{z} \in \partial^2_{1/8} f(\boldsymbol{x}_k)(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{w} \rangle = \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{z} \in \operatorname{conv}\left(\bigcup_{\boldsymbol{y} \in \boldsymbol{x}_k + \mathbb{B}^n/8} \partial^2_C f(\boldsymbol{y})(\boldsymbol{w}) \right)} \langle \boldsymbol{z}, \boldsymbol{w} \rangle \leq -0.0092714$$

for all $k = 0, \dots, m$; i.e., none of the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ is $(\varepsilon, \eta, \delta)$-Goldstein approximate second-order stationary for all $\varepsilon \geq 0$, $\eta \in [0, 0.0092714)$, and $\delta \in [0, 1/8]$.

The proof of Theorem 3.1 is (a bit involved and thus) deferred to Section 4 and Appendix B. Informally, we use a resisting oracle that confines zero-respecting iterates to the $\boldsymbol{e}_1$-axis,

[2] We note in passing that this direction corresponds to the minimum-norm solution of the linear system $\nabla^2 f(\boldsymbol{x})\boldsymbol{d} + \nabla f(\boldsymbol{x}) = \boldsymbol{0}$ (when it is satisfiable).

construct an ambient $C^{1,1}$-function with uniformly negative curvature near this trajectory, and apply a C^3-bump to maintain oracle compatibility while preserving the desired curvature.

3.2 Lower bound for second-order (general) deterministic algorithms

While Theorem 3.1 only applies to $\mathbb{A}^n_{\text{det-zr}}(m)$, with the help of an orthogonal transformation as employed in, e.g., (Nemirovsky & Yudin, 1983, Section 7.2.2) and the proof of (Tian & So, 2024, Theorem 1), it can be further extended to $\mathbb{A}^n_{\text{det}}(m)$, albeit with a slight reduction in hardness; i.e., from finite-time intractability to dimension-free intractability.

Corollary 3.2 *We have*

$$\max_{\substack{m=1,2,\dots\\ n=m+3,m+4,\dots\\ \boldsymbol{\mathcal{A}}\in\mathbb{A}^n_{\text{det}}(m)}} \min_{f\in\mathbb{F}^n_{C^1}(4.086,17/256;\boldsymbol{\mathcal{A}})} \max_{k=0,\dots,m} \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{z}\in\partial^2_{1/8}f(\boldsymbol{x}_k)(\boldsymbol{w})} \langle \boldsymbol{z},\boldsymbol{w}\rangle \le -0.0092714,$$

where $\boldsymbol{x}_0,\dots,\boldsymbol{x}_m$ are the iterates generated by $\boldsymbol{\mathcal{A}}$ when applied to f; see Section 2.3.2.

The proof of Corollary 3.2, which relies on the second-order Goldstein subdifferential chain rule established in Section 2.2 (i.e., Corollary 2.9), is deferred to Appendix C.

3.3 Discussions

3.3.1 Relation to deterministic first-order lower bounds

Our lower bounds may be viewed as second-order counterparts of the deterministic first-order lower bounds of Tian & So (2024) and Jordan et al. (2023). The connection is thematic rather than technical: While both establish deterministic hardness for Goldstein-type stationarity, in the earlier constructions, nonsmooth hard functions keep the Goldstein subdifferential uniformly separated from the origin, preventing nearby subgradients from canceling, while our hard functions, by contrast, are $C^{1,1}$ and have vanishing gradients at all iterates. The hardness arises instead from a common negative-curvature direction across the relevant generalized Hessians, thereby violating the required PSD-type condition in (1).

3.3.2 Comparison with smooth saddle-escape results

It is now folkloric that computing second-order stationary points of *smooth* nonconvex functions is deterministically tractable; e.g., the cubic Newton methods (Griewank, 1981; Nesterov & Polyak, 2006; Cartis et al., 2011) and trust-region methods (Conn et al., 2000; Curtis et al., 2018; Jiang et al., 2026) already offer dimension-free guarantees. Even some first-order methods (e.g., gradient descent (Lee et al., 2016)) can converge to such points, albeit randomness is often introduced to obtain a polynomial or even polylogarithmic complexity (Ge et al., 2015; Jin et al., 2017; Du et al., 2017). By contrast, our results establish an $\Omega(n)$ deterministic lower bound for computing even Goldstein approximate second-order stationary points of $C^{1,1}$-functions, and a finite-time hardness for zero-respecting methods, revealing a separation between the smooth and nonsmooth regimes.

Relatedly, Dragomir et al. (2026, Proposition 9.2) establish an $\Omega(n)$ deterministic lower bound for escaping saddle points of C^∞-functions with Lipschitz Hessians under an oracle returning a function value, a gradient, and one Hessian–vector product per query. While stated in terms of function-value decrease, their construction also rules out Goldstein approximate second-order stationary points at the queried points, which is thus closely related to Corollary 3.2. By contrast, our Corollary 3.2 allows full-Hessian access and treats $C^{1,1}$-functions, while Theorem 3.1 further gives finite-time impossibility for zero-respecting methods.

4 Proof of Theorem 3.1

In this section, we outline the main steps of the proof of Theorem 3.1 and defer the proofs of the auxiliary results used along the way to Appendix B, where a complete and self-contained

proof of Theorem 3.1 is also provided. In what follows, we shall restrict attention to the case $n = 2$; for $n \geq 3$, it suffices to extend the construction to be given for $n = 2$ in the canonical way. Let $m \geq 1$ be arbitrary.

Consider the resisting oracle $\mathcal{O} : \mathbb{R}^2 \to \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}^{2\times 2}_{\mathrm{sym}}$ defined for all $\boldsymbol{x} \in \mathbb{R}^2$ by $\mathcal{O}(\boldsymbol{x}) := (0, \mathbf{0}, -\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}})$ and an arbitrary $\mathcal{A} \in \mathbb{A}^2_{\text{det-zr}}(m)$; here, $\mathcal{O}$ is referred to as "resisting" in that $-\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$ is not PSD. Let $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ be the iterates generated by $\mathcal{A}$ after m rounds of interaction with $\mathcal{O}$. By the zero-respecting property of $\mathcal{A}$, we have $\boldsymbol{x}_0 = \mathbf{0}$ and $\mathrm{supp}(\boldsymbol{x}_k) \subseteq \{1\}$ for all $k = 0, \ldots, m$; i.e., $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ lie on the x-axis. Without loss of generality, we may assume that $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ for all $i \neq j$, and, by relabeling (and abusing notation), that $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ are sorted in ascending order of their x-coordinates. In particular, this implies that $\boldsymbol{x}_0$ may no longer be $\mathbf{0}$ anymore. For ease of exposition, we also define $\boldsymbol{x}_{-1} := (-\infty, 0)^{\mathrm{T}}$ and $\boldsymbol{x}_{m+1} = (\infty, 0)^{\mathrm{T}}$, and with a further abuse of notation, we denote by x_k the x-coordinate of $\boldsymbol{x}_k$ for $k = -1, \ldots, m+1$; e.g., $x_{-1} = -\infty$. Inspired by existing constructions (Tian & So, 2024; Jordan et al., 2023) that lower bound the complexity of Lipschitz optimization, in what follows, we shall construct a hard function in $\mathbb{F}^2_{C^1}(4.086, 17/256; \mathcal{A})$ such that it is compatible with $\mathcal{O}$ along $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ but none of the iterates is Goldstein approximate second-order stationary up to the tolerances specified in Theorem 3.1.

Define

$$\zeta := \min\{\xi, 1\} \in (0, 1], \quad \text{where } \xi := \min\big\{|x_i - x_j| : i, j = 0, \ldots, m \text{ with } i \neq j\big\} > 0.$$

Consider the hard function $f(\bullet, \bullet; \zeta) : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y; \zeta) := \begin{cases} f_k(x, y; \zeta), & |x - x_k| \leq \zeta/4, \\ q(y; 1), & \text{otherwise}, \end{cases} \tag{3}$$

where for $k = 0, \ldots, m$, each $\delta, \gamma > 0$, and $(\bullet)_+ := \max\{\bullet, 0\}$,

$$f_k(x, y; \zeta) := p_2(x - x_k, y; \zeta/32) \cdot \left(-\frac{1}{2}(x - x_k)^2\right) + \big(1 - p_2(x - x_k, y; \zeta/32)\big) \cdot h(x - x_k, y; \zeta/4),$$

$$p_2(x, y; \delta) := \left(1 - \frac{x^2 + y^2}{\delta^2}\right)_+^4, \quad q(x; \gamma) := -\frac{1}{2}x^2 + \left(|x| - \frac{\gamma}{4}\right)_+^2 - \left(|x| - \frac{3\gamma}{4}\right)_+^2 + \frac{1}{2}(|x| - \gamma)_+^2,$$

and $h(x, y; \gamma) := q(x; \gamma) + q(y; 1)$. We remark that $p_2(\bullet, \bullet; \delta)$ is C^3, whereas $q(\bullet; \gamma)$ and $h(\bullet, \bullet; \gamma)$ are $C^{1,1}$. As

$$h(x, y; \gamma) = -\frac{1}{2}x^2 - \frac{1}{2}y^2 = -2^{-1}\|\bullet\|^2(x, y) \quad \text{for all } (x, y)^{\mathrm{T}} \in \left[-\frac{\gamma}{4}, \frac{\gamma}{4}\right] \times \left[-\frac{1}{4}, \frac{1}{4}\right],$$

we see that around each $\boldsymbol{x}_k$, the function $f(\bullet, \bullet; \zeta)$ is defined to coincide with the strongly concave function $-2^{-1}\|\bullet\|^2$ (upon recentering at $\boldsymbol{x}_k$), modulo the local surgery induced by $p_2(\bullet, \bullet; \delta)$ that smoothly interpolates between $-2^{-1}\|\bullet\|^2$ and $(x, y) \mapsto -2^{-1}x^2$ (that corresponds to $\mathcal{O}$) and yields the C^3-blend $g(\bullet, \bullet; \delta) : \mathbb{R}^2 \to \mathbb{R}$ (which is also $C^{1,1}$; see Proposition 4.1) defined by

$$g(x, y; \delta) := p_2(x, y; \delta) \cdot \left(-\frac{1}{2}x^2\right) + \big(1 - p_2(x, y; \delta)\big) \cdot \left(-\frac{1}{2}x^2 - \frac{1}{2}y^2\right).$$

By contrast, away from $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$, the function $f(\bullet, \bullet; \zeta)$ is interpolated such that it remains $C^{1,1}$, becomes lower-bounded, and features a sufficiently wide infinite horizontal strip around the x-axis with strong stationarity refutation; see Corollary 4.4. It is worth noting that the purpose of carrying out the local surgeries is to give rise to the compatibility of $f(\bullet, \bullet; \zeta)$ with $\mathcal{O}$ along $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$, which is otherwise absent as $\nabla^2(-2^{-1}\|\bullet\|^2)(\boldsymbol{x}) = -\boldsymbol{I} \neq -\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}$, while preserving the essential smoothness and refutation properties of $-2^{-1}\|\bullet\|^2$.

Before analyzing $f(\bullet, \bullet; \zeta)$, we first record some useful properties of $g(\bullet, \bullet; \delta)$ and $h(\bullet, \bullet; \gamma)$.

Proposition 4.1 *Let $\delta, \gamma > 0$ be constants. The following statements hold.*

- ***(Stationarity refutations)*** *Let* $\boldsymbol{w} := (10, 11)^{\mathrm{T}}/\sqrt{221} \in \mathbb{S}^2$. *We have*
$$\langle \nabla^2 g(x, y; \delta), \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \rangle \le -0.0092714 \quad \text{for all } x, y \in \mathbb{R},$$
$$\max\{\langle \boldsymbol{z}, \boldsymbol{w} \rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 h(x, y; \gamma)(\boldsymbol{w})\} \le -0.095 \quad \text{for all } (x, y)^{\mathrm{T}} \in \mathbb{R} \times (-1/4, 1/4).$$
- ***($C^{1,1}$-smoothness)*** $\mathrm{lip}\nabla g(x, y; \delta) \le 4.086$ *and* $\mathrm{lip}\nabla h(x, y; \gamma) \le 1$ *for all* $x, y \in \mathbb{R}$.
- ***(Oracle compatibility)*** $g(0, 0; \delta) = 0$, $\nabla g(0, 0; \delta) = \boldsymbol{0}$, *and* $\nabla^2 g(0, 0; \delta) = -\boldsymbol{e}_1 \boldsymbol{e}_1^{\mathrm{T}}$.

