On computing Goldstein approximate second-order stationary points of structured nonsmooth nonconvex programs

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we exhibit a randomized first-order algorithm to compute Goldstein approximate second-order stationary points of L-smooth functions, using tools from randomized smoothing. The algorithm has oracle complexity $\widetilde{O}(n^2/\varepsilon^9 + n^3/\varepsilon^7)$, where $n = 1, 2, \ldots$ is the input dimension and $\varepsilon > 0$ is the (common) tolerance. We also present extensions to weakly convex functions and applications to bilevel optimization.

1 Introduction

For nonconvex functions, computing second-order stationary points (SOSPs), or equivalently, saddle escape, is important. From first principles, they are the natural next target beyond (first-order) stationary points (FOSPs) in the stationarity hierarchy FOSPs $\supseteq$ SOSPs $\supseteq \cdots$, thereby offering the promise of solutions with higher quality. From a practical point of view, many problems arising in applications admit no spurious SOSPs (i.e., every such point is globally optimal; e.g., low-rank matrix recovery (Bhojanapalli et al., 2016), phase retrieval (Sun et al., 2018), and neural collapse (Zhu et al., 2021)), which implies that any algorithm with convergence to SOSPs will avoid all spurious FOSPs and converge to global optimality.

The computational landscape of SOSPs of *smooth* nonconvex functions is by now well understood. The cubic Newton (Griewank, 1981; Nesterov & Polyak, 2006; Cartis et al., 2011) and trust-region (Conn et al., 2000; Curtis et al., 2018; Jiang et al., 2026) methods, e.g., offer deterministic, dimension-free guarantees. Moreover, even gradient descent can asymptotically avoid strict saddle points (Lee et al., 2016), while random perturbations are commonly employed to obtain finite-time guarantees with polynomial or even polylogarithmic dependence on the dimension (Ge et al., 2015; Jin et al., 2017). However, once *nonsmoothness* enters the picture, the situation becomes rather different. While some recent advances show certain computability of SOSPs for weakly convex functions and beyond under assumptions such as active manifolds and linear tilting (see, e.g., (Davis & Drusvyatskiy, 2022; Bianchi et al., 2024; Davis et al., 2026)), without such assumptions, the recent work of Guan & So (2026) shows that even for L-smooth functions, even a highly permissive target, namely Goldstein SOSPs, already faces fundamental barriers to deterministic computations. In particular, the complexity of any such algorithm must be $\Omega(n)$, where $n = 1, 2, \ldots$ is the input dimension. This naturally leads one to ask what guarantees randomization can offer.

In this paper, while not bypassing the lower bounds, we establish the randomized upper bound $\widetilde{O}\big(\Delta L^8 n^2/\varepsilon^9 + \Delta L^6 n^3/\varepsilon^7\big)$ for the problem, where $\Delta > 0$ is the initial optimality gap and $\varepsilon > 0$ is the (common) tolerance, via (uniform) randomized smoothing, a technique that has recently been shown to be effective for finding Goldstein FOSPs of Lipschitz functions; see (Lin et al., 2022) and also, e.g., (Kornowski & Shamir, 2024; Xia et al., 2025; Lei et al., 2025) for follow-up works. To be specific, just as Lin et al. (2022) show that randomized smoothing smooths a Lipschitz function into an L-smooth function (Lin et al., 2022, Proposition 2.3) and that an FOSP of the smoothed function transfers to a Goldstein FOSP of the original function (Lin et al., 2022, Theorem 3.1), we show that the same technique likewise smooths an L-smooth function into a function with Lipschitz Hessian, together with a similar SOSP

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.

transfer property. With these in place, together with some standard estimators for the gradient and Hessian of the smoothed function, the desired algorithm then follows by applying the cubic Newton method to the smoothed function. In passing, we note that the algorithm is entirely first-order, in a similar spirit to (Lin et al., 2022, Algorithm 1), thereby obviating the need for cumbersome (generalized) Hessian manipulations. We also note the closely related recent work of Lei et al. (2026), which presents a Gaussian-smoothing-based zeroth-order method for computing Goldstein approximate SOSPs of L-smooth functions. Both their algorithm and ours employ randomized smoothing and cubic regularization, but they use different smoothing distributions and operate under different oracle models: Lei et al. (2026) use Gaussian smoothing and function-value information, while we use uniform-ball smoothing and first-order information. The two approaches thus offer complementary perspectives on the computation of Goldstein approximate SOSPs.

Building on our algorithm, we then turn to study the computation of Goldstein SOSPs of other structured functions beyond the L-smooth setting. In particular, we propose a new and proper notion of Goldstein SOSPs for weakly convex functions via the Moreau envelope (Moreau, 1965), which is nontrivial as a naïve extension of the notion of Clarke SOSPs (Hiriart-Urruty et al., 1984, Theorem 3.1) to such functions fails to be necessary for optimality. We then develop an algorithm for computing such points. The techniques we develop for tackling L-smooth and weakly convex functions are highly versatile and can be applied to tackle, e.g., bilevel programs. In particular, they allow us to remove the assumption of third-order differentiability for the lower-level function in, e.g., (Yang et al., 2023; Chen et al., 2025b; Huang et al., 2025) in the nonconvex–strongly-convex (NC–SC) regime while still retaining a meaningful second-order guarantee, and to broaden the applicability of the set-smoothness machinery developed in (Chen et al., 2025a) in the nonconvex–(Polyak–Łojasiewicz) (NC–PŁ) regime.

The rest of the paper is organized as follows. We first lay out the notation and preliminaries in Section 2. Then, we present the second-order randomized smoothing in Section 3, the randomized upper bound in Section 4, and the extensions to weakly convex functions and applications to bilevel programs in Section 5. Finally, we conclude this paper in Section 6.

2 Notation and preliminaries

The notation in this paper adheres primarily to standard conventions, and we refer readers to Appendix A for a comprehensive overview. To streamline the presentation, we defer most proofs to the appendices.

2.1 Generalized differentiation theory and stationarity concepts

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

Definition 2.1 *Given a locally 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).
- *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).

We note that for a locally Lipschitz function f and a point $\boldsymbol{x}$, the sets $\partial_{\mathrm{C}} f(\boldsymbol{x})$ 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.[1] We begin with the Clarke generalized Hessian, defined analogously to $\partial_{\mathrm{C}} f(\boldsymbol{x})$.

[1]For a comprehensive treatment to this subject, we refer to (Mordukhovich, 2024).

Definition 2.2 *Given an L-smooth 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*

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

  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 an L-smooth function f and a point $\boldsymbol{x}$, the set $\partial_{\mathrm{C}}^2 f(\boldsymbol{x})$ is also nonempty, convex, and compact; see, e.g., (Clarke, 1990, Proposition 2.6.2(a)). Very naturally, the construction of $\partial_{\mathrm{C}}^2 f(\boldsymbol{x})$ gives rise to a second-order optimality condition that parallels the smooth case; i.e., vanishing gradient and positive semidefinite Hessian.

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

$$\nabla f(\boldsymbol{x}) = \boldsymbol{0} \quad \text{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial_{\mathrm{C}}^2 f(\boldsymbol{x})} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w} \geq 0. \tag{1}$$

For convenience, we refer to points satisfying (1) as Clarke SOSPs. We now turn to the central object in this subsection: The Goldstein generalized Hessian.

Definition 2.4 *Given a constant $\delta \geq 0$, an L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \mathbb{R}^n$, the Goldstein generalized Hessian of f at $\boldsymbol{x}$ is defined as*

$$\partial_\delta^2 f(\boldsymbol{x}) := \operatorname{conv}\Big(\bigcup\nolimits_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n} \partial_{\mathrm{C}}^2 f(\boldsymbol{y})\Big).$$

In other words, $\partial_\delta^2 f(\boldsymbol{x})$ is the smallest convex set that contains all possible $\partial_{\mathrm{C}}^2 f(\boldsymbol{y})$ 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". 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). Next, we introduce the stationarity concept that $\partial_\delta^2 f(\boldsymbol{x})$ induces.

Definition 2.5 *Given constants $\varepsilon_1, \varepsilon_2, \delta \geq 0$, an L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \mathbb{R}^n$, we say that $\boldsymbol{x}$ is an $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approximate SOSP of f if*

$$\|\nabla f(\boldsymbol{x})\| \leq \varepsilon_1 \quad \text{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial_\delta^2 f(\boldsymbol{x})} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{w} \geq -\varepsilon_2.$$

Lastly, we introduce second subderivatives, which, while not serving as our algorithmic target, will play a crucial role in our analysis.

Definition 2.6 *Given a function $f : \mathbb{R}^n \to \mathbb{R}$, a point $\boldsymbol{x} \in \mathbb{R}^n$, and vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$:*

- *The second subderivative of f at $\boldsymbol{x}$ for $\boldsymbol{v}$ and $\boldsymbol{w}$ is defined as*

$$\mathrm{d}^2 f(\boldsymbol{x}; \boldsymbol{v})(\boldsymbol{w}) := \liminf_{t\searrow 0,\, \boldsymbol{w}'\to\boldsymbol{w}} \frac{f(\boldsymbol{x} + t\boldsymbol{w}') - f(\boldsymbol{x}) - t\cdot \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}'}{\frac{1}{2}t^2};$$

  see, e.g., (Rockafellar & Wets, 2009, Definition 13.3).

As in Lemma 2.3, second subderivatives likewise give rise to a second-order optimality condition, whose solutions we refer to as Rockafellar SOSPs.

Lemma 2.7 ((Rockafellar & Wets, 2009, Theorem 13.24(a))) *If $\boldsymbol{x} \in \mathbb{R}^n$ is a local minimizer of $f : \mathbb{R}^n \to \mathbb{R}$, then*[2]

$$\boldsymbol{0} \in \partial f(\boldsymbol{x}) \quad \text{and} \quad \mathrm{d}^2 f(\boldsymbol{x}; \boldsymbol{0})(\boldsymbol{w}) \geq 0 \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^n.$$

[2]Here, the subdifferential is understood in the sense of Mordukhovich (2006, Definition 1.77).

3 Second-order randomized smoothing

3.1 From differentiability to twice differentiability

The theorem below is the main result of this section and underpins everything that follows.

Theorem 3.1 *Let $L, \sigma > 0$ be arbitrary and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. For $f_\sigma : \mathbb{R}^n \to \mathbb{R}$ given by*

$$f_\sigma(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{B}^n)}[f(\boldsymbol{x}+\sigma\boldsymbol{u})] \quad \text{for all } \boldsymbol{x}\in\mathbb{R}^n,$$

the following statements hold.

- *The function $f_\sigma : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.*
- *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the mapping $\nabla f_\sigma : \mathbb{R}^n \to \mathbb{R}^n$ satisfies*

$$\nabla f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})] \quad \text{and} \quad \|\nabla f_\sigma(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{y})\| \le L\|\boldsymbol{x}-\boldsymbol{y}\|.$$

- *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the mapping $\nabla^2 f_\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}_{\mathrm{sym}}$ satisfies*

$$\nabla^2 f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{B}^n)}[\nabla^2 f(\boldsymbol{x}+\sigma\boldsymbol{u})] \quad \text{and} \quad \|\nabla^2 f_\sigma(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{y})\| \le \frac{cL\sqrt{n}}{\sigma}\|\boldsymbol{x}-\boldsymbol{y}\|,$$

where $c > 0$ is a fixed constant.

3.2 Gradient and Hessian estimators, and connecting $\nabla^2 f_\sigma(\boldsymbol{x})$ with $\partial^2_\delta f(\boldsymbol{x})$

With Theorem 3.1 in place, we next develop estimators for the gradient and Hessian of f_σ.

Proposition 3.2 *Let $L, \sigma > 0$, $\boldsymbol{x} \in \mathbb{R}^n$, and $b = 1, 2, \ldots$ be arbitrary, $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and*

$$\widehat{\boldsymbol{g}}_b(\boldsymbol{x}) := \frac{1}{b}\sum_{j=1}^{b} \frac{\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}_j) + \nabla f(\boldsymbol{x}-\sigma\boldsymbol{u}_j)}{2}, \quad \text{where } \boldsymbol{u}_1, \ldots, \boldsymbol{u}_b \sim \mathrm{Unif}(\mathbb{B}^n) \text{ independently.}$$

Then, $\widehat{\boldsymbol{g}}_b(\boldsymbol{x})$ is unbiased and satisfies

$$\mathbb{P}\{\|\widehat{\boldsymbol{g}}_b(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\| \ge t\} \le 2\exp\Big(-\frac{bt^2}{2L^2\sigma^2 + 4L\sigma t/3}\Big) \quad \text{for all } t > 0.$$

Proposition 3.3 *Let $L, \sigma > 0$, $\boldsymbol{x} \in \mathbb{R}^n$, and $b = 1, 2, \ldots$ be arbitrary, $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and*

$$\widehat{\boldsymbol{H}}_b(\boldsymbol{x}) := \frac{1}{b}\sum_{j=1}^{b} \mathrm{Sym}\Big(\frac{n}{2\sigma}\big(\nabla f(\boldsymbol{x}+\sigma\boldsymbol{s}_j) - \nabla f(\boldsymbol{x}-\sigma\boldsymbol{s}_j)\big)\boldsymbol{s}_j^{\mathrm{T}}\Big),$$

where $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_b \sim \mathrm{Unif}(\mathbb{S}^n)$ independently and $\mathrm{Sym}(\boldsymbol{A}) := (\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})/2$. Then, $\widehat{\boldsymbol{H}}_b(\boldsymbol{x})$ is unbiased and satisfies

$$\mathbb{P}\Big\{\big\|\widehat{\boldsymbol{H}}_b(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big\| \ge t\Big\} \le 2n\exp\Big(-\frac{bt^2/2}{(n+1)^2L^2 + (n+1)Lt/3}\Big) \quad \text{for all } t > 0.$$

We close this section with a second-order stationarity transfer property from f_σ to f, which parallels (Lin et al., 2022, Theorem 3.1) and will prove useful in later developments.

Proposition 3.4 *Let $L, \varepsilon_1, \varepsilon_2, \delta > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. If $\|\nabla f_\sigma(\boldsymbol{x})\| \le \varepsilon_g$ and $\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x})\big) \ge -\varepsilon_H$, where $\varepsilon_g := \varepsilon_1/3$, $\varepsilon_H := \varepsilon_2$, and $\sigma := \min\big\{\delta, \varepsilon_1/(3L)\big\}$, then*

$$\|\nabla f(\boldsymbol{x})\| \le \varepsilon_1 \quad \text{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial^2_\delta f(\boldsymbol{x})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w}\rangle \ge -\varepsilon_2.$$

4 Computing Goldstein approximate SOSPs

Building on the developments in Section 3, we are now ready to apply the cubic Newton method to f_σ to compute Goldstein approximate SOSPs of L-smooth functions. The resulting procedure is described in Algorithm 1. We remark that the cubic Newton subproblem in Step 6 is polynomial-time solvable; see, e.g., (Griewank, 1981) and (Nesterov & Polyak, 2006, Section 5.1).

Algorithm 1 Computing $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approximate SOSPs of L-smooth functions

Input: An L-smooth $f : \mathbb{R}^n \to \mathbb{R}$, an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, and tolerances $\varepsilon_1, \varepsilon_2, \delta, \eta > 0$.

1: Let
$$\varepsilon_g := \frac{\varepsilon_1}{3}, \quad \varepsilon_H := \varepsilon_2, \quad \text{and} \quad \sigma := \min\Big\{\delta, \frac{\varepsilon_1}{3L}\Big\};$$

2: Let $\xi := cL\sqrt{n}/\sigma$ and
$$K_\star := \left\lceil \frac{24\Delta}{\xi r_\star^3} \right\rceil,$$
where
$$r_\star := \min\left\{\sqrt{\frac{16\varepsilon_g}{17\xi}}, \frac{24\varepsilon_H}{37\xi}\right\} > 0 \quad \text{and} \quad \Delta := f_\sigma(\boldsymbol{x}_0) - \min_{\boldsymbol{x}\in\mathbb{R}^n} f_\sigma(\boldsymbol{x});$$

3: **for** $k = 0, \ldots, K_\star - 1$ **do**

4: Sample $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{b_g} \sim \mathrm{Unif}(\mathbb{B}^n)$ independently, and construct
$$\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) := \frac{1}{b_g}\sum_{j=1}^{b_g} \frac{\nabla f(\boldsymbol{x}_k + \sigma\boldsymbol{u}_j) + \nabla f(\boldsymbol{x}_k - \sigma\boldsymbol{u}_j)}{2},$$
where
$$b_g := \left\lceil \left(\frac{4608L^2\sigma^2}{\xi^2 r_\star^4} + \frac{64L\sigma}{\xi r_\star^2}\right)\log\left(\frac{4K_\star}{\eta}\right)\right\rceil;$$

5: Sample $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{b_H} \sim \mathrm{Unif}(\mathbb{S}^n)$ independently, and construct
$$\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) := \frac{1}{b_H}\sum_{j=1}^{b_H} \mathrm{Sym}\Big(\frac{n}{2\sigma}\big(\nabla f(\boldsymbol{x}_k + \sigma\boldsymbol{s}_j) - \nabla f(\boldsymbol{x}_k - \sigma\boldsymbol{s}_j)\big)\boldsymbol{s}_j^{\mathrm{T}}\Big),$$
where
$$b_H := \left\lceil \left(\frac{1152(n+1)^2L^2}{\xi^2 r_\star^2} + \frac{16(n+1)L}{\xi r_\star}\right)\log\left(\frac{4nK_\star}{\eta}\right)\right\rceil;$$

6: Compute $\boldsymbol{p}_k \in \mathbb{R}^n$ as an arbitrary global minimizer of
$$\min\Big\{\widehat{m}_k(\boldsymbol{p}) := \boldsymbol{p}^{\mathrm{T}}\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p} + \frac{\xi}{6}\|\boldsymbol{p}\|^3 : \boldsymbol{p}\in\mathbb{R}^n\Big\}; \tag{2}$$

7: Update $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{p}_k$;

The main result of this section is as follows.

Theorem 4.1 *Consider the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 1. With probability at least $1-\eta$, there exists some $k_\star = 0, \ldots, K_\star - 1$ such that*
$$\|\nabla f(\boldsymbol{x}_{k_\star+1})\| \le \varepsilon_1 \quad \textit{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial_\delta^2 f(\boldsymbol{x}_{k_\star+1})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w}\rangle \ge -\varepsilon_2.$$

In other words, $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein SOSP is computable in a randomized manner within
$$K_\star \lesssim \max\left\{\frac{\Delta L^{1/2} n^{1/4}}{\sigma^{1/2}\varepsilon_g^{3/2}}, \frac{\Delta L^2 n}{\sigma^2\varepsilon_H^3}\right\} \asymp \max\left\{\frac{\Delta L^{1/2} n^{1/4}}{\min\{\delta, \varepsilon_1/L\}^{1/2}\varepsilon_1^{3/2}}, \frac{\Delta L^2 n}{\min\{\delta, \varepsilon_1/L\}^2\varepsilon_2^3}\right\}$$
outer loops of Algorithm 1, which is independent of η. In particular, when $\varepsilon_1, \varepsilon_2, \delta \asymp \varepsilon$ with ε small enough and $L \gg 1$, we have $K_\star \lesssim \Delta L^4 n/\varepsilon^5$, and the total number of first-order oracle calls is, modulo logarithmic factors, of order $\Delta L^8 n^2/\varepsilon^9 + \Delta L^6 n^3/\varepsilon^7$.

At a high level, the proof of Theorem 4.1 rests on a descent-or-stationarity dichotomy of the cubic Newton steps when applied to f_σ, whose Hessian is Lipschitz by Theorem 3.1. We first condition on the event that the sampled gradient and Hessian estimators are uniformly accurate over all iterations. On this event, the cubic Newton analysis of (2) makes this dichotomy precise: A large step produces a definite decrease in f_σ, while a small step certifies approximate stationarity at the next iterate. When $\Delta < \infty$, not every step can be large, as the resulting cumulative decrease would exceed this gap. Hence, some iterate must be an $(\varepsilon_g, \varepsilon_H)$-approximate SOSP of f_σ. Concentration inequalities and a union bound show that the estimators are accurate with high probability. Finally, Proposition 3.4 converts this point into an $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approximate SOSP of f.

5 Extensions and applications

5.1 Weakly convex functions

The goal of this subsection is to extend the framework developed earlier to weakly convex functions, which arise in various applications; e.g., sparse dictionary learning (Davis & Drusvyatskiy, 2019, Example 2.4) and robust low-rank matrix recovery (Li et al., 2020b). As weakly convex functions are a.e. twice differentiable by Alexandrov's theorem (see, e.g., (Rockafellar & Wets, 2009, Theorem 13.51)), the Clarke generalized Hessian remains well defined. However, as it turns out, the construction no longer gives rise to optimality conditions for this extended class of functions.

Example 5.1 *Consider $f(x) := |x| - x^2$, which is 2-weakly convex. It is easy to see that 0 is a (strict) local minimizer of f. However, $\partial_{\mathrm{C}}^2 f(0) = \mathrm{conv}\big(\limsup_{y\to 0,\, y\neq 0}\{f''(y)\}\big) = \mathrm{conv}\big(\limsup_{y\to 0,\, y\neq 0}\{-2\}\big) = \{-2\}$, which is negative definite.*

In fact, without further assumptions, the Clarke generalized Hessian could even be empty.

Example 5.2 *Consider $f(x) := |x|^{4/3} - x^2$, which is 2-weakly convex. We have $\partial_{\mathrm{C}}^2 f(0) = \mathrm{conv}\big(\limsup_{y\to 0,\, y\neq 0}\{f''(y)\}\big) = \mathrm{conv}\big(\limsup_{y\to 0,\, y\neq 0}\big\{\frac{4}{9}|y|^{-2/3} - 2\big\}\big) = \emptyset$.*

That said, it is well known that a weakly convex function is closely connected to its Moreau envelope, which is L-smooth; see, e.g., (Davis & Drusvyatskiy, 2022, Lemma 2.5). In particular, their FOSPs and local/global minimizers all coincide; see, e.g., (Davis & Drusvyatskiy, 2022, Lemma 2.5(4)), (Khanh et al., 2025a, Theorem 3.5), and (Hoheisel et al., 2020, Corollary 3.4(d)), respectively. In fact, a weakly convex function and its Moreau envelope share the same Rockafellar SOSPs, which, as an aside, suggests new approaches to analyzing the landscape of nonsmooth functions.

Theorem 5.3 *Let $\rho > 0$ and $\lambda \in (0, 1/\rho)$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be ρ-weakly convex. For every $\boldsymbol{x} \in \mathbb{R}^n$, the following two statements are equivalent.*

- *We have $\mathbf{0} \in \partial f(\boldsymbol{x})$ and $\mathrm{d}^2 f(\boldsymbol{x}; \mathbf{0})(\boldsymbol{w}) \geq 0$ for all $\boldsymbol{w} \in \mathbb{R}^n$.*
- *We have $\nabla f_\lambda(\boldsymbol{x}) = \mathbf{0}$ and $\mathrm{d}^2 f_\lambda(\boldsymbol{x}; \mathbf{0})(\boldsymbol{w}) \geq 0$ for all $\boldsymbol{w} \in \mathbb{R}^n$.*

Drawing on this, we propose to define the notion of Clarke SOSPs for weakly convex functions in terms of their Moreau envelopes, which is justified by the following proposition and also addresses in part the second open question raised in (Khanh et al., 2025a, Section 5).

Proposition 5.4 *Let $\rho > 0$ and $\lambda \in (0, 1/\rho)$ be arbitrary, $f : \mathbb{R}^n \to \mathbb{R}$ be ρ-weakly convex, and $\boldsymbol{x} \in \mathbb{R}^n$ be a local minimizer of f. We have*

$$\nabla f_\lambda(\boldsymbol{x}) = \mathbf{0} \quad \text{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{z}\in\partial_{\mathrm{C}}^2 f_\lambda(\boldsymbol{x})(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{w}\rangle \geq 0.$$

We now proceed to define Goldstein approximate SOSPs for weakly convex functions.

Definition 5.5 *Given constants $\varepsilon_1, \varepsilon_2, \delta \geq 0$, a ρ-weakly convex function $f : \mathbb{R}^n \to \mathbb{R}$, and a point $\boldsymbol{x} \in \mathbb{R}^n$, we say that $\boldsymbol{x}$ is an $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approximate SOSP of f if*

$$\|\nabla f_{1/(2\rho)}(\boldsymbol{x})\| \leq \varepsilon_1 \quad \textit{and} \quad \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{A} \in \partial^2_\delta f_{1/(2\rho)}(\boldsymbol{x})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w} \rangle \geq -\varepsilon_2.$$

We now apply Algorithm 1 to compute such points of weakly convex functions.

Algorithm 2 Computing $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approximate SOSPs of weakly convex functions

Input: A ρ-weakly convex $g : \mathbb{R}^n \to \mathbb{R}$, an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, tolerances $\varepsilon_1, \varepsilon_2, \delta, \eta > 0$.

1: Run Algorithm 1 with $f := g_{1/(2\rho)}$ and $L := 2\rho$, while keeping all the other inputs the same as the ones here;

We remark that the only additional ingredient needed in the implementation is the evaluation of $\nabla g_{1/(2\rho)}$. As $\nabla g_{1/(2\rho)}(\boldsymbol{x}) = 2\rho\big(\boldsymbol{x} - \mathrm{Prox}_{g/(2\rho)}(\boldsymbol{x})\big)$ by (Davis & Drusvyatskiy, 2022, Lemma 2.5(1)), it suffices to compute $\mathrm{Prox}_{g/(2\rho)}(\boldsymbol{x}) = \arg\min\big\{g(\boldsymbol{y}) + \rho\|\boldsymbol{y} - \boldsymbol{x}\|^2 : \boldsymbol{y} \in \mathbb{R}^n\big\}$, which is strongly convex and can thus be solved efficiently (Bubeck, 2015, Section 3.4).

Corollary 5.6 *Consider the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 2. With probability at least $1 - \eta$, there exists some $k_\star = 0, \ldots, K_\star - 1$ such that*

$$\|\nabla g_{1/(2\rho)}(\boldsymbol{x}_{k_\star+1})\| \leq \varepsilon_1 \quad \textit{and} \quad \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{A} \in \partial^2_\delta g_{1/(2\rho)}(\boldsymbol{x}_{k_\star+1})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w} \rangle \geq -\varepsilon_2.$$

Proof. It suffices to combine Theorem 4.1 with the fact that $g_{1/(2\rho)}$ is 2ρ-smooth by (Davis & Drusvyatskiy, 2022, Lemma 2.5(3)). □

In particular, we have $K_\star \lesssim \Delta\rho^4 n/\varepsilon^5$ when $\varepsilon_1, \varepsilon_2, \delta \asymp \varepsilon$, with ε small enough and $\rho \gg 1$.