We now proceed to analyze $f(\bullet, \bullet; \zeta)$, starting with a piecewise structure of $f_k(\bullet, \bullet; \zeta)$.

Lemma 4.2 *For all* $k = 0, \dots, m$, *we have*
$$f_k(x, y; \zeta) = \begin{cases} h(x - x_k, y; \zeta/4), & (x, y)^{\mathrm{T}} \notin (x_k, 0)^{\mathrm{T}} + \dfrac{\zeta}{32}\mathbb{B}^2, \\ g(x - x_k, y; \zeta/32), & (x, y)^{\mathrm{T}} \in \left[x_k - \dfrac{\zeta/4}{4}, x_k + \dfrac{\zeta/4}{4}\right] \times \left[-\dfrac{1}{4}, \dfrac{1}{4}\right]. \end{cases}$$

With Lemma 4.2 in place, we are now ready to derive a piecewise representation of $f(\bullet, \bullet; \zeta)$.

Proposition 4.3 *Consider the open sets* $\mathbb{H}_0, \dots, \mathbb{H}_m$ *and* $\mathbb{P}_0, \dots, \mathbb{P}_m$ *defined by*
$$\mathbb{H}_k := \left(\left(x_{k-1} + \frac{\zeta}{4}, x_{k+1} - \frac{\zeta}{4}\right) \times \mathbb{R}\right) \cap \mathbb{B}_k^{\complement} \quad \text{and} \quad \mathbb{P}_k := \left(x_k - \frac{\zeta/4}{4}, x_k + \frac{\zeta/4}{4}\right) \times \left(-\frac{1}{4}, \frac{1}{4}\right),$$
where $\mathbb{B}_k := (x_k, 0)^{\mathrm{T}} + \zeta\mathbb{B}^2/32$. *We have* $\bigcup_{k=0}^m \mathbb{H}_k \cup \bigcup_{k=0}^m \mathbb{P}_k = \mathbb{R}^2$ *and*
$$f(x, y; \zeta) = \begin{cases} h(x - x_k, y; \zeta/4), & (x, y)^{\mathrm{T}} \in \mathbb{H}_k, \\ g(x - x_k, y; \zeta/32), & (x, y)^{\mathrm{T}} \in \mathbb{P}_k. \end{cases}$$

By combining Proposition 4.1 and Proposition 4.3, we have the following properties.

Corollary 4.4 *The following statements hold.*

- ***($C^{1,1}$-smoothness)*** *The function* $f(\bullet, \bullet; \zeta)$ *is* C^1 *with* $\nabla f(\bullet, \bullet; \zeta)$ *being* 4.086*-Lipschitz.*
- ***(Lower boundedness)*** *We have* $f(0, 0; \zeta) - \min\{f(x, y; \zeta) : x, y \in \mathbb{R}\} \le 17/256$.
- ***(Local C^3-smoothness)*** *The function* $f(\bullet, \bullet; \zeta)$ *is* C^3 *(and thus* C^2*) around* $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$.
- ***(Oracle compatibility)*** *The function* $f(\bullet, \bullet; \zeta)$ *is compatible with* $\mathcal{O}$ *along* $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$.
- ***(Band refutation)*** *Let* $\boldsymbol{w} := (10, 11)^{\mathrm{T}}/\sqrt{221} \in \mathbb{S}^2$. *We have*
$$\max\left\{\langle \boldsymbol{z}, \boldsymbol{w} \rangle : \boldsymbol{z} \in \mathrm{conv}\left(\bigcup_{(x,y)^{\mathrm{T}} \in \mathbb{R} \times (-1/4, 1/4)} \partial_{\mathrm{C}}^2 f(x, y; \zeta)(\boldsymbol{w})\right)\right\} \le -0.0092714.$$

We note that the first three properties together imply $f(\bullet, \bullet; \zeta) \in \mathbb{F}_{C^1}^2(4.086, 17/256; \boldsymbol{\mathcal{A}})$. By combining this membership, the other two properties in Corollary 4.4, and the fact that $(x_k, 0)^{\mathrm{T}} + \mathbb{B}^2/8 \subseteq \mathbb{R} \times (-1/4, 1/4)$ for all $k = 0, \dots, m$, the whole proof is complete.

5 Concluding remarks

In this paper, we have shown the intractability of deterministic second-order $C^{1,1}$-optimization (and thus weakly convex optimization). Specifically, we have shown that for any $m \ge 1$, any $n \ge 2$ (resp., $n \ge m + 3$), and any $\boldsymbol{\mathcal{A}} \in \mathbb{A}_{\text{det-zr}}^n(m)$ (resp., $\boldsymbol{\mathcal{A}} \in \mathbb{A}_{\text{det}}^n(m)$), there exists an $f \in \mathbb{F}_{C^1}^n(4.086, 17/256; \boldsymbol{\mathcal{A}})$ such that the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ generated by $\boldsymbol{\mathcal{A}}$ when applied to f are not $(\varepsilon, \eta, \delta)$-Goldstein approximate second-order stationary for all $\varepsilon \ge 0$, $\eta \in [0, 0.0092714)$, and $\delta \in [0, 1/8]$. Recently, Guan & So (2026) have established a polynomial randomized upper bound for the problem, but it does not bypass the lower bounds here. An intriguing question is whether we can use randomization to bypass the lower bounds; see, e.g., (Zhang et al., 2020; Tian et al., 2022; Davis et al., 2022; Lin et al., 2022; Metel & Takeda, 2022; Kornowski & Shamir, 2024) for first-order counterparts.

AI use statement

In this work, we used generative AI tools for "assist with translation". We have not used generative AI tools for "help develop theoretical models or conceptual frameworks", "formulate mathematical claims", "provide critical ingredients for proving mathematical claims", "assist in the writing of proofs", "propose or refine hypotheses", "interpret results", and "generate synthetic data sets", "design or provide feedback on research methodology or experiments", "implement methods", "clean and reformat dataset", "support qualitative and thematic data analysis" are not applicable to this work. We have reviewed all AI-assisted work, to the best of our linguistic abilities. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

Coralia Cartis, Nicholas I M Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate descent method for large-scale L2-loss linear support vector machines. *Journal of Machine Learning Research*, 9(45): 1369–1398, 2008.

Bilian Chen and Zhening Li. On the tensor spectral p-norm and its dual norm via partitions. *Computational Optimization and Applications*, 75(3):609–628, 2020.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

Andrew R Conn, Nicholas I M Gould, and Philippe L Toint. *Trust-Region Methods*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2000.

Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2021.

Frank E Curtis, Zachary Lubberts, and Daniel P Robinson. Concise complexity analyses for trust region methods. *Optimization Letters*, 12:1713–1724, 2018.

Aris Daniilidis and Dmitriy Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 6692–6703, 2022.

Radu-Alexandru Dragomir, Xiaowen Jiang, Bonan Sun, and Nicolas Boumal. A practical randomized trust-region method to escape saddle points in high dimension. *arXiv preprint arXiv:2603.15494*, 2026.

Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1067–1077, 2017.

Antonio Fuduli, Manlio Gaudioso, and Giovanni Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3):743–756, 2004.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Annual Conference on Learning Theory*, pp. 797–842, 2015.

Bennet Gebken. Using second-order information in gradient sampling methods for nonsmooth optimization. *arXiv preprint arXiv:2210.04579*, 2022.

A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

Andreas Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981.

Jiewen Guan and Anthony Man-Cho So. On computing Goldstein approximate second-order stationary points of structured nonsmooth nonconvex programs. *arXiv preprint arXiv:2607.24122*, 2026.

Jean-Baptiste Hiriart-Urruty, Jean-Jacques Strodiot, and V. Hien Nguyen. Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. *Applied Mathematics and Optimization*, 11(1):43–56, 1984.

Tim Hoheisel, Christian Kanzow, Boris S. Mordukhovich, and Hung Phan. Generalized Newton's method based on graphical derivatives. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1324–1340, 2012.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 408–415, 2008.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Yuntian Jiang, Chang He, Chuwen Zhang, Dongdong Ge, Bo Jiang, and Yinyu Ye. Beyond nonconvexity: A universal trust-region method with new analyses. *Journal of Scientific Computing*, 106(1):28, 2026.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1724–1732, 2017.

Michael I. Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Proceedings of the 36th Annual Conference on Learning Theory*, pp. 4570–4597, 2023.

Pham Duy Khanh, Boris S. Mordukhovich, Vo Thanh Phat, and Dat Ba Tran. Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials. *Journal of Global Optimization*, 86(1):93–122, 2023.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 324–334, 2021.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122):1–14, 2024.

Bernd Kummer. Newton's method for non-differentiable functions. *Mathematical Research*, 45:114–125, 1988.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 1246–1257, 2016.

Ming Lei, Ting Kei Pong, Man-Chung Yue, Liaoyuan Zeng, and Hao Zhang. A Gaussian smoothing-based zeroth-order method for Goldstein second-order stationarity. *arXiv preprint arXiv:2607.21258*, 2026.

Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020a.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and René Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020b.

Tianyi Lin, Zeyu Zheng, and Michael I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 26160–26175, 2022.

Michael R. Metel and Akiko Takeda. Perturbed iterate SGD for Lipschitz continuous loss functions. *Journal of Optimization Theory and Applications*, 195(2):504–547, 2022.

Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin, Heidelberg, first edition, 2006.

Boris S. Mordukhovich. *Second-Order Variational Analysis in Optimization, Variational Stability, and Control: Theory, Algorithms, Applications*. Springer Series in Operations Research and Financial Engineering. Springer Nature Switzerland AG, Cham, Switzerland, 2024.

Boris S. Mordukhovich and M. Ebrahim Sarabi. Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization. *SIAM Journal on Optimization*, 31(2):1184–1214, 2021.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons Ltd., Chichester, 1983. English translation from "Slozhnost' Zadach i Effektivnost' Metodov Optimizatsii" by E. R. Dawson; original Russian publication in Glavnaya Redaktsiya Fiziko–Matematicheskoy Literatury Izdatel'stva "Nauka", 1979.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk*, 269(3):543–547, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer Nature Switzerland AG, Cham, Switzerland, second edition, 2018.

Yurii Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer Science+Business Media, New York, 2006.

Jong-Shi Pang. Newton's method for B-differentiable equations. *Mathematics of Operations Research*, 15(2):311–341, 1990.

Liqun Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18(1):227–244, 1993.

Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1):353–367, 1993.

R. Tyrrell Rockafellar. Extensions of subgradient calculus with applications to optimization. *Nonlinear Analysis: Theory, Methods & Applications*, 9(7):665–698, 1985.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton, New Jersey, 1997.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics*. Springer Berlin, Heidelberg, 2009.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Lai Tian and Anthony Man-Cho 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.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

Lai Tian and Anthony Man-Cho So. Computational nonsmooth analysis: Modern applications and recent developments. *IEEE Signal Processing Magazine*, 43(3):102–114, 2026. Special Issue on the Mathematics of Deep Learning.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 21360–21379, 2022.

Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11173–11182, 2020.