5.2 Second-order bilevel optimization

In this subsection, we focus on the following optimistic/pessimistic bilevel problems

$$\min_{\boldsymbol{x} \in \mathbb{X}} \min_{\boldsymbol{y} \in \mathbb{Y}^\star(\boldsymbol{x})} u(\boldsymbol{x}, \boldsymbol{y}) \quad \text{and} \quad \min_{\boldsymbol{x} \in \mathbb{X}} \max_{\boldsymbol{y} \in \mathbb{Y}^\star(\boldsymbol{x})} u(\boldsymbol{x}, \boldsymbol{y}), \quad \text{where } \mathbb{Y}^\star(\boldsymbol{x}) := \mathrm{Arg}\min\{\ell(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} \in \mathbb{Y}\},$$

$u, \ell : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$, $\mathbb{X} \subseteq \mathbb{R}^n$, and $\mathbb{Y} \subseteq \mathbb{R}^m$. Equivalently, one may also pose them as

$$\min\Big\{h_o(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathbb{Y}^\star(\boldsymbol{x})} u(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in \mathbb{X}\Big\} \quad \text{and} \quad \min\Big\{h_p(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathbb{Y}^\star(\boldsymbol{x})} u(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in \mathbb{X}\Big\},$$

respectively, where $h_o, h_p : \mathbb{X} \to \mathbb{R}$ are referred to as the hyperobjective functions. In cases where $h_o = h_p$, we simply write h for both; e.g., when $|\mathbb{Y}^\star(\boldsymbol{x})| = 1$ for all $\boldsymbol{x} \in \mathbb{X}$, and in such a case, we also write $\boldsymbol{y}^\star$ for $\mathbb{Y}^\star$. Bilevel programs arise in many contemporary applications; e.g., hyperparameter optimization (Chen et al., 2024). In the sequel, we use NC–SC (resp., NC–PŁ) to refer to the setting where h is nonconvex (in general) and $\ell(\boldsymbol{x}, \bullet)$ is strongly convex (resp., satisfies the PŁ inequality) for all $\boldsymbol{x} \in \mathbb{X}$. When discussing the NC–SC and NC–PŁ settings, we take $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \mathbb{R}^m$, whereas in other settings $\mathbb{X}$ and $\mathbb{Y}$ may be more general. As an aside, as bilevel optimization is not the main focus of this paper, we leave issues such as inexactness to interested readers and future work.

5.2.1 The NC–SC setting without third-order differentiability

Existing studies on second-order NC–SC bilevel optimization are typically based on the third-order differentiability of ℓ to ensure that h is twice differentiable, thereby enabling the use of the machinery of second-order smooth optimization; see, e.g., (Yang et al., 2023, Assumption 3), (Chen et al., 2025b, Assumption 3.2), and (Huang et al., 2025, Assumption 3). However, the third-order differentiability condition can be rather demanding, and it is thus very natural to ask whether we can still say something meaningful in its absence. With the techniques developed in this paper, the answer is in the affirmative, as h remains L-smooth even without the condition, which follows from, e.g., (Ghadimi & Wang, 2018, Lemma 2.2) and (Chen et al., 2025b, Proposition 3.1), and is quoted below.

Lemma 5.7 ((Chen et al., 2025b, Proposition 3.1)) *Suppose that $\ell(\boldsymbol{x}, \bullet)$ is μ-strongly convex, ℓ has L_ℓ-Lipschitz gradient and ρ_ℓ-Lipschitz Hessian, $u(\boldsymbol{x}, \bullet)$ is C_u-Lipschitz, u is twice continuously differentiable and L_u-smooth, and h is lower bounded. Then, h is L_h-smooth, where $L_h \lesssim \max\{C_u, L_u, L_\ell, \rho_\ell\}^4/\mu^3$.*

Of course, away from the condition, the twice differentiability of h can no longer be guaranteed; see Example H.1. We now apply Algorithm 1 to NC–SC bilevel programs.

Algorithm 3 Computing $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approx. SOSPs of NC–SC bilevel programs

Input: Functions $u, \ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ satisfying all the assumptions in Lemma 5.7, an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, tolerances $\varepsilon_1, \varepsilon_2, \delta, \eta > 0$.

1: Run Algorithm 1 with $f := h = u\big(\bullet, \boldsymbol{y}^\star(\bullet)\big)$ and $L := L_h$, while keeping all the other inputs the same as the ones here;

We remark that the only additional ingredient needed in the implementation is the evaluation of ∇h; i.e., the hypergradient. As ∇h admits an explicit closed-form expression (see, e.g., (Chen et al., 2025b, (2))), its evaluation amounts to nothing more than matrix multiplications, which can be done efficiently.

Corollary 5.8 *Consider the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 3. With probability at least $1 - \eta$, there exists some $k_\star = 0, \dots, K_\star - 1$ such that*

$$\|\nabla h(\boldsymbol{x}_{k_\star+1})\| \le \varepsilon_1 \quad \text{and} \quad \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{A} \in \partial_\delta^2 h(\boldsymbol{x}_{k_\star+1})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w} \rangle \ge -\varepsilon_2.$$

Proof. It suffices to combine Lemma 5.7 with Theorem 4.1. □

In particular, when $\varepsilon_1, \varepsilon_2, \delta \asymp \varepsilon$, with ε small enough and $L_h \gg 1$, we have

$$K_\star \lesssim \frac{\Delta \max\{C_u, L_u, L_\ell, \rho_\ell\}^{16} n}{\mu^{12} \varepsilon^5}.$$

5.2.2 The NC–PŁ setting

While the NC–SC setting gives rise to an L-smooth hyperobjective function, the strong convexity assumption can be demanding. In fact, with the help of the PŁ inequality (Polyak, 1963), one can move beyond strong convexity, albeit with a slight reduction in smoothness. In what follows, we restrict our attention to the pessimistic case, as the other case is a mirror image.

Lemma 5.9 ((Chen et al., 2025a, Theorem 3)) *Suppose that ℓ is L_ℓ-smooth and twice differentiable with $\nabla\nabla_{\boldsymbol{y}}\ell$ being H_ℓ-Lipschitz, $\mathbb{Y}^\star(\boldsymbol{x})$ is nonempty, closed, and convex for all $\boldsymbol{x} \in \mathbb{R}^n$, u is M_u-Lipschitz and L_u-smooth, $\mathrm{dom}(h_p) \neq \emptyset$, and there exists a constant $\tau > 0$ such that*

$$\mathrm{dist}\big(\boldsymbol{y}, \mathbb{Y}^\star(\boldsymbol{x})\big) \le \tau \|\nabla_{\boldsymbol{y}} \ell(\boldsymbol{x}, \boldsymbol{y})\| \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \text{ and } \boldsymbol{y} \in \mathbb{R}^m.$$

Then, h_p is $\big(M_u L_{\mathbb{Y}^\star} + L_u(1 + L_{\mathbb{Y}^\star})\big)$-weakly convex for $L_{\mathbb{Y}^\star} := \max\big\{2H_\ell\tau(1 + 9L_\ell^2\tau^2), 4L_\ell^2\tau^2\big\}$.

We now apply Algorithm 2 to NC–PŁ bilevel programs.

Algorithm 4 Computing $(\varepsilon_1, \varepsilon_2, \delta)$-Goldstein approx. SOSPs of NC–PŁ bilevel programs

Input: Functions $u, \ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ satisfying all the assumptions in Lemma 5.9, an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, tolerances $\varepsilon_1, \varepsilon_2, \delta, \eta > 0$.

1: Run Algorithm 2 with $g := h_p = \max_{\boldsymbol{y} \in \mathbb{Y}^\star(\bullet)} u(\bullet, \boldsymbol{y})$ and $\rho := M_u L_{\mathbb{Y}^\star} + L_u(1 + L_{\mathbb{Y}^\star})$, while keeping all the other inputs the same as the ones here;

Corollary 5.10 *Consider the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 4. With probability at least $1-\eta$, there exists some $k_\star = 0, \dots, K_\star - 1$ such that*

$$\|\nabla (h_p)_{1/(2\rho)}(\boldsymbol{x}_{k_\star+1})\| \leq \varepsilon_1 \quad \textit{and} \quad \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial_\delta^2 (h_p)_{1/(2\rho)}(\boldsymbol{x}_{k_\star+1})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w}\rangle \geq -\varepsilon_2.$$

Proof. It suffices to combine Lemma 5.9 with Corollary 5.6. □

In particular, when $\varepsilon_1, \varepsilon_2, \delta \asymp \varepsilon$, with ε small enough and $\rho \gg 1$, we have

$$K_\star \lesssim \frac{\Delta\big(M_u L_{\mathbb{Y}^\star} + L_u(1+L_{\mathbb{Y}^\star})\big)^4 n}{\varepsilon^5}.$$

Similar to Section 5.1, the only additional ingredient needed in the implementation is the evaluation of $\nabla (h_p)_{1/(2\rho)}$, or equivalently $\mathrm{Prox}_{h_p/(2\rho)}(\boldsymbol{x}) = \arg\min\big\{h_p(\boldsymbol{y}) + \rho\|\boldsymbol{y}-\boldsymbol{x}\|^2 : \boldsymbol{y} \in \mathbb{R}^n\big\}$. While this is a strongly convex program, typically, we do not have access to ∂h_p, as h_p arises as the marginal function of a nonconvex program. This hinders the application of the subgradient method. Fortunately, the evaluation subroutine for h_p in (Chen et al., 2025a, Algorithm 2) allows us to apply the zeroth-order algorithm in (Shamir, 2017, Algorithm 1) (see also (Shamir, 2017, Corollary 2)) to compute $\mathrm{Prox}_{h_p/(2\rho)}(\boldsymbol{x})$. We leave the details to interested readers. As an aside, under further assumptions (e.g., availability of a local upper Lipschitz selection of $\mathrm{Arg\,max}\big\{u(\boldsymbol{x},\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{Y}^\star(\boldsymbol{x})\big\}$), one may apply the subdifferential calculi for marginal functions developed in, e.g., (Mordukhovich et al., 2009; Phuoc Hai et al., 2025), to compute ∂h_p, thereby enabling the use of first-order methods.

6 Concluding Remarks

In this paper, we have exhibited a randomized algorithm (i.e., Algorithm 1) that computes Goldstein approximate SOSPs of L-smooth functions in a polynomial number of first-order oracle calls. In particular, in the regime where $\varepsilon_1, \varepsilon_2, \delta \asymp \varepsilon$, the complexity admits the explicit form $\widetilde{O}\big(\Delta L^8 n^2/\varepsilon^9 + \Delta L^6 n^3/\varepsilon^7\big)$. In the course of developing Algorithm 1, we have also established some underpinning theory; e.g., a framework for second-order randomized smoothing, which may be of independent interest. Furthermore, with Algorithm 1 as a foundation, we have extended the computability of Goldstein SOSPs to weakly convex functions, NC–SC and NC–PŁ bilevel programs, and beyond, together with some auxiliary results that may be of independent interest, such as a second-order stationarity coincidence between a weakly convex function and its Moreau envelope.

This work also points to some directions for future work. First, it is natural to ask whether randomization enables Goldstein approximate SOSPs of L-smooth functions to be computed with dimension-free complexity, as is possible for Goldstein approximate FOSPs (Zhang et al., 2020; Tian et al., 2022; Davis et al., 2022). We conjecture that the answer is negative.

Conjecture 6.1 *No dimension-free randomized algorithm computes Goldstein approximate SOSPs of L-smooth functions.*

Second, it would be valuable to identify structured subclasses of L-smooth functions for which near-approximate SOSPs, i.e.,

$$\|\nabla f(\boldsymbol{x})\| \leq \varepsilon_1 \quad \text{and} \quad \max_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n} \min_{\boldsymbol{w}\in\mathbb{S}^n} \max_{\boldsymbol{A}\in\partial_{\mathrm{C}}^2 f(\boldsymbol{y})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w}\rangle \geq -\varepsilon_2,$$

can be computed efficiently. Finally, it would be interesting to investigate the landscapes of nonsmooth nonconvex functions, particularly those arising in applications, through the lens of Goldstein SOSPs.

AI USE STATEMENT

In this work, we 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", "assist with translation". We have not used generative AI tools for "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 mathematical and linguistic abilities. In particular, we have verified the correctness of the generated content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Alexander Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2002.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3880–3888, 2016.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research*, 49(3):1761–1790, 2024.

Vladimir I Bogachev. *Measure Theory, Volume I*. Springer-Verlag, Berlin, Heidelberg, 2007.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3–4):231–357, 2015.

Vivien A Cabannes, Francis Bach, and Alessandro Rudi. Fast rates for structured prediction. In *Proceedings of the 34th Annual Conference on Learning Theory*, pp. 823–865, 2021.

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.

He Chen, Haochen Xu, Rujun Jiang, and Anthony Man-Cho So. Lower-level duality based reformulation and majorization minimization algorithm for hyperparameter optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 784–792, 2024.

He Chen, Jiajin Li, and Anthony Man-Cho So. Set smoothness unlocks Clarke hyper-stationarity in bilevel optimization. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, pp. 134964–134995, 2025a.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109): 1–56, 2025b.

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

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(8):1713–1724, 2018.

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

Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22(2):561–606, 2022.

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.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. *Foundations of Computational Mathematics*, 26(2):779–861, 2026.

Rick Durrett. *Probability: Theory and Examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fifth edition, 2019.

Lawrence Craig Evans and Ronald F Gariepy. *Measure Theory and Fine Properties of Functions*. Textbooks in Mathematics. CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742, revised edition, 2015.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs, and Tracts. John Wiley & Sons, Inc., New York, second edition, 1999.

Gerald B Folland. *Advanced Calculus*. Second edition, 2024. https://sites.math.washington.edu//~folland/AdvCalc24.pdf.

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.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Allen 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 the hardness of deterministic second-order optimization of functions with Lipschitz gradients. *arXiv preprint arXiv:2607.24120*, 2026.

Nguyen TV Hang, Boris S Mordukhovich, and M Ebrahim Sarabi. Augmented Lagrangian method for second-order cone programs under second-order sufficiency. *Journal of Global Optimization*, 82(1):51–81, 2022.

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, Maxime Laborde, and Adam Oberman. A regularization interpretation of the proximal point method for weakly convex functions. *Journal of Dynamics and Games*, 7 (1):79–96, 2020.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2012.

Minhui Huang, Xuxing Chen, Kaiyi Ji, Shiqian Ma, and Lifeng Lai. Efficiently escaping saddle points in bilevel optimization. *Journal of Machine Learning Research*, 26(1):1–61, 2025.

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.

Pham Duy Khanh, Vu VH Khoa, Boris S Mordukhovich, and Vo Thanh Phat. Local minimizers of nonconvex functions in Banach spaces via Moreau envelopes. *Vietnam Journal of Mathematics*, 53(4):803–813, 2025a.

Pham Duy Khanh, Boris S Mordukhovich, Vo Thanh Phat, and Dat Ba Tran. Inexact proximal methods for weakly convex functions. *Journal of Global Optimization*, 91(3): 611–646, 2025b.

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.

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, Shuqin Sun, and Man-Chung Yue. Subdifferentially polynomially bounded functions and Gaussian smoothing-based zeroth-order optimization. *SIAM Journal on Optimization*, 35(2):1393–1418, 2025.

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 Rene 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.

Boris S Mordukhovich, Nguyen Mau Nam, and Nguyen Dong Yen. Subgradients of marginal functions in parametric mathematical programming. *Mathematical Programming*, 116(1): 369–396, 2009.

Boris Sholimovich 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 Sholimovich 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.

Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Yurii Nesterov and Boris 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.

Le Phuoc Hai, Felipe Lara, and Boris S Mordukhovich. Regular subgradients of marginal functions with applications to calculus and bilevel programming. *Journal of Optimization Theory and Applications*, 205(2):36, 2025.

Iosif F Pinelis and Aleksandr Ivanovich Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

B T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Spyridon Pougkakiotis and Dionysis Kalogerias. A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. *SIAM Journal on Scientific Computing*, 45(5):A2679–A2702, 2023.

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

Ralph 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.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Book Company, New York, third edition, 1976.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

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.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

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

Jingfan Xia, Zhenwei Lin, and Qi Deng. Revisiting randomized smoothing: Nonsmooth nonconvex optimization beyond global Lipschitz continuity. *arXiv preprint arXiv:2508.13496*, 2025.

Haikuo Yang, Luo Luo, Chris Junchi Li, Michael Jordan, and Maryam Fazel. Accelerating inexact hypergradient descent for bilevel optimization. In *OPT 2023: Optimization for Machine Learning (NeurIPS 2023 Workshop)*, 2023.

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.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 29820–29834, 2021.

Vladimir A Zorich. *Mathematical Analysis I*. Universitext. Springer-Verlag, Berlin, Heidelberg, second edition, 2015. Original Russian edition: Matematicheskij Analiz (Part I, 6th corrected edition, Moscow, 2012) MCCME (Moscow Center for Continuous Mathematical Education Publ.).

A Notation

In this paper, we use lowercase letters (e.g., x), boldface lowercase letters (e.g., $\boldsymbol{x} = (x_i)$), and boldface capital letters (e.g., $\boldsymbol{X} = (x_{ij})$), to denote scalars, vectors, and matrices, respectively. We use $\|\boldsymbol{x}\|$ to denote the ℓ_2-norm of any $\boldsymbol{x} \in \mathbb{R}^n$, and $\mathbb{B}^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$ and $\mathbb{S}^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$ to denote the unit ball and unit sphere in $\mathbb{R}^n$. For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we write $\|\boldsymbol{A}\|$ for its spectral norm, $\|\boldsymbol{A}\|_1$ for its (entrywise) ℓ_1-norm, and $\boldsymbol{A}^{\mathrm{T}}$ for its transpose. When $\boldsymbol{A}$ is symmetric, $\lambda_{\min}(\boldsymbol{A})$ denotes its smallest eigenvalue. The all-zero vector in $\mathbb{R}^n$ is denoted by $\boldsymbol{0}_n$, the i-th standard basis vector in $\mathbb{R}^n$ is denoted by $\boldsymbol{e}_i^n$, and the identity matrix in $\mathbb{R}^{n\times n}$ is denoted by $\boldsymbol{I}_n$; their subscripts and superscripts are often omitted as long as there is no ambiguity. We use $\vee$ to denote vector concatenation; i.e., $\boldsymbol{x} \vee \boldsymbol{y} = (\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}})^{\mathrm{T}}$. We also adopt $\langle \bullet, \bullet \rangle$ to denote the (Frobenius) inner product; i.e., $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \sum_{i=1}^m \sum_{j=1}^n x_{ij} y_{ij}$. As an aside, $\succeq$ denotes the Löwner order.

We use $\sum_{j>i}$ and $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n}$ interchangeably as long as n is clear from the context. Besides, $\lceil \bullet \rceil$ denotes the ceiling operation and log denotes the natural logarithm. We use Arg min to denote the set of (global) minimizers of a program, and, when the minimizer is unique, arg min to denote that unique element; the same convention applies to maximizers as well. As an aside, in this paper, we do not distinguish between the maximum and supremum, nor between the minimum and infimum. For any $\mathbb{X} \subseteq \mathbb{R}^n$, we denote by $\operatorname{conv}(\mathbb{X})$ its convex hull, by

$$\operatorname{dist}(\boldsymbol{y}, \mathbb{X}) := \min\big\{\|\boldsymbol{y} - \boldsymbol{x}\| : \boldsymbol{x} \in \mathbb{X}\big\}$$

the distance from $\boldsymbol{y}$ to $\mathbb{X}$, by $\operatorname{vol}(\mathbb{X})$ its volume (i.e., its Lebesgue measure),[3] by $\mathbb{X}^{\complement}$ its complement, and by $|\mathbb{X}|$ its cardinality. When $|\mathbb{X}| = 1$, we identify $\mathbb{X}$ with its unique element. We also denote by $\emptyset$ the empty set. For any $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$, we use $\operatorname{dom}(f)$ to denote its effective domain; see, e.g., (Rockafellar & Wets, 2009, Section 1.A). For any $\lambda > 0$, $\boldsymbol{x} \in \mathbb{R}^n$, and lower semicontinuous $f : \mathbb{R}^n \to \mathbb{R}$, its Moreau envelope and proximal mapping are defined by (see, e.g., (Rockafellar & Wets, 2009, Definition 1.22))

$$f_\lambda(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\mathbb{R}^n} \left(f(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \right) \quad \text{and} \quad \operatorname{Prox}_{\lambda f}(\boldsymbol{x}) := \operatorname*{Arg\,min}_{\boldsymbol{y}\in\mathbb{R}^n} \left(f(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \right).$$

For any $f : \mathbb{R}^n \to \mathbb{R}$, we say that it is L-smooth for an $L \geq 0$ if it is differentiable with ∇f being L-Lipschitz, whereas that it is ρ-weakly convex for a $\rho \geq 0$ if $f + 2^{-1}\rho\|\bullet\|^2$ is convex (Vial, 1983, Definition 4.1). In passing, we remark that as weakly convex functions are subdifferentially regular (by, e.g., (Rockafellar & Wets, 2009, Example 7.27) and (Rockafellar & Wets, 2009, Exercise 8.20(b))) and locally Lipschitz (by, e.g., (Vial, 1983, Proposition 4.4)), their Fréchet (Rockafellar & Wets, 2009, Definition 8.3(a)), Mordukhovich (Mordukhovich, 2006, Definition 1.77), and Clarke (Clarke, 1990, Theorem 2.5.1) subdifferentials all collapse to the same object (by, e.g., (Li et al., 2020a, Fact 5)), and we thus use ∂f to denote any of them. For a vector field $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, we use $\operatorname{div} \boldsymbol{F}$ to denote its divergence.

We adopt $\mathbb{P}$ to denote a probability measure and $\mathbb{E}$ the corresponding expectation, and, when necessary, the underlying law is specified by the subscript; e.g., $\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}$, where $\mathrm{Unif}(\mathbb{B}^n)$ is the uniform distribution on $\mathbb{B}^n$, and $\sim$ means "is distributed according to." We also use $\overset{\text{iid}}{\sim}$ to denote independent and identically distributed sampling, and $\#$ to denote the pushforward operator. While asymptotic notation is used extensively in this paper, we refrain from reviewing it in detail, as it adheres primarily to standard conventions in the field. All set-valued limit superiors are understood in the sense of Painlevé–Kuratowski (see, e.g., (Rockafellar & Wets, 2009, Section 5.B)), whereas scalar limit superiors are taken in the usual sense; the same convention applies to limit inferiors as well.

[3] By a slight abuse of notation, we also write $\operatorname{vol}(\mathbb{S}^n)$ for the surface area of $\mathbb{S}^n$.

B Proof of Theorem 3.1

Proof of Theorem 3.1. We shall first deal with the differentiability of f_σ. To this end, for some given $\boldsymbol{x} \in \mathbb{R}^n$, let us consider the functions $f_{\boldsymbol{x},i} : (-1,1) \times \mathbb{B}^n \to \mathbb{R}$ for $i = 1, \ldots, n$ defined by $f_{\boldsymbol{x},i}(t, \boldsymbol{u}) := f(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i)$. First, by (Rudin, 1976, Theorem 4.14), it follows that for any $t \in (-1,1)$, there exists some $M > 0$, such that $|f_{\boldsymbol{x},i}(t, \boldsymbol{u})| \leq M$ for all $\boldsymbol{u} \in \mathbb{B}^n$. As a result, for all such t, we have $\mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}[|f_{\boldsymbol{x},i}(t, \boldsymbol{u})|] \leq M < \infty$; i.e., $f_{\boldsymbol{x},i}$ is Lebesgue-integrable w.r.t. $\boldsymbol{u}$. Besides, as f is L-smooth as assumed, we know that $\partial f_{\boldsymbol{x},i}(t, \boldsymbol{u})/\partial t$ exists and equals $\partial f(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i)/\partial x_i$ for all $t \in (-1,1)$ and $\boldsymbol{u} \in \mathbb{B}^n$. Moreover, by (Rudin, 1976, Theorem 4.14) again, there further exists some $M' > 0$ such that $|\partial f_{\boldsymbol{x},i}(t, \boldsymbol{u})/\partial t| \leq M'$ for all $t \in (-1,1)$ and $\boldsymbol{u} \in \mathbb{B}^n$. It then follows from (Folland, 1999, Theorem 2.27(b)) that for all $i = 1, \ldots, n$, the function $t \mapsto f_\sigma(\boldsymbol{x} + t\boldsymbol{e}_i) = \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}[f_{\boldsymbol{x},i}(t, \boldsymbol{u})]$ is differentiable at 0 with

$$\frac{\partial f_\sigma(\boldsymbol{x})}{\partial x_i} = \big(t \mapsto f_\sigma(\boldsymbol{x} + t\boldsymbol{e}_i)\big)'(0) = \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial f_{\boldsymbol{x},i}}{\partial t}(0, \boldsymbol{u})\right] = \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial f}{\partial x_i}(\boldsymbol{x} + \sigma \boldsymbol{u})\right];$$

i.e., $\nabla f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x} + \sigma \boldsymbol{u})]$. It remains to show that $\partial f_\sigma(\boldsymbol{x})/\partial x_i$ for $i = 1, \ldots, n$ are all Lipschitz continuous.

Indeed, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\begin{aligned}\left\| \begin{pmatrix} \partial f_\sigma(\boldsymbol{x})/\partial x_1 \\ \vdots \\ \partial f_\sigma(\boldsymbol{x})/\partial x_n \end{pmatrix} - \begin{pmatrix} \partial f_\sigma(\boldsymbol{y})/\partial x_1 \\ \vdots \\ \partial f_\sigma(\boldsymbol{y})/\partial x_n \end{pmatrix} \right\| &= \big\| \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\big[\nabla f(\boldsymbol{x} + \sigma \boldsymbol{u})\big] - \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\big[\nabla f(\boldsymbol{y} + \sigma \boldsymbol{u})\big] \big\| \\ &\leq \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\big[\|\nabla f(\boldsymbol{x} + \sigma \boldsymbol{u}) - \nabla f(\boldsymbol{y} + \sigma \boldsymbol{u})\|\big] \leq L\|\boldsymbol{x} - \boldsymbol{y}\|;\end{aligned}$$

i.e., ∇f_σ is L-Lipschitz, as desired. Here, the first inequality follows from Jensen's inequality; see, e.g., (Durrett, 2019, Theorem 1.6.2).