A Proof of Lemma 2.8

Proof of Lemma 2.8. Extend $\boldsymbol{U}$ to an orthogonal matrix $\boldsymbol{U}' \in \mathbb{R}^{n\times n}$, and let $h : \mathbb{R}^n \to \mathbb{R}$ be such that $h(\boldsymbol{x}) = f(\boldsymbol{x}_{[m]})$; as a quick sanity check, we have $g(\boldsymbol{x}) = h(\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})$. It follows from the proof of (Tian & So, 2024, Theorem 1) that $\partial_\delta g(\boldsymbol{x}) = \boldsymbol{U}'\partial_\delta h(\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})$. However, it also follows from the separable structure of h and (Rockafellar, 1985, Proposition 2.5) that $\partial_{\mathrm{C}} h(\boldsymbol{x}) = \partial_{\mathrm{C}} f(\boldsymbol{x}_{[m]}) \times \{\boldsymbol{0}\}$. As a result, we have

$$\begin{aligned}\partial_\delta h(\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x}) &= \operatorname{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x}+\delta\mathbb{B}^n} \partial_{\mathrm{C}} h(\boldsymbol{y})\right) = \operatorname{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x}+\delta\mathbb{B}^n} \begin{pmatrix}\partial_{\mathrm{C}} f(\boldsymbol{y}_{[m]})\\ \boldsymbol{0}\end{pmatrix}\right)\\ &= \begin{pmatrix}\operatorname{conv}\left(\bigcup_{\boldsymbol{z}\in(\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})_{[m]}+\delta\mathbb{B}^m} \partial_{\mathrm{C}} f(\boldsymbol{z})\right)\\ \boldsymbol{0}\end{pmatrix} = \begin{pmatrix}\partial_\delta f\left((\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})_{[m]}\right)\\ \boldsymbol{0}\end{pmatrix} = \begin{pmatrix}\partial_\delta f(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x})\\ \boldsymbol{0}\end{pmatrix},\end{aligned}$$

where the third equality follows from the fact that

$$\begin{aligned}\left\{\boldsymbol{y}_{[m]} : \boldsymbol{y} \in \boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x} + \delta\mathbb{B}^n\right\} &= \left\{\boldsymbol{z} : \left\|\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{z}'\end{pmatrix} - \boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x}\right\| \le \delta \text{ for some } \boldsymbol{z}' \in \mathbb{R}^{n-m}\right\}\\ &= \left\{\boldsymbol{z} : \min_{\boldsymbol{z}'\in\mathbb{R}^{n-m}} \left\|\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{z}'\end{pmatrix} - \boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x}\right\| \le \delta\right\}\\ &= \left\{\boldsymbol{z} : \|\boldsymbol{z} - (\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})_{[m]}\| \le \delta\right\} = (\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})_{[m]} + \delta\mathbb{B}^m.\end{aligned}$$

This, together with $\partial_\delta g(\boldsymbol{x}) = \boldsymbol{U}'\partial_\delta h(\boldsymbol{U}'^{\mathrm{T}}\boldsymbol{x})$ mentioned earlier, immediately implies that

$$\partial_\delta g(\boldsymbol{x}) = \boldsymbol{U}'\begin{pmatrix}\partial_\delta f(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x})\\ \boldsymbol{0}\end{pmatrix} = \boldsymbol{U}\partial_\delta f(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}),$$

as desired. □

B Full proof of Theorem 3.1

B.1 Preparation: The local surgery

Recall from Section 4 that we shall carry out certain local surgeries on an ambient function that is locally identical to $-2^{-1}\|\bullet\|^2$. In this subsection, we describe the local surgery in full detail.

To begin, consider the one-dimensional C^3 bump function $p_1(\bullet;\delta):\mathbb{R}\to\mathbb{R}$ defined by

$$p_1(x;\delta) := \begin{cases} \left(\frac{x^2}{\delta^2}-1\right)^4, & |x|\le\delta, \\ 0, & \text{otherwise}, \end{cases}$$

where $\delta>0$, and its induced two-dimensional bump function $p_2(\bullet,\bullet;\delta):\mathbb{R}^2\to\mathbb{R}$ defined by[3]

$$p_2(x,y;\delta) := p_1\big(\sqrt{x^2+y^2};\delta\big) = \begin{cases} \left(\frac{x^2+y^2}{\delta^2}-1\right)^4, & \sqrt{x^2+y^2}\le\delta, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

which inherits the C^3-smoothness of $p_1(\bullet;\delta)$: In light of the composition

$$p_2(\bullet,\bullet;\delta) = p_1(\bullet;\delta)\circ\sqrt{\bullet}\circ\big((x,y)\mapsto x^2+y^2\big)$$

and the fact that $\sqrt{\bullet}$ is C^∞ on $(0,\infty)$, we know that $p_2(\bullet,\bullet;\delta)$ is C^3 on $\mathbb{R}^2\setminus\{\mathbf{0}\}$. On the other hand, we see by (4) that there exists an open neighborhood of $\mathbf{0}$ (i.e., $\delta\mathrm{int}(\mathbb{B}^2)$) on which $p_2(\bullet,\bullet;\delta)$ is a polynomial; hence, $p_2(\bullet,\bullet;\delta)$ is C^∞ around $\mathbf{0}$. By putting the above two pieces together, it follows that $p_2(\bullet,\bullet;\delta)$ is (globally) C^3.

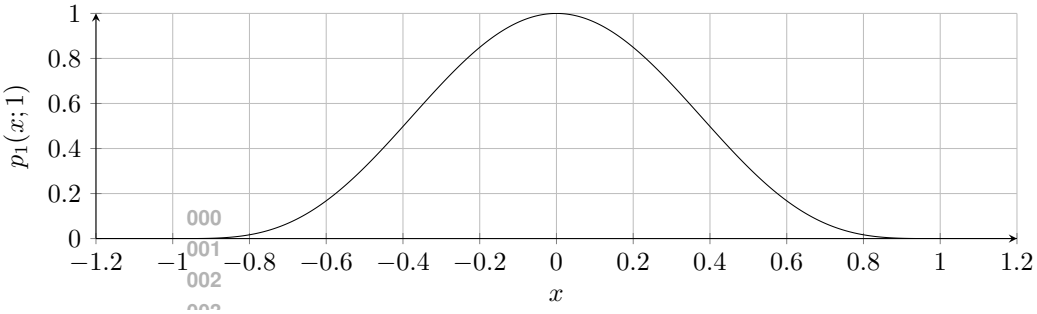

Figure 1: The landscape of $p_1(\bullet;1)$.

Recall from Section 4 that the local surgery pertains to interpolating between $-2^{-1}\|\bullet\|^2$ and $(x,y)\mapsto -2^{-1}x^2$. To this end, consider the smooth blend $g(\bullet,\bullet;\delta):\mathbb{R}^2\to\mathbb{R}$ defined by

$$\begin{aligned} g(x,y;\delta) &:= p_2(x,y;\delta)\cdot\left(-\frac{1}{2}x^2\right)+\big(1-p_2(x,y;\delta)\big)\cdot\left(-\frac{1}{2}x^2-\frac{1}{2}y^2\right) \\ &= \begin{cases} \frac{1}{2}\left(y^2\left(\frac{x^2+y^2}{\delta^2}-1\right)^4-x^2-y^2\right), & \sqrt{x^2+y^2}\le\delta, \\ \frac{1}{2}(-x^2-y^2), & \text{otherwise}. \end{cases} \end{aligned}$$

[3]It is worth noting that the exponent 4 in (4) is necessary and essential, and its purpose is to enforce a rapid decay within a small vicinity near $\mathbf{0}$, ensuring that certain undesired magnitudes that arise in later constructions remain controllable when gated by $p_2(\bullet,\bullet;\delta)$; if we replace 4 with, e.g., 3 or 2, then the validity of the remaining analysis would break down.

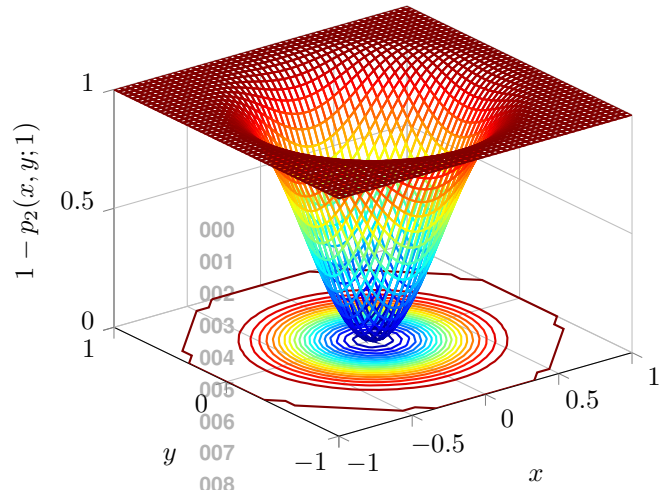

Figure 2: The landscape of $1 - p_2(\bullet, \bullet; 1)$.

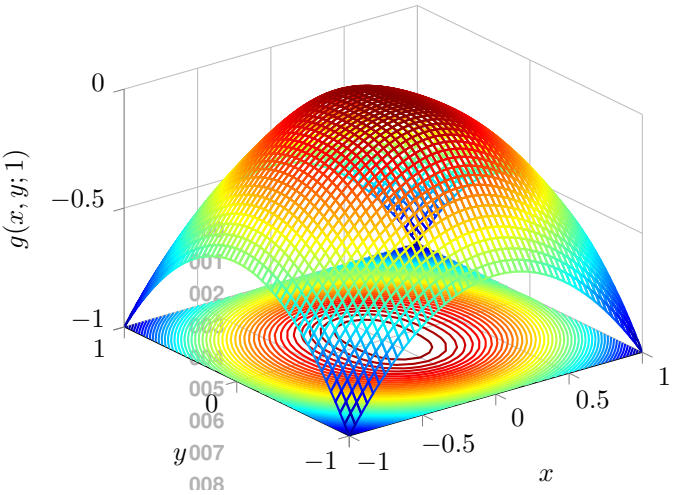

Figure 3: The landscape of $g(\bullet, \bullet; 1)$.

By applying Leibniz's rule, it follows that $g(\bullet, \bullet; \delta)$ is C^3 as well, with

$$\nabla g(x, y; \delta) = \begin{cases} \begin{pmatrix} x\left(\frac{4y^2(-\delta^2+x^2+y^2)^3}{\delta^8} - 1\right) \\ y\left(\left(\frac{x^2+y^2}{\delta^2} - 1\right)^4 + \frac{4y^2(-\delta^2+x^2+y^2)^3}{\delta^8} - 1\right) \end{pmatrix}, & \sqrt{x^2+y^2} \le \delta, \\ \begin{pmatrix} -x \\ -y \end{pmatrix}, & \text{otherwise,} \end{cases}$$

and

$$\nabla^2 g(x, y; \delta) = \begin{cases} \boldsymbol{H}(x, y; \delta) = \begin{pmatrix} h_{11}(x, y; \delta) & h_{12}(x, y; \delta) \\ h_{21}(x, y; \delta) & h_{22}(x, y; \delta) \end{pmatrix}, & \sqrt{x^2+y^2} \le \delta, \\ -\boldsymbol{I}, & \text{otherwise,} \end{cases}$$

where

$$\begin{cases} h_{11}(x,y;\delta) := \dfrac{-\delta^8 + 4y^2(-\delta^2+x^2+y^2)^2(-\delta^2+7x^2+y^2)}{\delta^8}, \\ h_{12}(x,y;\delta) := \dfrac{8xy(-\delta^2+x^2+y^2)^2(-\delta^2+x^2+4y^2)}{\delta^8}, \\ h_{21}(x,y;\delta) := h_{12}(x,y;\delta), \\ h_{22}(x,y;\delta) := \dfrac{\begin{pmatrix} -\delta^8 + (-\delta^2+x^2+y^2)^4 + 16y^2(-\delta^2+x^2+y^2)^3 \\ +4y^2(-\delta^2+x^2+y^2)^2(-\delta^2+x^2+7y^2) \end{pmatrix}}{\delta^8}, \end{cases}$$

which immediately imply the desired oracle compatibility mentioned in Section 4.

Lemma B.1 *Let $\delta > 0$ be a constant. We have*

$$g(0,0;\delta) = 0, \quad \nabla g(0,0;\delta) = \mathbf{0}, \quad \textit{and} \quad \nabla^2 g(0,0;\delta) = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}.$$

With Lemma B.1 in place, as discussed in Section 4, it then remains to verify the (stationarity) refutation property of $g(\bullet,\bullet;\delta)$ and to estimate the Lipschitz modulus of $\nabla g(\bullet,\bullet;\delta)$. To this end, a few interesting technical preparations are in order.

B.1.1 Technical bounds

In this part, we shall bound the elements of $\boldsymbol{H}(x,y;\delta)$ on $\delta\mathbb{B}^2$. In view of the fact that

$$h_{ij}(x,y;\delta) = h_{ij}(\delta^{-1}x, \delta^{-1}y; 1) \quad \text{for all } i,j = 1,2$$

and the following trivial lemma (with $\boldsymbol{A} := \delta^{-1}\boldsymbol{I}$), it suffices to assume that $\delta = 1$ throughout the proofs that follow.