We next turn to the twice differentiability of f_σ. Given $\boldsymbol{x} \in \mathbb{R}^n$, we see that

$$\begin{aligned}\frac{1}{t}\left(\frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x} + t\boldsymbol{e}_i) - \frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x})\right) &= \frac{1}{t}\left(\mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i)\right] - \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right]\right) \\ &= \int_{\mathbb{B}^n} \frac{1}{t}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right) \mathbb{P}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}(\mathrm{d}\boldsymbol{u})\end{aligned}$$

for all $j = 1, \ldots, n$. In what follows, we shall take limits on both sides and pass the limit on the right-hand side by applying the dominated convergence theorem (DCT); see, e.g., (Folland, 1999, Theorem 2.24). To begin, as f is L-smooth, we know from Rademacher's theorem (Rockafellar & Wets, 2009, Theorem 9.60) that ∇f is a.e. differentiable. Hence, for all $i = 1, \ldots, n$, the limit

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u}) = \lim_{t \to 0} \frac{1}{t}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right)$$

exists for $\mathbb{P}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}$-a.e. $\boldsymbol{u} \in \mathbb{B}^n$. This, together with Heine's theorem (Zorich, 2015, Proposition 1), further implies that for all sequences $t_1, t_2, \ldots$ with $\lim_{k \to \infty} t_k = 0$ and all such $\boldsymbol{u} \in \mathbb{B}^n$, the limit

$$\lim_{k \to \infty} \frac{1}{t_k}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t_k\boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right)$$

exists and equals $\partial^2 f(\boldsymbol{x} + \sigma \boldsymbol{u})/\partial x_i \partial x_j$. Besides, for all $t \neq 0$, we have

$$\left|\frac{1}{t}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right)\right| \leq \frac{1}{|t|}\|\nabla f(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i) - \nabla f(\boldsymbol{x} + \sigma \boldsymbol{u})\| \leq L,$$

and thus

$$\int_{\mathbb{B}^n} \left|\frac{1}{t}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u} + t\boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma \boldsymbol{u})\right)\right| \mathbb{P}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^n)}(\mathrm{d}\boldsymbol{u}) < \infty.$$

It then follows from the DCT that for all sequences $t_1, t_2, \ldots$ with $\lim_{k\to\infty} t_k = 0$,

$$\begin{aligned}
&\lim_{k\to\infty} \frac{1}{t_k}\left(\frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x} + t_k \boldsymbol{e}_i) - \frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x})\right) \\
&= \lim_{k\to\infty} \int_{\mathbb{B}^n} \frac{1}{t_k}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u} + t_k \boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u})\right) \mathbb{P}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}(\mathrm{d}\boldsymbol{u}) \\
&= \int_{\mathbb{B}^n} \lim_{k\to\infty} \frac{1}{t_k}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u} + t_k \boldsymbol{e}_i) - \frac{\partial f}{\partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u})\right) \mathbb{P}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}(\mathrm{d}\boldsymbol{u}) \\
&= \int_{\mathbb{B}^n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u}) \mathbb{P}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}(\mathrm{d}\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u})\right].
\end{aligned}$$

Hence, we have by Heine's theorem (Zorich, 2015, Proposition 1) again that for all $i, j = 1, \ldots, n$,

$$\frac{\partial^2 f_\sigma}{\partial x_i \partial x_j}(\boldsymbol{x}) = \lim_{t\to 0} \frac{1}{t}\left(\frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x} + t\boldsymbol{e}_i) - \frac{\partial f_\sigma}{\partial x_j}(\boldsymbol{x})\right) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + \sigma\boldsymbol{u})\right];$$

i.e., $\nabla^2 f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla^2 f(\boldsymbol{x} + \sigma\boldsymbol{u})]$. It remains to show that $\partial^2 f_\sigma(\boldsymbol{x})/\partial x_i \partial x_j$ for $i, j = 1, \ldots, n$ are all Lipschitz continuous in the spectral norm.

To this end, for each direction $\boldsymbol{d} \in \mathbb{S}^n$, let us consider the marginal function $h(\bullet; \boldsymbol{d}) : \mathbb{R}^n \to \mathbb{R}$ defined by $h(\boldsymbol{x}; \boldsymbol{d}) := \boldsymbol{d}^{\mathrm{T}} \nabla f(\boldsymbol{x})$; as a sanity check, we have $\nabla h(\boldsymbol{x}; \boldsymbol{d}) = \nabla^2 f(\boldsymbol{x})\boldsymbol{d}$ a.e. (In passing, we note that as f is L-smooth and thus prox-regular (see, e.g., (Rockafellar & Wets, 2009, Proposition 13.34)), it follows from (Rockafellar & Wets, 2009, Corollary 13.42) that $\nabla^2 f(\boldsymbol{x})$ must be symmetric at every point where it exists.) As ∇f is L-Lipschitz as assumed, it follows from (Rockafellar & Wets, 2009, Exercise 9.9) that all the functions $h(\bullet; \boldsymbol{d})$ are L-Lipschitz. As a result, we know from (Lin et al., 2022, Proposition 2.3) that for every $\boldsymbol{d} \in \mathbb{S}^n$, the function $h_\sigma(\bullet; \boldsymbol{d}) : \mathbb{R}^n \to \mathbb{R}$ defined by $h_\sigma(\boldsymbol{x}; \boldsymbol{d}) := \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[h(\boldsymbol{x} + \sigma\boldsymbol{u}; \boldsymbol{d})]$ is $cL\sqrt{n}/\sigma$-smooth for some fixed $c > 0$. Besides, by definition, we have

$$h_\sigma(\boldsymbol{x}; \boldsymbol{d}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\boldsymbol{d}^{\mathrm{T}} \nabla f(\boldsymbol{x} + \sigma\boldsymbol{u})] = \boldsymbol{d}^{\mathrm{T}} \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x} + \sigma\boldsymbol{u})] = \boldsymbol{d}^{\mathrm{T}} \nabla f_\sigma(\boldsymbol{x}),$$

which, together with the twice differentiability of f_σ shown earlier, further implies that $h_\sigma(\bullet; \boldsymbol{d})$ is differentiable with $\nabla h_\sigma(\boldsymbol{x}; \boldsymbol{d}) = \nabla^2 f_\sigma(\boldsymbol{x})\boldsymbol{d}$. As a result, we have

$$\begin{aligned}
\|\nabla^2 f_\sigma(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{y})\| &= \max_{\boldsymbol{d}\in\mathbb{S}^n} \|\nabla^2 f_\sigma(\boldsymbol{x})\boldsymbol{d} - \nabla^2 f_\sigma(\boldsymbol{y})\boldsymbol{d}\| \\
&= \max_{\boldsymbol{d}\in\mathbb{S}^n} \|\nabla h_\sigma(\boldsymbol{x}; \boldsymbol{d}) - \nabla h_\sigma(\boldsymbol{y}; \boldsymbol{d})\| \le \frac{cL\sqrt{n}}{\sigma}\|\boldsymbol{x} - \boldsymbol{y}\| \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,
\end{aligned}$$

as desired. □

C Gradient and Hessian estimators

C.1 Gradient estimator

Recall that $\nabla f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})]$, and thus

$$\nabla f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}) + \nabla f(\boldsymbol{x}-\sigma\boldsymbol{u})}{2}\right],$$

as $\boldsymbol{u}$ and $-\boldsymbol{u}$ are equal in distribution. This suggests the estimator in Proposition 3.2.

Proof of Proposition 3.2. We have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_b\overset{\mathrm{iid}}{\sim}\mathrm{Unif}(\mathbb{B}^n)}[\widehat{\boldsymbol{g}}_b(\boldsymbol{x})] &= \mathbb{E}_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_b\overset{\mathrm{iid}}{\sim}\mathrm{Unif}(\mathbb{B}^n)}\left[\frac{1}{b}\sum_{j=1}^b \frac{\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}_j) + \nabla f(\boldsymbol{x}-\sigma\boldsymbol{u}_j)}{2}\right] \\ &= \frac{1}{b}\sum_{j=1}^b \mathbb{E}_{\boldsymbol{u}_j\sim\mathrm{Unif}(\mathbb{B}^n)}\left[\frac{\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}_j) + \nabla f(\boldsymbol{x}-\sigma\boldsymbol{u}_j)}{2}\right] = \nabla f_\sigma(\boldsymbol{x});\end{aligned}$$

i.e., $\widehat{\boldsymbol{g}}_b(\boldsymbol{x})$ is unbiased.

Let us write $\widehat{\boldsymbol{g}}_b(\boldsymbol{x}) = \frac{1}{b}\sum_{j=1}^b \widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x})$. We shall apply the vector Bernstein inequality (see, e.g., (Cabannes et al., 2021, Theorem 32) and (Pinelis & Sakhanenko, 1986)) to obtain the concentration bound. To this end, we first observe that

$$\|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\| \le \|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| + \|\nabla f(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\|.$$

On the one hand,

$$\|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| = \left\|\frac{\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}_j) + \nabla f(\boldsymbol{x}-\sigma\boldsymbol{u}_j)}{2} - \nabla f(\boldsymbol{x})\right\| \le L\|\sigma\boldsymbol{u}_j\| \le L\sigma. \quad (3)$$

On the other hand,

$$\begin{aligned}\|\nabla f(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\| &= \|\nabla f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})]\| \\ &= \|\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})]\| \\ &\le \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})\|] \le \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[L\|\sigma\boldsymbol{u}\|] \le L\sigma.\end{aligned}$$

By putting the above two pieces together, it follows that

$$\|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\| \le 2L\sigma.$$

Besides, as the mean minimizes the mean-square error, we have

$$\mathbb{E}_{\boldsymbol{u}_j\sim\mathrm{Unif}(\mathbb{B}^n)}[\|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\|^2] \le \mathbb{E}_{\boldsymbol{u}_j\sim\mathrm{Unif}(\mathbb{B}^n)}[\|\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2] \le L^2\sigma^2,$$

where the last inequality follows from (3). As a result, the vector Bernstein inequality implies that

$$\mathbb{P}\left\{\left\|\sum_{j=1}^b \left(\widehat{\boldsymbol{g}}_{b,j}(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\right)\right\| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2bL^2\sigma^2 + 4tL\sigma/3}\right) \quad \text{for all } t > 0,$$

from which the desired concentration bound immediately follows. We are done. □

C.2 Hessian estimator

While $\nabla^2 f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^n)}[\nabla^2 f(\boldsymbol{x}+\sigma\boldsymbol{u})]$ also suggests a second-order estimator for $\nabla^2 f_\sigma(\boldsymbol{x})$, as $\nabla^2 f(\boldsymbol{x})$ may not be available or can be expensive in practice, and to match the order of information used by the gradient estimator $\widehat{\boldsymbol{g}}_b(\boldsymbol{x})$ in Proposition 3.2, we resort to a first-order Hessian estimator. To this end, we begin with a technical tool connecting $\nabla^2 f_\sigma(\boldsymbol{x})$ with $\nabla f(\boldsymbol{x})$ instead, which can also be taken as a second-order parallel of (Flaxman et al., 2005, Lemma 2.1).

Proposition C.1 *Let $L, \sigma > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. We have*

$$\nabla^2 f_\sigma(\boldsymbol{x}) = \frac{n}{\sigma} \mathbb{E}_{\boldsymbol{s} \sim \mathrm{Unif}(\mathbb{S}^n)}[\nabla f(\boldsymbol{x} + \sigma \boldsymbol{s}) \boldsymbol{s}^{\mathrm{T}}].$$

Proof. Given $\boldsymbol{v} \in \mathbb{R}^n$, let us consider the vector field $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ given by $\boldsymbol{F}(\boldsymbol{z}) := f(\boldsymbol{z})\boldsymbol{v}$. It holds that $\operatorname{div} \boldsymbol{F}(\boldsymbol{z}) = \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{z})$. As a result, the divergence theorem (see, e.g., (Folland, 2024, Section 5.9)) implies that

$$\int_{\boldsymbol{x} + \sigma \mathbb{B}^n} \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\boldsymbol{x} + \sigma \mathbb{B}^n} \operatorname{div} \boldsymbol{F}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \frac{1}{\sigma} \int_{\boldsymbol{x} + \sigma \mathbb{S}^n} f(\boldsymbol{z}) \boldsymbol{v}^{\mathrm{T}} (\boldsymbol{z} - \boldsymbol{x}) S(\mathrm{d}\boldsymbol{z}),$$

where S is the surface measure on $\boldsymbol{x} + \sigma \mathbb{S}^n$; i.e., the restriction of the Hausdorff measure to $\boldsymbol{x} + \sigma \mathbb{S}^n$ (see, e.g., (Evans & Gariepy, 2015, Section 3.3.4(D))). As $\boldsymbol{v}$ is arbitrary, this further implies that

$$\sigma^n \operatorname{vol}(\mathbb{B}^n) \nabla f_\sigma(\boldsymbol{x}) = \int_{\boldsymbol{x} + \sigma \mathbb{B}^n} \nabla f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \frac{1}{\sigma} \int_{\boldsymbol{x} + \sigma \mathbb{S}^n} f(\boldsymbol{z})(\boldsymbol{z} - \boldsymbol{x}) S(\mathrm{d}\boldsymbol{z}).$$