Lemma B.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^m$ be arbitrary, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be invertible. We have*

$$\big\{f(\boldsymbol{x}) : \boldsymbol{G}(\boldsymbol{x}) \leq \mathbf{0}\big\} = \big\{f(\boldsymbol{A}\boldsymbol{x}) : \boldsymbol{G}(\boldsymbol{A}\boldsymbol{x}) \leq \mathbf{0}\big\}.$$

To put it simple, linear isomorphisms preserve (constrained) function ranges.

Proof. Suppose that $v \in \big\{f(\boldsymbol{x}) : \boldsymbol{G}(\boldsymbol{x}) \leq \mathbf{0}\big\}$. Then, we have $v = f(\boldsymbol{x}_0)$ for some $\boldsymbol{x}_0 \in \mathbb{R}^n$ with $\boldsymbol{G}(\boldsymbol{x}_0) \leq \mathbf{0}$. However, we also have $f(\boldsymbol{x}_0) = f(\boldsymbol{A}\boldsymbol{x}_0')$ and $\boldsymbol{G}(\boldsymbol{x}_0) = \boldsymbol{G}(\boldsymbol{A}\boldsymbol{x}_0')$ for $\boldsymbol{x}_0' := \boldsymbol{A}^{-1}\boldsymbol{x}_0$. Hence, it follows that $v = f(\boldsymbol{A}\boldsymbol{x}_0')$ for some $\boldsymbol{x}_0' \in \mathbb{R}^n$ such that $\boldsymbol{G}(\boldsymbol{A}\boldsymbol{x}_0') \leq \mathbf{0}$; i.e., $v \in \big\{f(\boldsymbol{A}\boldsymbol{x}) : \boldsymbol{G}(\boldsymbol{A}\boldsymbol{x}) \leq \mathbf{0}\big\}$, as desired. The other direction is trivial. □

As an aside, we note for all $x, y \in \mathbb{R}$ with $\sqrt{x^2+y^2} = \delta$ that

$$h_{11}(x,y;\delta) = -1, \quad h_{12}(x,y;\delta) = h_{21}(x,y;\delta) = 0, \quad \text{and} \quad h_{22}(x,y;\delta) = -1;$$

hence, we shall also assume that $\sqrt{x^2+y^2} \neq \delta$ in the sequel.

Proposition B.3 *Let $\delta > 0$ be a constant. We have*

$$-\frac{91}{64} \leq h_{11}(x,y;\delta) \leq -\frac{71}{98} \quad \textit{for all } x,y \in \mathbb{R} \textit{ such that } \sqrt{x^2+y^2} \leq \delta.$$

Proof. To begin, we calculate

$$h_{11}(x,y;1) = 4y^2(x^2+y^2-1)^2(7x^2+y^2-1) - 1.$$

In what follows, we shall derive a set of candidate stationary points of $h_{11}(\bullet,\bullet;1)$ and then evaluate and compare at these candidates; as $h_{11}(x,0;1) = -1$ for all $x \in \mathbb{R}$, we shall also assume that $y \neq 0$.

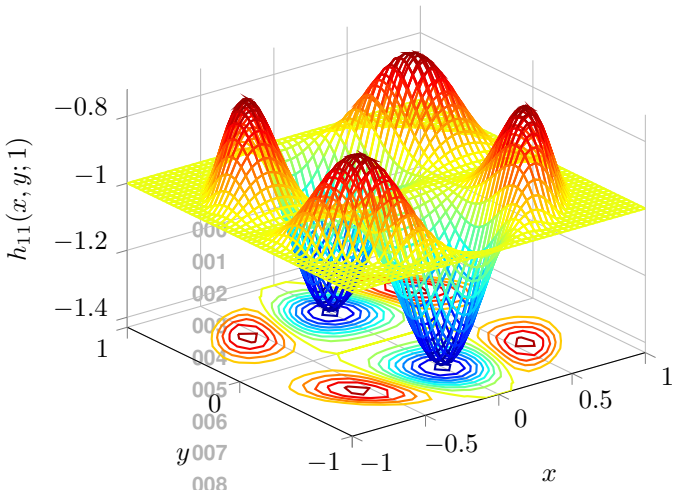

Figure 4: The landscape of $h_{11}(\bullet, \bullet; 1)$; for visual clarity, we overwrite $h_{11}(x, y; 1) = -1$ on $(\mathbb{B}^2)^{\complement}$.

It can be calculated that

$$\frac{\partial}{\partial x} h_{11}(x, y; 1) = 56xy^2(x^2 + y^2 - 1)^2 + 16xy^2(x^2 + y^2 - 1)(7x^2 + y^2 - 1)$$

and

$$\frac{\partial}{\partial y} h_{11}(x, y; 1) = 8y(x^2+y^2-1)^2(7x^2+y^2-1)+8y^3(x^2+y^2-1)^2+16y^3(x^2+y^2-1)(7x^2+y^2-1).$$

By setting these partial derivatives to zero and simplifying, we obtain

$$\begin{cases} 24xy^2(x^2 + y^2 - 1)(7x^2 + 3y^2 - 3) = 0, \\ 8y(x^2 + y^2 - 1)\big(7x^4 + x^2(23y^2 - 8) + 4y^4 - 5y^2 + 1\big) = 0. \end{cases}$$

By using the assumptions that $\sqrt{x^2 + y^2} \neq 1$ and $y \neq 0$, we can further simplify the system to

$$\begin{cases} x(7x^2 + 3y^2 - 3) = 0, \\ 7x^4 + x^2(23y^2 - 8) + 4y^4 - 5y^2 + 1 = 0. \end{cases} \tag{5}$$

As the second equation is quadratic in y^2, we can solve for y^2 and this yields

$$y^2 = \frac{1}{8}\left(-23x^2 \pm \sqrt{3}\sqrt{139x^4 - 34x^2 + 3} + 5\right). \tag{6}$$

We divide the subsequent discussions into two cases.

- Suppose that $y^2 = 8^{-1}\big(-23x^2 - \sqrt{3}\sqrt{139x^4 - 34x^2 + 3} + 5\big)$. Then, by substituting this into the first equation in (5) and simplifying, it follows that either $x = 0$ or
$$9 \leq 13x^2 + 3\sqrt{417x^4 - 102x^2 + 9} + 9 = 0,$$
which is impossible. Hence, we only have one pair of solutions in this case; i.e., $(0, \pm 2^{-1})$.
- Suppose that $y^2 = 8^{-1}\big(-23x^2 + \sqrt{3}\sqrt{139x^4 - 34x^2 + 3} + 5\big)$. Then, we similarly know that either $x = 0$ or
$$3\sqrt{417x^4 - 102x^2 + 9} = 13x^2 + 9 \implies x = 0, \pm\frac{3}{2\sqrt{7}}.$$
Hence, in this case, we have the following four solutions
$$\left(\pm\frac{3}{2\sqrt{7}}, \pm\frac{1}{2}\right);$$
the other pair of solutions $(0, \pm 1)$ is discarded as the assumption $\sqrt{x^2 + y^2} \neq 1$ is violated.

By comparing $h_{11}(x, y; 1)$ at all calculated points and boundary cases (that are explicitly excluded by the assumptions), the desired claim directly follows. The proof is complete. □

Proposition B.4 *Let $\delta > 0$ be a constant. For all $x, y \in \mathbb{R}$ with $\sqrt{x^2+y^2} \le \delta$, we have*

$$\left|h_{12}(x, y; \delta)\right| \le \frac{1}{2}\sqrt{\frac{1}{2}\left(\sqrt{177}+17\right)}\left(\frac{1}{64}\left(\sqrt{177}+17\right)-\frac{7}{8}\right)^2\left(\frac{1}{16}\left(\sqrt{177}+17\right)-\frac{7}{8}\right) \le 0.32.$$

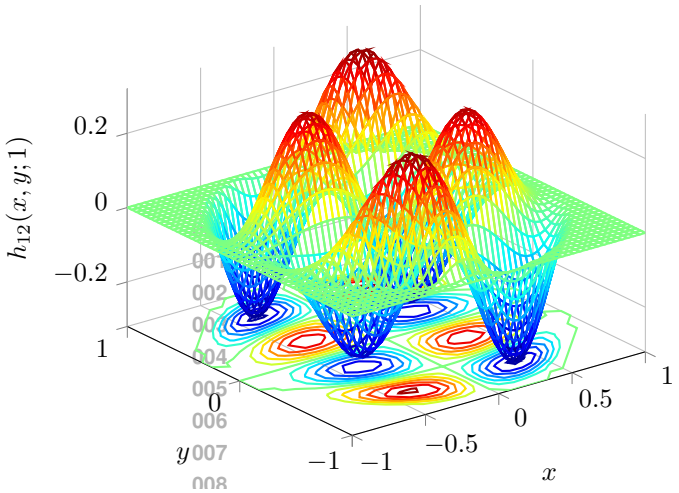

Figure 5: The landscape of $h_{12}(\bullet, \bullet; 1)$; for visual clarity, we overwrite $h_{12}(x, y; 1) = 0$ on $(\mathbb{B}^2)^{\complement}$.

Proof. Similar to the proof of Proposition B.3, we first calculate

$$h_{12}(x, y; 1) = 8xy(x^2+y^2-1)^2(x^2+4y^2-1),$$

together with

$$\frac{\partial}{\partial x}h_{12}(x, y; 1) = 16x^2y(x^2+y^2-1)^2+32x^2y(x^2+y^2-1)(x^2+4y^2-1)+8y(x^2+y^2-1)^2(x^2+4y^2-1)$$

and

$$\frac{\partial}{\partial y}h_{12}(x, y; 1) = 64xy^2(x^2+y^2-1)^2+32xy^2(x^2+y^2-1)(x^2+4y^2-1)+8x(x^2+y^2-1)^2(x^2+4y^2-1);$$

as $h_{12}(x, y; 1) = 0$ as soon as $x = 0$ or $y = 0$, we further suppose that $x \neq 0$ and $y \neq 0$ in the sequel. By setting these partial derivatives to zero and simplifying, we obtain

$$\begin{cases} 8y(x^2+y^2-1)\left(7x^4+x^2(23y^2-8)+4y^4-5y^2+1\right) = 0, \\ 8x(x^2+y^2-1)\left(17(x^2-1)y^2+(x^2-1)^2+28y^4\right) = 0, \end{cases}$$

which, upon incorporating all the assumptions mentioned earlier, further simplifies to

$$\begin{cases} 7x^4+x^2(23y^2-8)+4y^4-5y^2+1 = 0, \\ 17(x^2-1)y^2+(x^2-1)^2+28y^4 = 0. \end{cases}$$

As the first equation is the same to the second one in (5), we know that (6) remains valid here. Similarly, we divide the subsequent discussions into two cases.

- Suppose that $y^2 = 8^{-1}\left(-23x^2-\sqrt{3}\sqrt{139x^4-34x^2+3}+5\right)$. Then, we know from the second equation that

$$(8x^2-1)\left(61x^2+3\sqrt{417x^4-102x^2+9}-7\right) = 0;$$

i.e., either $x^2 = 1/8$ or

$$9 \leq 61x^2 + 3\sqrt{417x^4 - 102x^2 + 9} = 7,$$

which is impossible. Hence, the solutions in this case are merely

$$\left(\pm\frac{1}{2\sqrt{2}}, \pm\frac{1}{8}\sqrt{17 - \sqrt{177}}\right).$$

- Suppose that $y^2 = 8^{-1}\left(-23x^2 + \sqrt{3}\sqrt{139x^4 - 34x^2 + 3} + 5\right)$. Then, we similarly have either $x^2 = 1/8$ or

$$-61x^2 + 3\sqrt{417x^4 - 102x^2 + 9} + 7 = 0 \implies x = \pm 1,$$

which should be discarded as the assumption $\sqrt{x^2 + y^2} \neq 1$ is violated. Hence, only the below solution pairs arise in this case

$$\left(\pm\frac{1}{2\sqrt{2}}, \pm\frac{1}{8}\sqrt{17 + \sqrt{177}}\right).$$

The desired claim then directly follows by putting the above pieces together. We are done. □

Proposition B.5 *Let $\delta > 0$ be a constant. We have*

$$-\frac{253}{125} \leq h_{22}(x, y; \delta) \leq 0 \quad \textit{for all } x, y \in \mathbb{R} \textit{ such that } \sqrt{x^2 + y^2} \leq \delta.$$

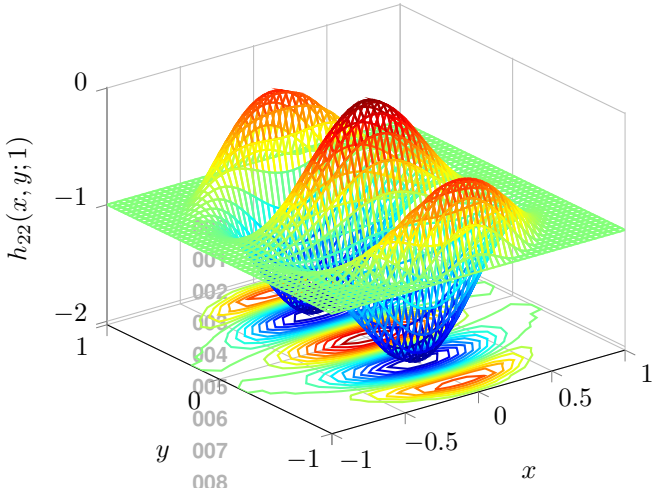

Figure 6: The landscape of $h_{22}(\bullet, \bullet; 1)$; for visual clarity, we overwrite $h_{22}(x, y; 1) = -1$ on $(\mathbb{B}^2)^{\complement}$.

Proof. In a similar vein, we begin by calculating

$$h_{22}(x, y; 1) = (x^2 + y^2 - 1)^4 + 16y^2(x^2 + y^2 - 1)^3 + 4y^2(x^2 + 7y^2 - 1)(x^2 + y^2 - 1)^2 - 1,$$

together with

$$\frac{\partial}{\partial x} h_{22}(x, y; 1) = 8x(x^2 + y^2 - 1)^3 + 104xy^2(x^2 + y^2 - 1)^2 + 16xy^2(x^2 + 7y^2 - 1)(x^2 + y^2 - 1)$$

and

$$\begin{aligned}\frac{\partial}{\partial y} h_{22}(x, y; 1) = {} & 40y(x^2 + y^2 - 1)^3 + 8y(x^2 + y^2 - 1)^2(x^2 + 7y^2 - 1) \\ & + 152y^3(x^2 + y^2 - 1)^2 + 16y^3(x^2 + y^2 - 1)(x^2 + 7y^2 - 1).\end{aligned}$$

By setting these partial derivatives to zero and simplifying, we obtain

$$\begin{cases} 8x(x^2+y^2-1)\big(17(x^2-1)y^2+(x^2-1)^2+28y^4\big)=0, \\ 24y(x^2+y^2-1)(2x^2+3y^2-2)(x^2+5y^2-1)=0, \end{cases}$$

which can be further simplified to

$$\begin{cases} x\big(17(x^2-1)y^2+(x^2-1)^2+28y^4\big)=0, \\ y(2x^2+3y^2-2)(x^2+5y^2-1)=0. \end{cases}$$

As the second equation is clearly more amenable to analysis, we shall exploit it first. This time, the subsequent discussions are divided into three cases.

- Suppose that $y=0$. Then, we know from the other equation that
$$x(x^2-1)^2=0 \implies x=0,\pm 1;$$
i.e., two pairs of solutions, namely $(0,0)$ and $(\pm 1,0)$, arise in this case, but the latter should be discarded as it violates the assumption $\sqrt{x^2+y^2}\neq 1$.
- Suppose that $y^2=2(1-x^2)/3$. Then, we have
$$x^3=x \implies x=0,\pm 1;$$
hence, $(0,\pm\sqrt{2/3})$ are the only solutions in this case, as $(\pm 1,0)$ should be discarded as well.
- Suppose that $y^2=(1-x^2)/5$. Then, it follows, again, that
$$x^3=x \implies x=0,\pm 1;$$
likewise, the solutions in this case are merely $(0,\pm\sqrt{1/5})$.

Summarizing the above discussions yields the desired result and thus completes the proof. □

B.1.2 Stationarity refutations

With Propositions B.3, B.4, and B.5 in place, we are ready to control $\langle \nabla^2 g(x,y;\delta), \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle$ along some specific $\boldsymbol{w}\in\mathbb{S}^2$ that will prove useful in later developments.

Corollary B.6 *Let $\delta>0$ be a constant and $\boldsymbol{w}:=(10,11)^{\mathrm{T}}/\sqrt{221}\in\mathbb{S}^2$. We have*

$$\begin{cases} \langle \boldsymbol{H}(x,y;\delta), \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle \le -0.0092714 & \text{for all } x,y\in\mathbb{R} \text{ such that } \sqrt{x^2+y^2}\le\delta, \\ \langle -\boldsymbol{I}, \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle = -1 \le -0.0092714 & \text{for all } x,y\in\mathbb{R} \text{ such that } \sqrt{x^2+y^2}\ge\delta. \end{cases} \tag{7}$$

As a consequence, we have $\langle \nabla^2 g(x,y;\delta), \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle \le -0.0092714$ for all $x,y\in\mathbb{R}$.

We remark that the said consequence is immediate from the C^3-smoothness of $g(\bullet,\bullet;\delta)$.

Proof. As the claim $\langle -\boldsymbol{I}, \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle = -1$ is trivial, we thus focus on the remaining claim in (7). Indeed, it follows from Propositions B.3, B.4, and B.5 that for all $x,y\in\mathbb{R}$ with $\sqrt{x^2+y^2}\le\delta$,

$$\begin{aligned} \langle \boldsymbol{H}(x,y;\delta), \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle &= \frac{1}{221}\big(100h_{11}(x,y;\delta)+220h_{12}(x,y;\delta)+121h_{22}(x,y;\delta)\big) \\ &\le \frac{1}{221}\left(100\left(-\frac{71}{98}\right)+220\cdot 0.32\right) \le -0.0092714, \end{aligned}$$

as desired. The proof is complete. □

B.1.3 Lipschitz modulus estimations

We next turn to estimate $\mathrm{lip}\nabla g(\bullet,\bullet;\delta)$.

Corollary B.7 *Let $\delta > 0$ be a constant. We have*

$$\begin{cases} \|\boldsymbol{H}(x,y;\delta)\|_\sigma \le 4.086 & \text{for all } x,y\in\mathbb{R} \text{ such that } \sqrt{x^2+y^2}\le\delta, \\ \|-\boldsymbol{I}\|_\sigma = 1 \le 4.086 & \text{for all } x,y\in\mathbb{R} \text{ such that } \sqrt{x^2+y^2}\ge\delta. \end{cases}$$

As a consequence, we have $\mathrm{lip}\nabla g(x,y;\delta)\le 4.086$ for all $x,y\in\mathbb{R}$.

We note that the said consequence is immediate from (Rockafellar & Wets, 2009, Theorem 9.7).

Proof. Recall that $\|\boldsymbol{X}\|_\sigma \le \|\boldsymbol{X}\|_1$ for every matrix $\boldsymbol{X}$; see, e.g., (Chen & Li, 2020, Proposition 4.2). Hence, it follows from Propositions B.3, B.4, and B.5 that for all $x,y\in\mathbb{R}$ with $\sqrt{x^2+y^2}\le\delta$,

$$\begin{aligned} \|\boldsymbol{H}(x,y;\delta)\|_\sigma \le \|\boldsymbol{H}(x,y;\delta)\|_1 &= \big|h_{11}(x,y;\delta)\big| + 2\big|h_{12}(x,y;\delta)\big| + \big|h_{22}(x,y;\delta)\big| \\ &\le \frac{91}{64} + 2\cdot 0.32 + \frac{253}{125} = 4.085875 \le 4.086, \end{aligned}$$

as desired. The proof is complete as the other claim is trivial. □

B.2 Preparation: The ambient function

An essential object that has appeared in preceding discussions but has not yet been made precise is the ambient function that is locally identical to $-2^{-1}\|\bullet\|^2$. In fact, this function is constructed by horizontally gluing shifted copies of the (cross-supported) local ambient function introduced below, with one copy placed at each iterate; recall from Section 4 that all the iterates lie on the x-axis.

To begin, consider the function $q(\bullet;\gamma):\mathbb{R}\to\mathbb{R}$ defined by

$$q(x;\gamma) := \begin{cases} -\dfrac{1}{2}x^2, & |x|\le\dfrac{\gamma}{4}, \\ \dfrac{1}{16}(-8\gamma|x|+\gamma^2+8x^2), & \dfrac{\gamma}{4}\le|x|\le\dfrac{3\gamma}{4}, \\ -\dfrac{1}{2}(\gamma-|x|)^2, & \dfrac{3\gamma}{4}\le|x|\le\gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where $\gamma>0$. This function is $C^{1,1}$, and it can be directly calculated that

$$q'(x;\gamma) = \begin{cases} -x, & |x|\le\dfrac{\gamma}{4}, \\ x-\dfrac{\gamma}{2}\mathrm{sign}(x), & \dfrac{\gamma}{4}\le|x|\le\dfrac{3\gamma}{4}, \\ \gamma\mathrm{sign}(x)-x, & \dfrac{3\gamma}{4}\le|x|\le\gamma, \\ 0, & \text{otherwise}, \end{cases} \quad\text{and}\quad q''(x;\gamma) = \begin{cases} -1, & |x|<\dfrac{\gamma}{4}, \\ 1, & \dfrac{\gamma}{4}<|x|<\dfrac{3\gamma}{4}, \\ -1, & \dfrac{3\gamma}{4}<|x|<\gamma, \\ 0, & |x|>\gamma. \end{cases} \tag{9}$$

With $q(\bullet;\gamma)$ in place, we are ready to introduce the local ambient function $h(\bullet,\bullet;\gamma):\mathbb{R}^2\to\mathbb{R}$ defined by

$$h(x,y;\gamma) := q(x;\gamma)+q(y;1), \tag{10}$$

which is (trivially) $C^{1,1}$ as well. We refer to this function as "ambient" as it satisfies that

$$h(x,y;\gamma) = -\frac{1}{2}x^2-\frac{1}{2}y^2 = -2^{-1}\|\bullet\|^2(x,y) \quad \text{for all } (x,y)^{\mathrm{T}}\in\left[-\frac{\gamma}{4},\frac{\gamma}{4}\right]\times\left[-\frac{1}{4},\frac{1}{4}\right]. \tag{11}$$

As in Section B.1, we shall proceed with the refutation and smoothness properties of $h(\bullet,\bullet;\gamma)$.

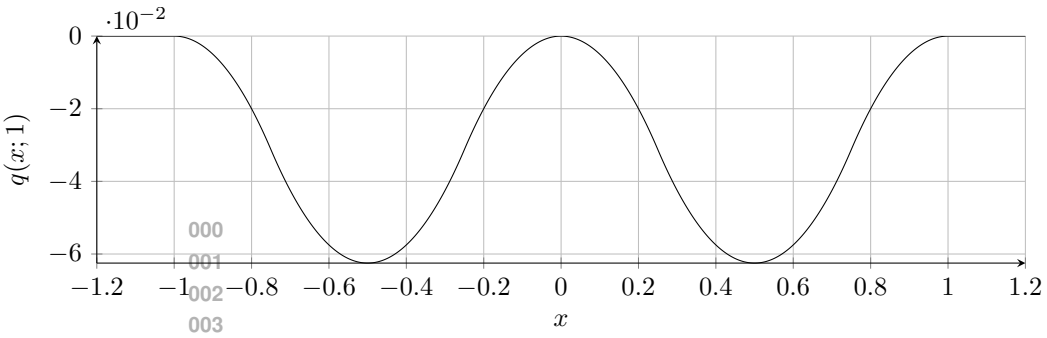

Figure 7: The landscape of $q(\bullet;1)$.