Consider the change of variables $\boldsymbol{z} = \boldsymbol{x} + \sigma \boldsymbol{s}$ where $\boldsymbol{s} \in \mathbb{S}^n$. Then, $S(\mathrm{d}\boldsymbol{z}) = \sigma^{n-1} S(\mathrm{d}\boldsymbol{s})$; see, e.g., (Evans & Gariepy, 2015, Theorem 2.2(iv–v)). As a result, it follows from (Bogachev, 2007, Theorem 3.6.1) that

$$\int_{\boldsymbol{x} + \sigma \mathbb{S}^n} f(\boldsymbol{z})(\boldsymbol{z} - \boldsymbol{x}) S(\mathrm{d}\boldsymbol{z}) = \sigma^n \int_{\mathbb{S}^n} f(\boldsymbol{x} + \sigma \boldsymbol{s}) \boldsymbol{s} S(\mathrm{d}\boldsymbol{s}),$$

and thus

$$\begin{aligned} \nabla f_\sigma(\boldsymbol{x}) &= \frac{1}{\sigma \operatorname{vol}(\mathbb{B}^n)} \int_{\mathbb{S}^n} f(\boldsymbol{x} + \sigma \boldsymbol{s}) \boldsymbol{s} S(\mathrm{d}\boldsymbol{s}) \\ &= \frac{n}{\sigma \operatorname{vol}(\mathbb{S}^n)} \int_{\mathbb{S}^n} f(\boldsymbol{x} + \sigma \boldsymbol{s}) \boldsymbol{s} S(\mathrm{d}\boldsymbol{s}) = \frac{n}{\sigma} \int_{\mathbb{S}^n} f(\boldsymbol{x} + \sigma \boldsymbol{s}) \boldsymbol{s} \mathbb{P}_{\boldsymbol{s} \sim \mathrm{Unif}(\mathbb{S}^n)}(\mathrm{d}\boldsymbol{s}). \end{aligned}$$

This, together with (Folland, 1999, Theorem 2.27(b)), further implies that

$$\nabla^2 f_\sigma(\boldsymbol{x}) = \frac{n}{\sigma} \int_{\mathbb{S}^n} \boldsymbol{s} \nabla f(\boldsymbol{x} + \sigma \boldsymbol{s})^{\mathrm{T}} \mathbb{P}_{\boldsymbol{s} \sim \mathrm{Unif}(\mathbb{S}^n)}(\mathrm{d}\boldsymbol{s}) = \frac{n}{\sigma} \mathbb{E}_{\boldsymbol{s} \sim \mathrm{Unif}(\mathbb{S}^n)}[\boldsymbol{s} \nabla f(\boldsymbol{x} + \sigma \boldsymbol{s})^{\mathrm{T}}].$$

As $\nabla^2 f_\sigma(\boldsymbol{x})$ is symmetric, the desired result then follows by taking transpose on both sides. □

As $\boldsymbol{s}$ and $-\boldsymbol{s}$ equal in distribution, Proposition C.1 further implies that

$$\nabla^2 f_\sigma(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{s} \sim \mathrm{Unif}(\mathbb{S}^n)} \Big[\frac{n}{2\sigma} \big(\nabla f(\boldsymbol{x} + \sigma \boldsymbol{s}) - \nabla f(\boldsymbol{x} - \sigma \boldsymbol{s}) \big) \boldsymbol{s}^{\mathrm{T}} \Big],$$

which motivates the estimator in Proposition 3.3.

Proof of Proposition 3.3. We have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_b \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\mathbb{S}^n)} [\widehat{\boldsymbol{H}}_b(\boldsymbol{x})] &= \mathbb{E}_{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_b \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(\mathbb{S}^n)} \left[\frac{1}{b} \sum_{j=1}^b \operatorname{Sym}\Big(\frac{n}{2\sigma} \big(\nabla f(\boldsymbol{x} + \sigma \boldsymbol{s}_j) - \nabla f(\boldsymbol{x} - \sigma \boldsymbol{s}_j) \big) \boldsymbol{s}_j^{\mathrm{T}} \Big) \right] \\ &= \frac{1}{b} \sum_{j=1}^b \operatorname{Sym}\Big(\mathbb{E}_{\boldsymbol{s}_j \sim \mathrm{Unif}(\mathbb{S}^n)} \Big[\frac{n}{2\sigma} \big(\nabla f(\boldsymbol{x} + \sigma \boldsymbol{s}_j) - \nabla f(\boldsymbol{x} - \sigma \boldsymbol{s}_j) \big) \boldsymbol{s}_j^{\mathrm{T}} \Big] \Big) \\ &= \frac{1}{b} \sum_{j=1}^b \operatorname{Sym}\big(\nabla^2 f_\sigma(\boldsymbol{x}) \big) = \operatorname{Sym}\big(\nabla^2 f_\sigma(\boldsymbol{x}) \big) = \nabla^2 f_\sigma(\boldsymbol{x}); \end{aligned}$$

i.e., $\widehat{\boldsymbol{H}}_b(\boldsymbol{x})$ is unbiased.

Let us write $\widehat{\boldsymbol{H}}_b(\boldsymbol{x}) = \frac{1}{b}\sum_{j=1}^{b}\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x})$. We shall apply the matrix Bernstein inequality (see, e.g., (Vershynin, 2018, Theorem 5.4.1)) to obtain the concentration bound. To this end, we first observe that

$$\big\|\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big\| \le \big\|\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x})\big\| + \big\|\nabla^2 f_\sigma(\boldsymbol{x})\big\|.$$

On the one hand,

$$\begin{aligned}\big\|\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x})\big\| &= \Big\|\mathrm{Sym}\Big(\frac{n}{2\sigma}\big(\nabla f(\boldsymbol{x}+\sigma\boldsymbol{s}_j) - \nabla f(\boldsymbol{x}-\sigma\boldsymbol{s}_j)\big)\boldsymbol{s}_j^{\mathrm{T}}\Big)\Big\| \\ &\le \Big\|\frac{n}{2\sigma}\big(\nabla f(\boldsymbol{x}+\sigma\boldsymbol{s}_j) - \nabla f(\boldsymbol{x}-\sigma\boldsymbol{s}_j)\big)\boldsymbol{s}_j^{\mathrm{T}}\Big\| \\ &= \frac{n}{2\sigma}\big\|\nabla f(\boldsymbol{x}+\sigma\boldsymbol{s}_j) - \nabla f(\boldsymbol{x}-\sigma\boldsymbol{s}_j)\big\| \cdot \|\boldsymbol{s}_j\| \le \frac{n}{2\sigma}L\|2\sigma\boldsymbol{s}_j\| = nL,\end{aligned} \tag{4}$$

where we have used the fact that $\|\boldsymbol{A}\| = \|\boldsymbol{A}^{\mathrm{T}}\|$ for all matrices $\boldsymbol{A}$ (which may be rectangular and not symmetric). On the other hand, as f_σ is L-smooth and twice continuously differentiable, it follows from (Rockafellar & Wets, 2009, Theorem 9.7) that $\big\|\nabla^2 f_\sigma(\boldsymbol{x})\big\| \le L$ for all $\boldsymbol{x} \in \mathbb{R}^n$. By putting the above two pieces together, it follows that

$$\big\|\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big\| \le (n+1)L.$$

Besides, by Jensen's inequality (see, e.g., (Durrett, 2019, Theorem 1.6.2)), we have

$$\begin{aligned}\Big\|\mathbb{E}_{\boldsymbol{s}_j\sim\mathrm{Unif}(\mathbb{S}^n)}\Big[\big(\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big)^2\Big]\Big\| &\le \mathbb{E}_{\boldsymbol{s}_j\sim\mathrm{Unif}(\mathbb{S}^n)}\Big[\Big\|\big(\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big)^2\Big\|\Big] \\ &\le \mathbb{E}_{\boldsymbol{s}_j\sim\mathrm{Unif}(\mathbb{S}^n)}\Big[\Big\|\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\Big\|^2\Big] \\ &\le \mathbb{E}_{\boldsymbol{s}_j\sim\mathrm{Unif}(\mathbb{S}^n)}\big[(n+1)^2L^2\big] = (n+1)^2L^2.\end{aligned}$$

As a result, the matrix Bernstein inequality implies that

$$\mathbb{P}\left\{\left\|\sum_{j=1}^{b}\big(\widehat{\boldsymbol{H}}_{b,j}(\boldsymbol{x}) - \nabla^2 f_\sigma(\boldsymbol{x})\big)\right\| \ge t\right\} \le 2n\exp\Big(-\frac{t^2/2}{b(n+1)^2L^2 + (n+1)Lt/3}\Big) \quad \text{for all } t>0,$$

from which the desired concentration bound immediately follows. We are done. □

D CONNECTING $\nabla^2 f_\sigma(\boldsymbol{x})$ WITH $\operatorname{conv}\big(\bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n}\partial_{\mathrm{C}}^2 f(\boldsymbol{y})\big)$

In this section, our goal is to prove Proposition 3.4. To this end, we begin with some topological properties of $\partial_\delta f(\boldsymbol{x})$ and $\partial_\delta^2 f(\boldsymbol{x})$.

Lemma D.1 *Let $L, \delta > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. The sets $\partial_\delta f(\boldsymbol{x})$ and $\partial_\delta^2 f(\boldsymbol{x})$ are both nonempty, convex, and compact.*

Proof. As L-smooth functions are locally Lipschitz by (Rockafellar & Wets, 2009, Theorem 9.7) and (Rockafellar & Wets, 2009, Theorem 9.2), $\partial_\delta f(\boldsymbol{x})$ is compact by (Goldstein, 1977, Proposition 2.3). As $\partial_{\mathrm{C}} f(\boldsymbol{x}) \subseteq \partial_\delta f(\boldsymbol{x})$ and $\partial_{\mathrm{C}} f(\boldsymbol{x}) \neq \emptyset$ by (Cui & Pang, 2021, Proposition 4.3.1), we have $\partial_\delta f(\boldsymbol{x}) \neq \emptyset$. In a similar vein, as $\partial_{\mathrm{C}}^2 f(\boldsymbol{x}) \subseteq \partial_\delta^2 f(\boldsymbol{x})$ and $\partial_{\mathrm{C}}^2 f(\boldsymbol{x}) \neq \emptyset$ by (Clarke, 1990, Proposition 2.6.2(a)), it also follows that $\partial_\delta^2 f(\boldsymbol{x}) \neq \emptyset$. It remains to show that $\partial_\delta^2 f(\boldsymbol{x})$ is compact.

Towards that end, we begin by showing that $\bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n}\partial_{\mathrm{C}}^2 f(\boldsymbol{y})$ is bounded and closed. As ∇f is L-Lipschitz, by (Clarke, 1990, Proposition 2.6.2(d)), the sets $\partial_{\mathrm{C}}^2 f(\boldsymbol{y})$ are uniformly bounded for $\boldsymbol{y} \in \mathbb{R}^n$; thus, $\bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n}\partial_{\mathrm{C}}^2 f(\boldsymbol{y})$ is bounded. Besides, as $\boldsymbol{x}+\delta\mathbb{B}^n$ is compact and $\partial_{\mathrm{C}}^2 f$ is outer semicontinuous (Clarke, 1990, Proposition 2.6.2(b)), it follows from (Rockafellar & Wets, 2009, Theorem 5.25(a)) that $\bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\delta\mathbb{B}^n}\partial_{\mathrm{C}}^2 f(\boldsymbol{y})$ is closed, and thus compact. As a result, as the convex hull of any compact set is compact (see, e.g., (Barvinok, 2002, Corollary I.2.4)), it follows that $\partial_\delta^2 f(\boldsymbol{x})$ is compact, as desired. □

The next result comes without any surprise, as $\nabla f_\sigma(\boldsymbol{x})$ and $\nabla^2 f_\sigma(\boldsymbol{x})$ are the barycenters of the pushforward measures $(\nabla f)_{\#}\operatorname{Unif}(\boldsymbol{x}+\sigma\mathbb{B}^n)$ and $(\nabla^2 f)_{\#}\operatorname{Unif}(\boldsymbol{x}+\sigma\mathbb{B}^n)$; see Theorem 3.1.

Corollary D.2 *Let $L, \sigma > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. We have*

$$\nabla f_\sigma(\boldsymbol{x}) \in \partial_\sigma f(\boldsymbol{x}) \quad \text{and} \quad \nabla^2 f_\sigma(\boldsymbol{x}) \in \partial_\sigma^2 f(\boldsymbol{x}).$$

While the proof is almost verbatim that of (Lin et al., 2022, Theorem 3.1), we include it here for completeness.