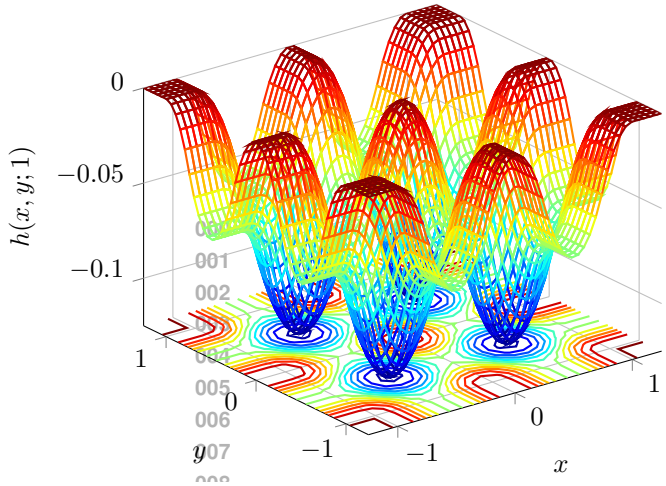

Figure 8: The landscape of $h(\bullet,\bullet;1)$.

B.2.1 Stationarity refutations

By construction, the function $h(\bullet,\bullet;\gamma)$ enjoys a similar stationarity refutation to Corollary B.6.

Proposition B.8 *Let $\gamma > 0$ be a constant and $\boldsymbol{w} := (10,11)^{\mathrm{T}}/\sqrt{221} \in \mathbb{S}^2$. We have*

$$\max\{\langle \boldsymbol{z}, \boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 h(x,y;\gamma)(\boldsymbol{w})\} \leq -0.095 \quad \textit{for all } (x,y)^{\mathrm{T}} \in \mathbb{R} \times \left(-\frac{1}{4},\frac{1}{4}\right).$$

Proof. As $h(\bullet,\bullet;\gamma)$ carries a separable structure, for all $x \in \mathbb{R}$ and $y \in (-4^{-1},4^{-1})$, we have

$$\begin{aligned}\limsup_{\substack{(x',y')^{\mathrm{T}} \to (x,y)^{\mathrm{T}} \\ (x',y')^{\mathrm{T}} \in \mathbb{D}\times(-1/4,1/4)}} \{\nabla^2 h(x',y';\gamma)\} &= \limsup_{\substack{(x',y')^{\mathrm{T}} \to (x,y)^{\mathrm{T}} \\ (x',y')^{\mathrm{T}} \in \mathbb{D}\times(-1/4,1/4)}} \left\{\begin{pmatrix} q''(x';\gamma) & 0 \\ 0 & -1 \end{pmatrix}\right\} \\ &= \begin{pmatrix} \limsup_{\substack{x' \to x \\ x' \in \mathbb{D}}} \{q''(x';\gamma)\} & 0 \\ 0 & -1 \end{pmatrix},\end{aligned}$$

where $\mathbb{D} \subseteq \mathbb{R}$ is the set of points at which $q(\bullet;\gamma)$ is twice differentiable. By taking convex hulls on both sides, it further follows that

$$\partial_{\mathrm{C}}^2 h(x,y;\gamma) = \begin{pmatrix} \partial_{\mathrm{C}}^2 q(x;\gamma) & 0 \\ 0 & -1 \end{pmatrix} \quad \text{for all } (x,y)^{\mathrm{T}} \in \mathbb{R} \times \left(-\frac{1}{4},\frac{1}{4}\right).$$

However, we know from (9) that $\partial_{\mathrm{C}}^2 q(x;\gamma) \subseteq [-1,1]$ for all $x \in \mathbb{R}$. As a result, it follows that

$$\begin{aligned}\max\big\{\langle \boldsymbol{z}, \boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 h(x,y;\gamma)(\boldsymbol{w})\big\} &= \max\left\{\frac{1}{221}(100a-121) : a \in \partial_{\mathrm{C}}^2 q(x;\gamma)\right\}\\ &\le \max\left\{\frac{1}{221}(100a-121) : a \in [-1,1]\right\}\\ &= -\frac{21}{221} \le -0.095 \quad \text{for all } (x,y)^{\mathrm{T}} \in \mathbb{R} \times \left(-\frac{1}{4}, \frac{1}{4}\right),\end{aligned}$$

as desired. The proof is complete. □

B.2.2 Lipschitz modulus estimations

We next take some time to understand $\mathrm{lip}\nabla h(\bullet,\bullet;\gamma)$.

Lemma B.9 *Let $\gamma > 0$ be a constant. We have $\mathrm{lip}\nabla h(x,y;\gamma) \le 1$ for all $x, y \in \mathbb{R}$.*

Proof. As $q'(\bullet;\gamma)$ is 1-Lipschitz continuous by (Rockafellar & Wets, 2009, Theorem 9.13), it immediately follows that

$$\begin{aligned}\|\nabla h(x_1,y_1;\gamma) - \nabla h(x_2,y_2;\gamma)\| &= \left\|\begin{pmatrix} q'(x_1;\gamma)\\ q'(y_1;1)\end{pmatrix} - \begin{pmatrix} q'(x_2;\gamma)\\ q'(y_2;1)\end{pmatrix}\right\|\\ &= \sqrt{\big(q'(x_1;\gamma) - q'(x_2;\gamma)\big)^2 + \big(q'(y_1;1) - q'(y_2;1)\big)^2}\\ &\le \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2} = \left\|\begin{pmatrix} x_1\\ y_1\end{pmatrix} - \begin{pmatrix} x_2\\ y_2\end{pmatrix}\right\|;\end{aligned}$$

i.e., $\nabla h(\bullet,\bullet;\gamma)$ is 1-Lipschitz continuous as well, from which the desired claim directly follows. □

B.3 Proof of Theorem 3.1

In this subsection, we prove Theorem 3.1 in a discussion manner. In what follows, we shall restrict attention to the case $n = 2$; for $n \ge 3$, it suffices to extend the construction to be given for $n = 2$ in the canonical way. Let $m \ge 1$ be arbitrary.

To begin, consider the resisting oracle $\boldsymbol{\mathcal{O}} : \mathbb{R}^2 \to \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}^{2\times 2}_{\mathrm{sym}}$ defined by

$$\boldsymbol{\mathcal{O}}(\boldsymbol{x}) := \left(0, \boldsymbol{0}, \begin{pmatrix} -1 & 0\\ 0 & 0\end{pmatrix}\right) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^2$$

and an arbitrary $\boldsymbol{\mathcal{A}} \in \mathbb{A}^2_{\text{det-zr}}(m)$. Let $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ be the iterates generated by $\boldsymbol{\mathcal{A}}$ after m rounds of interaction with $\boldsymbol{\mathcal{O}}$. By the zero-respecting property of $\boldsymbol{\mathcal{A}}$, we have $\boldsymbol{x}_0 = \boldsymbol{0}$ and $\mathrm{supp}(\boldsymbol{x}_k) \subseteq \{1\}$ for all $k = 0, \ldots, m$; i.e., $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ lie on the x-axis. Without loss of generality, we may assume that $\boldsymbol{x}_i \ne \boldsymbol{x}_j$ for all $i \ne j$, and, by relabeling (and abusing notation), that $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ are sorted in ascending order of their x-coordinates. In particular, this implies that $\boldsymbol{x}_0$ may no longer be $\boldsymbol{0}$ anymore. For ease of exposition, we also define $\boldsymbol{x}_{-1} := (-\infty, 0)^{\mathrm{T}}$ and $\boldsymbol{x}_{m+1} = (\infty, 0)^{\mathrm{T}}$, and with a further abuse of notation, we denote by x_k the x-coordinate of $\boldsymbol{x}_k$ for $k = -1, \ldots, m+1$; e.g., $x_{-1} = -\infty$. In what follows, as discussed in Section 4, we shall construct a hard function that is compatible with $\boldsymbol{\mathcal{O}}$ along $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ and possesses all the desired properties stated in Theorem 3.1.

Define

$$\zeta := \min\{\xi, 1\} \in (0,1], \quad \text{where } \xi := \min\big\{|x_i - x_j| : i, j = 0, \ldots, m \text{ with } i \ne j\big\} > 0.$$

Consider the hard function $f(\bullet,\bullet;\zeta) : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y;\zeta) := \begin{cases} f_k(x,y;\zeta), & |x - x_k| \le \zeta/4,\\ q(y;1), & \text{otherwise,}\end{cases} \tag{12}$$

where for $k = 0, \dots, m$,

$$f_k(x, y; \zeta) := p_2(x - x_k, y; \zeta/32) \cdot \left(-\frac{1}{2}(x - x_k)^2 \right) + \big(1 - p_2(x - x_k, y; \zeta/32)\big) \cdot h(x - x_k, y; \zeta/4).$$

We begin the analysis of $f(\bullet, \bullet; \zeta)$ with a basic observation on the piecewise structure of $f_k(\bullet, \bullet; \zeta)$, which makes clear that it results from the local surgery performed on the local ambient function.

Lemma B.10 *For all $k = 0, \dots, m$, we have*

$$f_k(x, y; \zeta) = \begin{cases} h(x - x_k, y; \zeta/4), & (x, y)^{\mathrm{T}} \notin (x_k, 0)^{\mathrm{T}} + \dfrac{\zeta}{32}\mathbb{B}^2, \\ g(x - x_k, y; \zeta/32), & (x, y)^{\mathrm{T}} \in \left[x_k - \dfrac{\zeta/4}{4}, x_k + \dfrac{\zeta/4}{4} \right] \times \left[-\dfrac{1}{4}, \dfrac{1}{4} \right]. \end{cases} \tag{13}$$

Before presenting the proof, for convenience, let us introduce for all $k = 0, \dots, m$ the shorthand

$$\mathbb{B}_k := (x_k, 0)^{\mathrm{T}} + \frac{\zeta}{32}\mathbb{B}^2 \subseteq (x_k, 0)^{\mathrm{T}} + \frac{1}{2} \cdot \left(\left[-\frac{\zeta/4}{4}, \frac{\zeta/4}{4} \right] \times \left[-\frac{1}{4}, \frac{1}{4} \right] \right), \tag{14}$$

where the containment follows, as $\zeta \in (0, 1]$, from

$$\sqrt{x^2 + y^2} \le \frac{\zeta}{32} \implies \left(|x| \le \frac{\zeta}{32} = \frac{1}{2} \cdot \frac{\zeta/4}{4} \quad \text{and} \quad |y| \le \frac{\zeta}{32} \le \frac{1}{32} \le \frac{1}{2} \cdot \frac{1}{4} \right).$$

Proof. If $(x, y)^{\mathrm{T}} \notin \mathbb{B}_k$ (i.e., $\sqrt{(x - x_k)^2 + y^2} > \zeta/32$), then we have $p_2(x - x_k, y; \zeta/32) = 0$ by (4), and hence $f_k(x, y; \zeta) = h(x - x_k, y; \zeta/4)$, as desired. If $(x, y)^{\mathrm{T}} \in [x_k - \zeta/16, x_k + \zeta/16] \times [-1/4, 1/4]$, then it follows from (11) that $h(x - x_k, y; \zeta/4) = -2^{-1}(x - x_k)^2 - 2^{-1}y^2$, and hence

$$\begin{aligned} f_k(x, y; \zeta) &:= p_2(x - x_k, y; \zeta/32) \cdot \left(-\frac{1}{2}(x - x_k)^2 \right) \\ &\quad + \big(1 - p_2(x - x_k, y; \zeta/32)\big) \cdot \left(-\frac{1}{2}(x - x_k)^2 - \frac{1}{2}y^2 \right) = g(x - x_k, y; \zeta/32), \end{aligned}$$

as desired. The proof is complete. □

With Lemma B.10 in place, we are now ready to derive a piecewise representation of $f(\bullet, \bullet; \zeta)$, revealing that it results from the local surgeries performed on the ambient function around every $\boldsymbol{x}_k$.