Proof. If $\nabla f_\sigma(\boldsymbol{x}) \notin \partial_\sigma f(\boldsymbol{x})$, then (Rockafellar, 1997, Corollary 11.4.2) implies that there exists a $\boldsymbol{y} \in \mathbb{R}^n$ for which

$$\max_{\boldsymbol{z}\in\partial_\sigma f(\boldsymbol{x})} \boldsymbol{y}^{\mathrm{T}}\boldsymbol{z} < \boldsymbol{y}^{\mathrm{T}}\nabla f_\sigma(\boldsymbol{x}) = \boldsymbol{y}^{\mathrm{T}}\mathbb{E}_{\boldsymbol{u}\sim\operatorname{Unif}(\mathbb{B}^n)}[\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})] = \mathbb{E}_{\boldsymbol{u}\sim\operatorname{Unif}(\mathbb{B}^n)}[\boldsymbol{y}^{\mathrm{T}}\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})].$$

However, as

$$\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u}) \in \bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\sigma\mathbb{B}^n}\partial_{\mathrm{C}} f(\boldsymbol{y}) \subseteq \operatorname{conv}\left(\bigcup_{\boldsymbol{y}\in\boldsymbol{x}+\sigma\mathbb{B}^n}\partial_{\mathrm{C}} f(\boldsymbol{y})\right) = \partial_\sigma f(\boldsymbol{x})$$

whenever it exists, we have

$$\mathbb{E}_{\boldsymbol{u}\sim\operatorname{Unif}(\mathbb{B}^n)}[\boldsymbol{y}^{\mathrm{T}}\nabla f(\boldsymbol{x}+\sigma\boldsymbol{u})] \leq \mathbb{E}_{\boldsymbol{u}\sim\operatorname{Unif}(\mathbb{B}^n)}\left[\max_{\boldsymbol{z}\in\partial_\sigma f(\boldsymbol{x})}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{z}\right] = \max_{\boldsymbol{z}\in\partial_\sigma f(\boldsymbol{x})}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{z},$$

a contradiction. In a similar vein, if $\nabla^2 f_\sigma(\boldsymbol{x}) \notin \partial_\sigma^2 f(\boldsymbol{x})$, then there exists a $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ such that

$$\max_{\boldsymbol{Z}\in\partial_\sigma^2 f(\boldsymbol{x})}\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle < \langle \boldsymbol{Y}, \nabla^2 f_\sigma(\boldsymbol{x})\rangle = \mathbb{E}_{\boldsymbol{u}\sim\operatorname{Unif}(\mathbb{B}^n)}[\langle \boldsymbol{Y}, \nabla^2 f(\boldsymbol{x}+\sigma\boldsymbol{u})\rangle] \leq \max_{\boldsymbol{Z}\in\partial_\sigma^2 f(\boldsymbol{x})}\langle \boldsymbol{Y}, \boldsymbol{Z}\rangle,$$

a contradiction. □

The last technical preparation is the following lemma, which establishes a fundamental connection between the notion of Goldstein FOSPs and that of (ordinary) FOSPs for L-smooth functions.

Lemma D.3 ((Zhang et al., 2020, Proposition 6(ii))) *Let $L, \varepsilon > 0$ be arbitrary and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. If $\operatorname{dist}\big(\mathbf{0}, \partial_{\frac{\varepsilon}{3L}} f(\boldsymbol{x})\big) \leq \varepsilon/3$, then $\|\nabla f(\boldsymbol{x})\| \leq \varepsilon$.*

We are now in a position to prove Proposition 3.4.

Proof of Proposition 3.4. As $\sigma \leq \varepsilon_1/(3L)$, Corollary D.2 implies that $\nabla f_\sigma(\boldsymbol{x}) \in \partial_\sigma f(\boldsymbol{x}) \subseteq \partial_{\frac{\varepsilon_1}{3L}} f(\boldsymbol{x})$, and thus

$$\operatorname{dist}\Big(\mathbf{0}, \partial_{\frac{\varepsilon_1}{3L}} f(\boldsymbol{x})\Big) \leq \operatorname{dist}\big(\mathbf{0}, \partial_\sigma f(\boldsymbol{x})\big) \leq \|\nabla f_\sigma(\boldsymbol{x})\| \leq \varepsilon_g = \frac{\varepsilon_1}{3};$$

i.e., $\|\nabla f(\boldsymbol{x})\| \leq \varepsilon_1$ by Lemma D.3. On the other hand, as $\sigma \leq \delta$, Corollary D.2 also implies that $\nabla^2 f_\sigma(\boldsymbol{x}) \in \partial^2_\sigma f(\boldsymbol{x}) \subseteq \partial^2_\delta f(\boldsymbol{x})$, and thus

$$\min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{A} \in \partial^2_\delta f(\boldsymbol{x})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w} \rangle \geq \min_{\boldsymbol{w} \in \mathbb{S}^n} \max_{\boldsymbol{A} \in \partial^2_\sigma f(\boldsymbol{x})} \langle \boldsymbol{A}\boldsymbol{w}, \boldsymbol{w} \rangle \geq \min_{\boldsymbol{w} \in \mathbb{S}^n} \boldsymbol{w}^{\mathrm{T}} \nabla^2 f_\sigma(\boldsymbol{x}) \boldsymbol{w} \geq -\varepsilon_H = -\varepsilon_2,$$

as desired. □

E Proof of Theorem 4.1: Computing SOSPs of f_σ

For ease of exposition in the analysis of Algorithm 1, we introduce the following shorthands:

$$r_k := \|\boldsymbol{p}_k\|, \quad \boldsymbol{q}_k := \widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \nabla f_\sigma(\boldsymbol{x}_k), \quad \boldsymbol{Q}_k := \widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k),$$
$$\alpha_k := \|\boldsymbol{q}_k\|, \quad \delta_k := \|\boldsymbol{Q}_k\|.$$

In what follows, we shall prove in a discussion manner the theorem below, which, together with Proposition 3.4, immediately implies Theorem 4.1.

Theorem E.1 *Consider the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 1. With probability at least $1-\eta$, there exists some $k_\star = 0, \ldots, K_\star - 1$ such that*

$$\|\nabla f_\sigma(\boldsymbol{x}_{k_\star+1})\| \le \varepsilon_g \quad \text{and} \quad \lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_{k_\star+1})\big) \ge -\varepsilon_H.$$

To this end, we first recall (Cartis et al., 2011, Theorem 3.1) (adapted to our setting), which characterizes the global minimizers of (2).

Lemma E.2 ((Cartis et al., 2011, Theorem 3.1)) *A vector $\boldsymbol{p} \in \mathbb{R}^n$ is a global minimizer of (2) if and only if*

$$\big(\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) + \lambda \boldsymbol{I}\big)\boldsymbol{p} = -\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) \quad \text{and} \quad \widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) + \lambda \boldsymbol{I} \succeq \boldsymbol{O}, \quad \text{where } \lambda := \xi\|\boldsymbol{p}\|/2.$$

Proof. It suffices to note that $\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)$ is symmetric, and then apply (Cartis et al., 2011, Theorem 3.1). □

We begin the analysis by understanding the model decrease for (2). While it is also known, we nonetheless include a short proof here to familiarize the readers with the notation.

Lemma E.3 ((Nesterov & Polyak, 2006, Lemma 4)) *We have $\widehat{m}_k(\boldsymbol{p}_k) \le -\xi r_k^3/12$.*

Proof. As $\boldsymbol{p}_k$ is a global minimizer of (2), it follows from Lemma E.2 that

$$\big(\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) + \lambda \boldsymbol{I}\big)\boldsymbol{p}_k + \widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) = \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p}_k + \lambda r_k^2 + \boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) = 0,$$

where $\lambda = \xi r_k/2$; i.e.,

$$\boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) = -\boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p}_k - \frac{\xi}{2}r_k^3.$$

Hence,

$$\widehat{m}_k(\boldsymbol{p}_k) = \boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) + \frac{1}{2}\boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p}_k + \frac{\xi}{6}r_k^3 = -\frac{1}{2}\boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p}_k - \frac{\xi}{3}r_k^3.$$

However, Lemma E.2 also implies that

$$\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) \succeq -\xi r_k \boldsymbol{I}/2 \quad \Longrightarrow \quad \boldsymbol{p}_k^{\mathrm{T}}\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\boldsymbol{p}_k \ge -\xi r_k^3/2.$$

As a result, it follows that

$$\widehat{m}_k(\boldsymbol{p}_k) \le \frac{\xi}{4}r_k^3 - \frac{\xi}{3}r_k^3 = -\frac{\xi}{12}r_k^3,$$

as desired. □

With the model decrease well understood, we next figure out how it transfers to a decrease in f_σ.

Lemma E.4 *We have $f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k) \le f_\sigma(\boldsymbol{x}_k) - \xi r_k^3/12 + \alpha_k r_k + \delta_k r_k^2/2$.*

In comparison with (Nesterov & Polyak, 2006, Lemma 4), we see that the extra term $\alpha_k r_k + \delta_k r_k^2/2$ is exactly the price we need to pay for using inexact gradients and Hessians.

Proof. As $\nabla^2 f_\sigma$ is ξ-Lipschitz (w.r.t. the spectral norm), it follows from (Nesterov & Polyak, 2006, Lemma 1) that

$$\begin{aligned} f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k) &\le f_\sigma(\boldsymbol{x}_k) + \boldsymbol{p}_k^{\mathrm{T}} \nabla f_\sigma(\boldsymbol{x}_k) + \frac{1}{2} \boldsymbol{p}_k^{\mathrm{T}} \nabla^2 f_\sigma(\boldsymbol{x}_k) \boldsymbol{p}_k + \frac{\xi}{6} r_k^3 \\ &= f_\sigma(\boldsymbol{x}_k) + \boldsymbol{p}_k^{\mathrm{T}} \big(\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \boldsymbol{q}_k\big) + \frac{1}{2} \boldsymbol{p}_k^{\mathrm{T}} \big(\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \boldsymbol{Q}_k\big) \boldsymbol{p}_k + \frac{\xi}{6} r_k^3 \\ &= f_\sigma(\boldsymbol{x}_k) + \widehat{m}_k(\boldsymbol{p}_k) - \boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{q}_k - \frac{1}{2} \boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{Q}_k \boldsymbol{p}_k \\ &\le f_\sigma(\boldsymbol{x}_k) - \frac{\xi r_k^3}{12} + r_k \|\boldsymbol{q}_k\| + \frac{1}{2} r_k^2 \|\boldsymbol{Q}_k\|, \end{aligned}$$

where the second inequality follows from Lemma E.3 and the Cauchy–Schwarz inequality. □

With Lemma E.4 in place, we next turn to bound the gradient and Hessian of f_σ at $\boldsymbol{x}_k + \boldsymbol{p}_k$.

Lemma E.5 *We have* $\|\nabla f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k)\| \le \alpha_k + \delta_k r_k + \xi r_k^2$ *and* $\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k)\big) \ge -\delta_k - 3\xi r_k/2$.

Proof. On the one hand, by Taylor's Theorem (see, e.g., (Nocedal & Wright, 2006, Theorem 2.1)), we have

$$\begin{aligned} \nabla f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k) &= \nabla f_\sigma(\boldsymbol{x}_k) + \int_0^1 \nabla^2 f_\sigma(\boldsymbol{x}_k + t\boldsymbol{p}_k) \boldsymbol{p}_k \mathrm{d}t \\ &= \widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \boldsymbol{q}_k + \big(\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \boldsymbol{Q}_k\big) \boldsymbol{p}_k + \int_0^1 \big(\nabla^2 f_\sigma(\boldsymbol{x}_k + t\boldsymbol{p}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\big) \boldsymbol{p}_k \mathrm{d}t \\ &= -\frac{\xi r_k}{2} \boldsymbol{p}_k - \boldsymbol{q}_k - \boldsymbol{Q}_k \boldsymbol{p}_k + \int_0^1 \big(\nabla^2 f_\sigma(\boldsymbol{x}_k + t\boldsymbol{p}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\big) \boldsymbol{p}_k \mathrm{d}t, \end{aligned}$$

and thus

$$\begin{aligned} \|\nabla f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k)\| &\le \frac{\xi}{2} r_k^2 + \alpha_k + \delta_k r_k + \left\| \int_0^1 \big(\nabla^2 f_\sigma(\boldsymbol{x}_k + t\boldsymbol{p}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\big) \boldsymbol{p}_k \mathrm{d}t \right\| \\ &\le \frac{\xi}{2} r_k^2 + \alpha_k + \delta_k r_k + r_k \int_0^1 \big\|\nabla^2 f_\sigma(\boldsymbol{x}_k + t\boldsymbol{p}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\big\| \mathrm{d}t \\ &\le \frac{\xi}{2} r_k^2 + \alpha_k + \delta_k r_k + \xi r_k^2 \int_0^1 t \mathrm{d}t = \alpha_k + \delta_k r_k + \xi r_k^2. \end{aligned}$$

On the other hand, as $\|\nabla^2 f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\| \le \xi r_k$, Weyl's inequality implies that

$$\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k)\big) \ge \lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_k)\big) - \xi r_k;$$

see, e.g., the second exercise after (Horn & Johnson, 2012, Corollary 6.3.4). However, it also follows from Lemma E.2 that $\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) \succeq -\xi r_k \boldsymbol{I}/2$, and thus by Weyl's inequality again,

$$\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_k)\big) \ge \lambda_{\min}\big(\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k)\big) - \delta_k \ge -\frac{\xi}{2} r_k - \delta_k.$$

By putting the above two pieces together, we have

$$\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_k + \boldsymbol{p}_k)\big) \ge -\frac{\xi}{2} r_k - \delta_k - \xi r_k = -\delta_k - \frac{3\xi}{2} r_k,$$

as desired. □

We are now at a position to show the soundness of Algorithm 1, conditional on the event that the estimators are sufficiently accurate.