Proposition B.11 *Consider the open sets $\mathbb{H}_0, \dots, \mathbb{H}_m$ and $\mathbb{P}_0, \dots, \mathbb{P}_m$ defined by*

$$\mathbb{H}_k := \left(\left(x_{k-1} + \frac{\zeta}{4}, x_{k+1} - \frac{\zeta}{4} \right) \times \mathbb{R} \right) \cap \mathbb{B}_k^{\complement} \quad \text{and} \quad \mathbb{P}_k := \left(x_k - \frac{\zeta/4}{4}, x_k + \frac{\zeta/4}{4} \right) \times \left(-\frac{1}{4}, \frac{1}{4} \right).$$

Their union constitutes an open cover of $\mathbb{R}^2$; i.e., $\bigcup_{k=0}^m \mathbb{H}_k \cup \bigcup_{k=0}^m \mathbb{P}_k = \mathbb{R}^2$. Besides, we have

$$f(x, y; \zeta) = \begin{cases} h(x - x_k, y; \zeta/4), & (x, y)^{\mathrm{T}} \in \mathbb{H}_k, \\ g(x - x_k, y; \zeta/32), & (x, y)^{\mathrm{T}} \in \mathbb{P}_k. \end{cases}$$

Proof. We shall first prove $\bigcup_{k=0}^m \mathbb{H}_k \cup \bigcup_{k=0}^m \mathbb{P}_k = \mathbb{R}^2$; as $\mathbb{R}^2 = \left(\bigcup_{k=0}^m \mathbb{B}_k \right) \cup \left(\bigcup_{k=0}^m \mathbb{B}_k \right)^{\complement}$, it suffices to cover these two sets by $\bigcup_{k=0}^m \mathbb{P}_k$ and $\bigcup_{k=0}^m \mathbb{H}_k$, respectively. By (14), we have $\mathbb{B}_k \subseteq \mathbb{P}_k$ for all $k = 0, \dots, m$, and hence $\bigcup_{k=0}^m \mathbb{B}_k \subseteq \bigcup_{k=0}^m \mathbb{P}_k$; it remains to show $\left(\bigcup_{k=0}^m \mathbb{B}_k \right)^{\complement} \subseteq \bigcup_{k=0}^m \mathbb{H}_k$. To this end, let $(x, y)^{\mathrm{T}} \in \left(\bigcup_{k=0}^m \mathbb{B}_k \right)^{\complement}$ be arbitrary; i.e., we have $(x, y)^{\mathrm{T}} \notin \mathbb{B}_k$ for all $k = 0, \dots, m$. Suppose that $|x - x_{k'}| \le \zeta/4$ for some $k' = 0, \dots, m$. Then, we must have $x \in (x_{k'-1} + \zeta/4, x_{k'+1} - \zeta/4)$, and hence $(x, y)^{\mathrm{T}} \in \mathbb{H}_{k'} \subseteq \bigcup_{k=0}^m \mathbb{H}_k$, as desired.

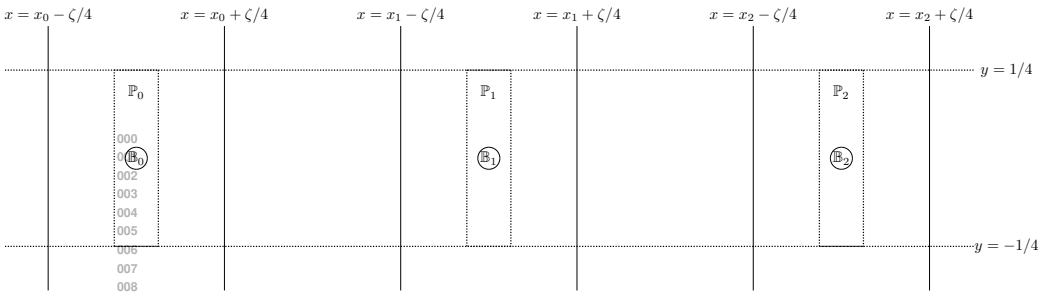

Figure 9: An illustration of $\mathbb{B}_k$, $\mathbb{P}_k$, and (implicitly) $\mathbb{H}_k$ for $k = 0, \ldots, m$ when $m = 2$ and $\zeta = 1$.

On the other hand, suppose that $|x - x_k| > \zeta/4$ for all $k = 0, \ldots, m$. Then, there must exist some $k' = 0, \ldots, m$ such that

$$x \in (x_{k'-1} + \zeta/4, x_{k'} - \zeta/4) \cup (x_{k'} + \zeta/4, x_{k'+1} - \zeta/4) \subseteq (x_{k'-1} + \zeta/4, x_{k'+1} - \zeta/4);$$

i.e., $(x, y)^{\mathrm{T}} \in \mathbb{H}_{k'} \subseteq \bigcup_{k=0}^{m} \mathbb{H}_k$, as desired. This establishes the covering property.

We next turn to study the piecewise behavior of $f(\bullet, \bullet; \zeta)$. Suppose that $(x, y)^{\mathrm{T}} \in \mathbb{H}_k$ for some $k = 0, \ldots, m$. Then, we divide the subsequent discussions into two cases.

- If $|x - x_k| \leq \zeta/4$, then we know by (12) that $f(x, y; \zeta) = f_k(x, y; \zeta)$. However, as $(x, y)^{\mathrm{T}} \notin \mathbb{B}_k$ by the definition of $\mathbb{H}_k$, it also follows from (13) that $f_k(x, y; \zeta) = h(x - x_k, y; \zeta/4)$. By putting the above two pieces together, we have $f(x, y; \zeta) = h(x - x_k, y; \zeta/4)$, as desired.
- If $|x - x_k| > \zeta/4$, then we know by the definition of $\mathbb{H}_k$ that

  $$x \in (x_{k-1} + \zeta/4, x_k - \zeta/4) \cup (x_k + \zeta/4, x_{k+1} - \zeta/4);$$

  as a result, there does not exist any $k' = 0, \ldots, m$ such that $|x - x_{k'}| \leq \zeta/4$. This, together with (12), implies that $f(x, y; \zeta) = q(y; 1)$. However, as $|x - x_k| > \zeta/4$, we also know by (10) and (8) that $h(x - x_k, y; \zeta/4) = q(x - x_k; \zeta/4) + q(y; 1) = q(y; 1)$. Hence, we have $f(x, y; \zeta) = h(x - x_k, y; \zeta/4)$, as desired.

On the other hand, suppose that $(x, y)^{\mathrm{T}} \in \mathbb{P}_k$ for some $k = 0, \ldots, m$. Then, we have $|x - x_k| \leq \zeta/16 \leq \zeta/4$, and hence $f(x, y; \zeta) = f_k(x, y; \zeta)$ by (12). However, by combining the definition of $\mathbb{P}_k$ and (13), we also know that $f_k(x, y; \zeta) = g(x - x_k, y; \zeta/32)$. Hence, we have $f(x, y; \zeta) = g(x - x_k, y; \zeta/32)$, as desired. The whole proof is now complete. □

As an immediate consequence of Proposition B.11, some desired properties of $f(\bullet, \bullet; \zeta)$ follow.

Corollary B.12 *The following two statements hold.*

- *The function $f(\bullet, \bullet; \zeta)$ is C^1 with $\nabla f(\bullet, \bullet; \zeta)$ being 4.086-Lipschitz continuous.*
- *The function $f(\bullet, \bullet; \zeta)$ is compatible with $\boldsymbol{\mathcal{O}}$ along $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$, and is C^3 around $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$.*

Proof. Let $(x, y)^{\mathrm{T}} \in \mathbb{R}^2$ be arbitrary. We know from Proposition B.11 that $(x, y)^{\mathrm{T}}$ must belong to some $\mathbb{H}_k$ or $\mathbb{P}_k$. If $(x, y)^{\mathrm{T}} \in \mathbb{H}_k$, then it follows from the facts that $\mathbb{H}_k$ is open and $h(\bullet - x_k, \bullet; \zeta/4)$ is C^1, Proposition B.11, and Lemma B.9 that $f(\bullet, \bullet; \zeta)$ is C^1 around $(x, y)^{\mathrm{T}}$ with

$$\mathrm{lip}\nabla f(x, y; \zeta) = \mathrm{lip}\nabla h(x - x_k, y; \zeta/4) \leq 1 \leq 4.086;$$

recall from (Rockafellar & Wets, 2009, Definition 9.1(b)) that $\mathrm{lip}\nabla f(\bullet, \bullet; \zeta)$ is defined only locally. If $(x, y)^{\mathrm{T}} \in \mathbb{P}_k$, then it follows from the facts that $\mathbb{P}_k$ is open and $g(\bullet - x_k, \bullet; \zeta/32)$ is C^3, Proposition B.11, and Corollary B.7 that $f(\bullet, \bullet; \zeta)$ is C^1 around $(x, y)^{\mathrm{T}}$ with

$$\mathrm{lip}\nabla f(x, y; \zeta) = \mathrm{lip}\nabla g(x - x_k, y; \zeta/32) \leq 4.086.$$

In summary, it follows that $f(\bullet,\bullet;\zeta)$ is C^1 with $\mathrm{lip}\nabla f(x,y;\zeta) \leq 4.086$ for all $x,y \in \mathbb{R}$; i.e., $\nabla f(\bullet,\bullet;\zeta)$ is Lipschitz continuous with constant 4.086 (Rockafellar & Wets, 2009, Theorem 9.2), as desired.

To see the other claim, it suffices to combine (12) and (13) with Lemma B.1 and the C^3-smoothness of $g(\bullet,\bullet;\delta)$. The proof is complete. □

Another consequence of Proposition B.11 is a strong stationarity refutation property of $f(\bullet,\bullet;\zeta)$.

Corollary B.13 (Band refutation) *Let* $\boldsymbol{w} := (10,11)^{\mathrm{T}}/\sqrt{221} \in \mathbb{S}^2$. *We have*

$$\max\left\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \mathrm{conv}\left(\bigcup_{(x,y)^{\mathrm{T}}\in\mathbb{R}\times(-1/4,1/4)} \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\right)\right\} \leq -0.0092714.$$

Proof. Let $\boldsymbol{z} \in \mathrm{conv}\big(\bigcup_{(x,y)^{\mathrm{T}}\in\mathbb{R}\times(-1/4,1/4)} \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\big)$ be arbitrary. By the definition of convex hulls, we have $\boldsymbol{z} = \sum_{k'=1}^{m'} t_{k'} \cdot \boldsymbol{z}_{k'}$ for some $m' \geq 1$, $\boldsymbol{z}_{k'} \in \bigcup_{(x,y)^{\mathrm{T}}\in\mathbb{R}\times(-1/4,1/4)} \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})$, and $t_{k'} \geq 0$ with $\sum_{k'=1}^{m'} t_{k'} = 1$. Besides, we also know by the definition of $\boldsymbol{z}_{k'}$ that there further exist some $x'_{k'} \in \mathbb{R}$ and $y'_{k'} \in (-1/4,1/4)$ such that $\boldsymbol{z}_{k'} \in \partial_{\mathrm{C}}^2 f(x'_{k'},y'_{k'};\zeta)(\boldsymbol{w})$. Hence, it follows that

$$\langle \boldsymbol{z},\boldsymbol{w}\rangle = \sum_{k'=1}^{m'} t_{k'} \cdot \langle \boldsymbol{z}_{k'},\boldsymbol{w}\rangle \leq \sum_{k'=1}^{m'} t_{k'} \cdot \max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x'_{k'},y'_{k'};\zeta)(\boldsymbol{w})\big\};$$

this motivates us to study $\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x'_{k'},y'_{k'};\zeta)(\boldsymbol{w})\big\}$ instead. However, as $x'_{k'} \in \mathbb{R}$ and $y'_{k'} \in (-1/4,1/4)$, it further suffices to study $\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\big\}$ for some arbitrary $x \in \mathbb{R}$ and $y \in (-1/4,1/4)$.

Let $(x,y)^{\mathrm{T}} \in \mathbb{R}\times(-1/4,1/4)$ be arbitrary. We know from Proposition B.11 that $(x,y)^{\mathrm{T}}$ must belong to some $\mathbb{H}_k$ or $\mathbb{P}_k$. If $(x,y)^{\mathrm{T}} \in \mathbb{H}_k$, then it follows from Propositions B.11 and B.8 that

$$\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\big\} = \max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 h(x-x_k,y;\zeta/4)(\boldsymbol{w})\big\} \leq -0.095.$$

If $(x,y)^{\mathrm{T}} \in \mathbb{P}_k$, then it follows from Proposition B.11 and Corollary B.6 that

$$\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\big\} = \langle \nabla^2 g(x-x_k,y;\zeta/32),\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle \leq -0.0092714.$$

By combining the above two pieces, we have $\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\big\} \leq -0.0092714$ for all $(x,y)^{\mathrm{T}} \in \mathbb{R}\times(-1/4,1/4)$, which implies that $\max\big\{\langle \boldsymbol{z},\boldsymbol{w}\rangle : \boldsymbol{z} \in \partial_{\mathrm{C}}^2 f(x'_{k'},y'_{k'};\zeta)(\boldsymbol{w})\big\} \leq -0.0092714$ for all $k' = 1,\ldots,m'$, and thus $\langle \boldsymbol{z},\boldsymbol{w}\rangle \leq -0.0092714$, as desired. We are done. □

With the previous two results in place and noting that for all $\boldsymbol{w} \in \mathbb{R}^2$ and $k = 0,\ldots,m$,

$$\mathrm{conv}\left(\bigcup_{(x,y)^{\mathrm{T}}\in(x_k,0)^{\mathrm{T}}+\mathbb{B}^2/8} \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\right) \subseteq \mathrm{conv}\left(\bigcup_{(x,y)^{\mathrm{T}}\in\mathbb{R}\times(-1/4,1/4)} \partial_{\mathrm{C}}^2 f(x,y;\zeta)(\boldsymbol{w})\right),$$

all that remains is to bound the initial optimality gap $f(0,0;\zeta) - \min\{f(x,y;\zeta) : x,y \in \mathbb{R}\}$.

Lemma B.14 *We have* $f(0,0;\zeta) - \min\{f(x,y;\zeta) : x,y \in \mathbb{R}\} \leq 17/256$.

We remark that as $f(0,0;\zeta) = 0$ (as $(0,0)^{\mathrm{T}}$ is some iterate and $f(\bullet,\bullet;\zeta)$ is compatible with $\boldsymbol{\mathcal{O}}$, whose zeroth-order output is always 0, along all the iterates; see Corollary B.12), it suffices to verify that $f(x,y;\zeta) \geq -17/256$ for all $x,y \in \mathbb{R}$.

Proof. Since $f(\bullet,\bullet;\zeta)$ takes values only from $f_k(\bullet,\bullet;\zeta)$ and $q(\bullet;1)$, and $q(x;\gamma) \geq -\gamma^2/16$ for all $x \in \mathbb{R}$, it suffices to bound $f_k(\bullet,\bullet;\zeta)$. Let us divide the subsequent analysis into two cases.

- If $(x, y)^{\mathrm{T}} \notin (x_k, 0)^{\mathrm{T}} + \zeta\mathbb{B}^2/32$, then it immediately follows from (13) and $\zeta \leq 1$ that
$$f_k(x, y; \zeta) = h(x - x_k, y; \zeta/4) \geq -1/256 - 1/16 = -17/256.$$
- If $(x, y)^{\mathrm{T}} \in (x_k, 0)^{\mathrm{T}} + \zeta\mathbb{B}^2/32$, then it follows from (13) and $\zeta \leq 1$ again that
$$\begin{aligned} f_k(x, y; \zeta) &= g(x - x_k, y; \zeta/32) \\ &= p_2(x - x_k, y; \zeta/32) \cdot \left(-\frac{1}{2}(x - x_k)^2\right) \\ &\quad + \big(1 - p_2(x - x_k, y; \zeta/32)\big) \cdot \left(-\frac{1}{2}(x - x_k)^2 - \frac{1}{2}y^2\right) \\ &\geq -\frac{1}{2} \cdot (\zeta/32)^2 - \frac{1}{2} \cdot (\zeta/32)^2 \geq -\frac{1}{1024} \geq -\frac{17}{256}. \end{aligned}$$

In summary, we see that $f_k(x, y; \zeta) \geq -17/256$ for all $x, y \in \mathbb{R}$, and thus $f(x, y; \zeta) \geq -17/256$ for all $x, y \in \mathbb{R}$, as desired. The proof is complete. □

C Proof of Corollary 3.2

Proof of Corollary 3.2. Let $m \geq 1$, $n \geq m+3$, and $\mathcal{A} \in \mathbb{A}^n_{\mathrm{det}}(m)$ be arbitrary. Consider the resisting oracle $\mathcal{O} : \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{n\times n}_{\mathrm{sym}}$ given by $\mathcal{O}(\boldsymbol{x}) := (0, \boldsymbol{0}, -\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}})$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ be the iterates generated by $\mathcal{A}$ after m rounds of interaction with $\mathcal{O}$, and $f_2 : \mathbb{R}^2 \to \mathbb{R}$ be the function constructed in (3) with $(\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{x}_0, 0)^{\mathrm{T}}, \dots, (\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{x}_m, 0)^{\mathrm{T}}$ as the iterates. In addition, let $\boldsymbol{v} \in \mathbb{S}^n \cap \mathrm{span}\{\boldsymbol{e}_1, \boldsymbol{x}_0, \dots, \boldsymbol{x}_m\}^{\perp}$ be arbitrary; this is possible as $n \geq m+3$. As a quick sanity check, for $\boldsymbol{U} := (\boldsymbol{e}_1, \boldsymbol{v}) \in \mathbb{R}^{n\times 2}$, we have $(\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{x}_k, 0)^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k$ for all $k = 0, \dots, m$. We also note that $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$. In the sequel, similar to Section 4, we shall show that the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\boldsymbol{x}) := f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x})$ is compatible with $\mathcal{O}$ along $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$ and possesses all the desired properties stated in Corollary 3.2.

We begin with the oracle compatibility of f along $\boldsymbol{x}_0, \dots, \boldsymbol{x}_m$. Indeed, it follows by combining the local C^3-smoothness and oracle compatibility of f_2 in Corollary 4.4 with the smooth chain rule that

$$\begin{cases} f(\boldsymbol{x}_k) = f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k) = 0, \\ \nabla f(\boldsymbol{x}_k) = \boldsymbol{U}\nabla f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k) = \boldsymbol{U}\boldsymbol{0} = \boldsymbol{0}, \\ \nabla^2 f(\boldsymbol{x}_k) = \boldsymbol{U}\nabla^2 f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k)\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}(-\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}})\boldsymbol{U}^{\mathrm{T}} = -\boldsymbol{e}_1\boldsymbol{e}_1^{\mathrm{T}}, \end{cases} \qquad \text{for all } k = 0, \dots, m,$$

as desired.

We next verify that $f \in \mathbb{F}^n_{C^1}(4.086, 17/256; \mathcal{A})$. To see the first property, it suffices to observe through the $C^{1,1}$-smoothness of f_2 in Corollary 4.4 and the smooth chain rule that

$$\begin{aligned} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| &= \|\boldsymbol{U}\nabla f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}) - \boldsymbol{U}\nabla f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y})\| = \|\nabla f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}) - \nabla f_2(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y})\| \\ &\leq 4.086 \cdot \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}\| \leq 4.086 \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n; \end{aligned}$$

the other two properties are trivial.

Finally, we deal with stationarity refutations. To start, we observe through Corollary 2.9 that for all $\boldsymbol{w} \in \mathbb{R}^n$ and $k = 0, \dots, m$,

$$\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{x}_k+\mathbb{B}^n/8} \partial^2_{\mathrm{C}} f(\boldsymbol{y})(\boldsymbol{w})\right) = \boldsymbol{U}\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k+\mathbb{B}^2/8} \partial^2_{\mathrm{C}} f_2(\boldsymbol{y})(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{w})\right).$$

As a result, for $\boldsymbol{w} := \boldsymbol{U}\boldsymbol{w}_0 \in \mathbb{S}^n$ with $\boldsymbol{w}_0 := (10, 11)^{\mathrm{T}}/\sqrt{221} \in \mathbb{S}^2$, we have

$$\begin{aligned} \max_{\boldsymbol{z}\in\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{x}_k+\mathbb{B}^n/8} \partial^2_{\mathrm{C}} f(\boldsymbol{y})(\boldsymbol{w})\right)} \langle \boldsymbol{z}, \boldsymbol{w}\rangle &= \max_{\boldsymbol{z}\in\boldsymbol{U}\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k+\mathbb{B}^2/8} \partial^2_{\mathrm{C}} f_2(\boldsymbol{y})(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{w})\right)} \langle \boldsymbol{z}, \boldsymbol{w}\rangle \\ &= \max_{\boldsymbol{z}\in\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_k+\mathbb{B}^2/8} \partial^2_{\mathrm{C}} f_2(\boldsymbol{y})(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{w})\right)} \langle \boldsymbol{z}, \boldsymbol{U}^{\mathrm{T}}\boldsymbol{w}\rangle \\ &= \max_{\boldsymbol{z}\in\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in(\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{x}_k, 0)^{\mathrm{T}}+\mathbb{B}^2/8} \partial^2_{\mathrm{C}} f_2(\boldsymbol{y})(\boldsymbol{w}_0)\right)} \langle \boldsymbol{z}, \boldsymbol{w}_0\rangle \\ &\leq \max_{\boldsymbol{z}\in\mathrm{conv}\left(\bigcup_{\boldsymbol{y}\in\mathbb{R}\times(-1/4,1/4)} \partial^2_{\mathrm{C}} f_2(\boldsymbol{y})(\boldsymbol{w}_0)\right)} \langle \boldsymbol{z}, \boldsymbol{w}_0\rangle \\ &\leq -0.0092714 \quad \text{for all } k = 0, \dots, m, \end{aligned}$$

where the last inequality is due to the band refutation of f_2 in Corollary 4.4. The proof is complete. □

D Conjecture on extension to local oracles

At first glance, it appears rather straightforward to extend Theorem 3.1 and Corollary 3.2 from second-order oracles to local oracles (see, e.g., (Tian & So, 2024, Section 2.3)): It should suffice to slightly flatten the landscape of $p_1(\bullet;\delta)$ near 0; e.g., one may consider the variant $\overline{p}_1(\bullet;\delta,\lambda):\mathbb{R}\to\mathbb{R}$ of $p_1(\bullet;\delta)$ defined by

$$\overline{p}_1(x;\delta,\lambda) := \begin{cases} 1, & |x|\leq \lambda\delta, \\ \left(\dfrac{(|x|-\lambda\delta)^2}{(\delta-\lambda\delta)^2}-1\right)^4, & \lambda\delta\leq|x|\leq\delta, \\ 0, & \text{otherwise}, \end{cases}$$

where $\lambda\in(0,1)$ determines the extent of such a flattening.

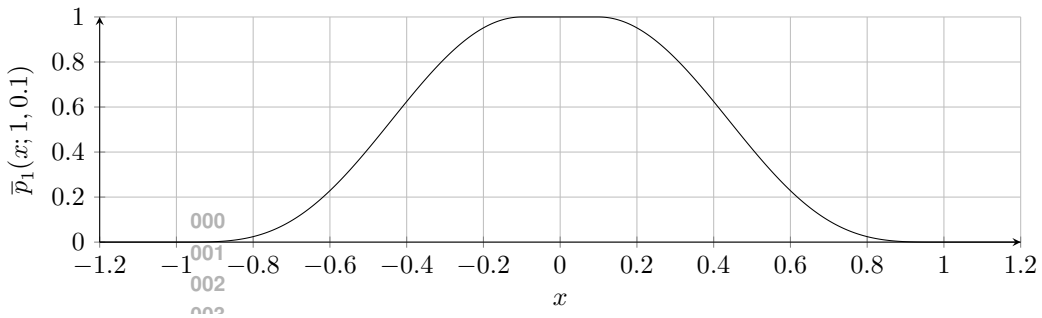

Figure 10: The landscape of $\overline{p}_1(\bullet;1,0.1)$.

Conceivably, as soon as λ is sufficiently small, $\overline{p}_1(\bullet;\delta,\lambda)$ should almost coincide with $p_1(\bullet;\delta)$, thereby inheriting all its desired properties; in fact, numerical evidence suggests that setting $\lambda=0.1$ already suffices by far. However, as it turns out, the analysis of the $\overline{p}_1(\bullet;\delta,\lambda)$-based construction is entirely a nightmare to carry out. This is mainly due to the fact that the corresponding $\overline{p}_2(\bullet,\bullet;\delta,\lambda)$ defined by $\overline{p}_2(x,y;\delta,\lambda) := \overline{p}_1\big(\sqrt{x^2+y^2};\delta,\lambda\big)$ is no longer a piecewise polynomial as $p_2(\bullet,\bullet;\delta)$, but rather a piecewise polynomial in *norms*. That said, we remain firmly convinced that the extension is valid.

Conjecture D.1 *Theorem 3.1 and Corollary 3.2 remain valid (in substance) even under local oracles.*