Proposition E.6 *Consider the iterates $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{K_\star}$ generated by Algorithm 1. Conditional on*

$$\mathcal{E} := \bigcap_{k=0,\dots,K_\star-1} \left\{ \|\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \nabla f_\sigma(\boldsymbol{x}_k)\| \le \frac{\xi r_\star^2}{48} \text{ and } \|\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\| \le \frac{\xi r_\star}{24} \right\}, \quad (5)$$

there must exist some $k_\star = 0, \dots, K_\star - 1$ such that

$$\|\nabla f_\sigma(\boldsymbol{x}_{k_\star+1})\| \le \varepsilon_g \quad \text{and} \quad \lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_{k_\star+1})\big) \ge -\varepsilon_H.$$

Proof. If $r_k > r_\star$ for some $k = 0, \dots, K_\star - 1$, then Lemma E.4 implies that

$$\begin{aligned} f_\sigma(\boldsymbol{x}_{k+1}) &\le f_\sigma(\boldsymbol{x}_k) - \frac{\xi}{12} r_k^3 + \alpha_k r_k + \frac{\delta_k}{2} r_k^2 \\ &\le f_\sigma(\boldsymbol{x}_k) - \frac{\xi}{12} r_k^3 + \frac{\xi r_\star^2}{48} r_k + \frac{\xi r_\star}{48} r_k^2 < f_\sigma(\boldsymbol{x}_k) - \frac{\xi}{24} r_k^3 < f_\sigma(\boldsymbol{x}_k) - \frac{\xi r_\star^3}{24}; \end{aligned}$$

i.e., the value of f_σ can be decreased by at least $\xi r_\star^3/24$. Thus, there must exist some $k_\star = 0, \dots, K_\star - 1$ such that $r_{k_\star} \le r_\star$, as otherwise we would have

$$\Delta \ge f_\sigma(\boldsymbol{x}_0) - f_\sigma(\boldsymbol{x}_{K_\star}) = \sum_{k=0,\dots,K_\star-1} \big(f_\sigma(\boldsymbol{x}_k) - f_\sigma(\boldsymbol{x}_{k+1})\big) > K_\star \frac{\xi r_\star^3}{24} \ge \frac{24\Delta}{\xi r_\star^3} \cdot \frac{\xi r_\star^3}{24} = \Delta,$$

a contradiction. It then immediately follows from Lemma E.5 that

$$\|\nabla f_\sigma(\boldsymbol{x}_{k_\star+1})\| \le \alpha_{k_\star} + \delta_{k_\star} r_{k_\star} + \xi r_{k_\star}^2 \le \frac{\xi r_\star^2}{48} + \frac{\xi r_\star}{24} r_{k_\star} + \xi r_{k_\star}^2 \le \frac{17 \xi r_\star^2}{16} \le \varepsilon_g,$$

$$\lambda_{\min}\big(\nabla^2 f_\sigma(\boldsymbol{x}_{k_\star+1})\big) \ge -\delta_{k_\star} - \frac{3\xi}{2} r_{k_\star} \ge -\frac{37 \xi r_\star}{24} \ge -\varepsilon_H,$$

as desired. □

To show Theorem E.1, it only remains to show that $\mathcal{E}$ (5) happens with high probability.

Proposition E.7 *We have $\mathbb{P}(\mathcal{E}) \ge 1 - \eta$.*

Proof. Let $\mathcal{F}_{k-1}$ be the filtration generated by all randomness up to iteration $k-1$ (inclusive). By Proposition 3.2 and Proposition 3.3, for all $k = 0, \dots, K_\star - 1$, we have

$$\mathbb{P}\left\{ \|\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \nabla f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star^2}{48} \middle| \mathcal{F}_{k-1} \right\} \le 2 \exp\left(- \frac{b_g \xi^2 r_\star^4 / 2304}{2L^2\sigma^2 + L\sigma\xi r_\star^2/36} \right) \le 2 \exp\left(\log\left(\frac{\eta}{4K_\star} \right) \right) = \frac{\eta}{2K_\star},$$

$$\begin{aligned} \mathbb{P}\left\{ \|\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star}{24} \middle| \mathcal{F}_{k-1} \right\} &\le 2n \exp\left(- \frac{b_H \xi^2 r_\star^2 / 1152}{(n+1)^2 L^2 + (n+1) L \xi r_\star / 72} \right) \\ &\le 2n \exp\left(\log\left(\frac{\eta}{4nK_\star} \right) \right) = \frac{\eta}{2K_\star}, \end{aligned}$$

and thus by the tower property of conditional probability,

$$\mathbb{P}\left\{ \|\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \nabla f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star^2}{48} \right\} \le \frac{\eta}{2K_\star} \quad \text{and} \quad \mathbb{P}\left\{ \|\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star}{24} \right\} \le \frac{\eta}{2K_\star}.$$

It then follows by applying a union bound that

$$\mathbb{P}(\mathcal{E}^\complement) \le \sum_{k=0}^{K_\star-1} \mathbb{P}\left\{ \|\widehat{\boldsymbol{g}}_{b_g}(\boldsymbol{x}_k) - \nabla f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star^2}{48} \right\} + \sum_{k=0}^{K_\star-1} \mathbb{P}\left\{ \|\widehat{\boldsymbol{H}}_{b_H}(\boldsymbol{x}_k) - \nabla^2 f_\sigma(\boldsymbol{x}_k)\| > \frac{\xi r_\star}{24} \right\} \le \eta;$$

i.e., $\mathbb{P}(\mathcal{E}) \ge 1 - \eta$, as desired. □

F Proof of Theorem 5.3

Theorem 5.3 follows immediately from the following proposition, which identifies the second subderivatives of Moreau envelopes with the Moreau envelopes of second subderivatives at FOSPs of weakly convex functions, offering insights complementary to, e.g., (Rockafellar & Wets, 2009, Exercise 13.45) (with "for some $\lambda > 0$ sufficiently small" sharpened) and (Hang et al., 2022, Proposition 4.1) (from convexity to weak convexity, and without twice epi-differentiability).

Proposition F.1 *Let $\rho > 0$ and $\lambda \in (0, 1/\rho)$ be arbitrary, and $f : \mathbb{R}^n \to \mathbb{R}$ be ρ-weakly convex. If $\boldsymbol{x} \in \mathbb{R}^n$ is an FOSP of f (i.e., $\mathbf{0} \in \partial f(\boldsymbol{x})$), then*

$$\frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x}; \mathbf{0})(\boldsymbol{w}) = \left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x}; \mathbf{0})\right)_\lambda (\boldsymbol{w}) \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^n.$$

Proof. Upon relocating the graph of f to the origin, we may assume without loss of generality that $\boldsymbol{x} = \mathbf{0}$ and $f(\boldsymbol{x}) = 0$. We begin with the "$\leq$" direction. As $\mathbf{0} \in \partial f(\boldsymbol{x})$, it follows from (Davis & Drusvyatskiy, 2022, Lemma 2.5(4)) that $\mathrm{Prox}_{\lambda f}(\boldsymbol{x}) = \boldsymbol{x}$, and thus

$$f_\lambda(\boldsymbol{x}) = f\big(\mathrm{Prox}_{\lambda f}(\boldsymbol{x})\big) + \frac{1}{2\lambda}\big\|\mathrm{Prox}_{\lambda f}(\boldsymbol{x}) - \boldsymbol{x}\big\|^2 = f(\boldsymbol{x}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{x}\|^2 = f(\boldsymbol{x}) = 0.$$

As a result, for all $t > 0$ and $\boldsymbol{w}' \in \mathbb{R}^n$, we have

$$\begin{aligned}
\frac{f_\lambda(\boldsymbol{x} + t\boldsymbol{w}') - f_\lambda(\boldsymbol{x}) - t\mathbf{0}^{\mathrm{T}}\boldsymbol{w}'}{t^2/2} &= \frac{f_\lambda(t\boldsymbol{w}')}{t^2/2} \\
&= \frac{2}{t^2}\min\left\{f(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - t\boldsymbol{w}'\|^2 : \boldsymbol{y} \in \mathbb{R}^n\right\} \\
&= \frac{2}{t^2}\min\left\{f(t\boldsymbol{y}) + \frac{t^2}{2\lambda}\|\boldsymbol{y} - \boldsymbol{w}'\|^2 : \boldsymbol{y} \in \mathbb{R}^n\right\} \\
&= \min\left\{\frac{f(t\boldsymbol{y})}{t^2/2} + \frac{1}{\lambda}\|\boldsymbol{y} - \boldsymbol{w}'\|^2 : \boldsymbol{y} \in \mathbb{R}^n\right\}.
\end{aligned} \tag{6}$$

Now, for all $\boldsymbol{u} \in \mathbb{R}^n$, as

$$\mathrm{d}^2 f(\boldsymbol{x}; \mathbf{0})(\boldsymbol{u}) = \liminf_{t \searrow 0,\, \boldsymbol{u}' \to \boldsymbol{u}} \frac{f(\boldsymbol{x} + t\boldsymbol{u}') - f(\boldsymbol{x}) - t\mathbf{0}^{\mathrm{T}}\boldsymbol{u}'}{t^2/2},$$

it follows from (Rockafellar & Wets, 2009, Lemma 1.7) that there must exist a sequence $(t_1, \boldsymbol{u}_1), (t_2, \boldsymbol{u}_2), \ldots$ with $t_k \searrow 0$ and $\boldsymbol{u}_k \to \boldsymbol{u}$ as $k \to \infty$ such that

$$\mathrm{d}^2 f(\boldsymbol{x}; \mathbf{0})(\boldsymbol{u}) = \lim_{k \to \infty} \frac{f(\boldsymbol{x} + t_k\boldsymbol{u}_k) - f(\boldsymbol{x}) - t_k\mathbf{0}^{\mathrm{T}}\boldsymbol{u}_k}{t_k^2/2} = \lim_{k \to \infty} \frac{f(t_k\boldsymbol{u}_k)}{t_k^2/2}.$$

On the other hand, for all $\boldsymbol{w} \in \mathbb{R}^n$, by letting $t = t_k$ and $\boldsymbol{w}' = \boldsymbol{w}$ in (6) and taking $\boldsymbol{u}_k$ as a feasible solution to (6), it also follows that

$$\frac{f_\lambda(\boldsymbol{x} + t_k\boldsymbol{w}) - f_\lambda(\boldsymbol{x}) - t_k\mathbf{0}^{\mathrm{T}}\boldsymbol{w}}{t_k^2/2} = \min\left\{\frac{f(t_k\boldsymbol{y})}{t_k^2/2} + \frac{1}{\lambda}\|\boldsymbol{y} - \boldsymbol{w}\|^2 : \boldsymbol{y} \in \mathbb{R}^n\right\} \leq \frac{f(t_k\boldsymbol{u}_k)}{t_k^2/2} + \frac{1}{\lambda}\|\boldsymbol{u}_k - \boldsymbol{w}\|^2.$$

By putting the above two pieces together, we have

$$\begin{aligned}
\liminf_{t \searrow 0,\, \boldsymbol{w}' \to \boldsymbol{w}} \frac{f_\lambda(\boldsymbol{x} + t\boldsymbol{w}') - f_\lambda(\boldsymbol{x}) - t\mathbf{0}^{\mathrm{T}}\boldsymbol{w}'}{t^2/2} &\leq \liminf_{k \to \infty} \frac{f_\lambda(\boldsymbol{x} + t_k\boldsymbol{w}) - f_\lambda(\boldsymbol{x}) - t_k\mathbf{0}^{\mathrm{T}}\boldsymbol{w}}{t_k^2/2} \\
&\leq \liminf_{k \to \infty}\left(\frac{f(t_k\boldsymbol{u}_k)}{t_k^2/2} + \frac{1}{\lambda}\|\boldsymbol{u}_k - \boldsymbol{w}\|^2\right) \\
&= \lim_{k \to \infty} \frac{f(t_k\boldsymbol{u}_k)}{t_k^2/2} + \lim_{k \to \infty} \frac{1}{\lambda}\|\boldsymbol{u}_k - \boldsymbol{w}\|^2 \\
&= \mathrm{d}^2 f(\boldsymbol{x}; \mathbf{0})(\boldsymbol{u}) + \frac{1}{\lambda}\|\boldsymbol{u} - \boldsymbol{w}\|^2;
\end{aligned}$$

i.e.,

$$\frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \leq \frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{u}) + \frac{1}{2\lambda}\|\boldsymbol{u}-\boldsymbol{w}\|^2 \quad \text{for all } \boldsymbol{u},\boldsymbol{w}\in\mathbb{R}^n,$$

or in other words, for all $\boldsymbol{w}\in\mathbb{R}^n$,

$$\frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \leq \min\left\{\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{u}) + \frac{1}{2\lambda}\|\boldsymbol{u}-\boldsymbol{w}\|^2 : \boldsymbol{u}\in\mathbb{R}^n\right\} = \left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})\right)_\lambda(\boldsymbol{w}),$$

as desired.

To see the other direction, we first observe through (Rockafellar & Wets, 2009, Lemma 1.7) again that for all $\boldsymbol{w}\in\mathbb{R}^n$ there must exist a sequence $(t_1,\boldsymbol{w}_1),(t_2,\boldsymbol{w}_2),\ldots$ with $t_k \searrow 0$ and $\boldsymbol{w}_k\to\boldsymbol{w}$ as $k\to\infty$ such that

$$\begin{aligned}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) &= \lim_{k\to\infty}\frac{f_\lambda(\boldsymbol{x}+t_k\boldsymbol{w}_k)-f_\lambda(\boldsymbol{x})-t_k\boldsymbol{0}^{\mathrm{T}}\boldsymbol{w}_k}{t_k^2/2}\\ &= \lim_{k\to\infty}\min\left\{\frac{f(t_k\boldsymbol{y})}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}-\boldsymbol{w}_k\|^2 : \boldsymbol{y}\in\mathbb{R}^n\right\},\end{aligned}$$

where the last equality follows from (6) with $t=t_k$ and $\boldsymbol{w}'=\boldsymbol{w}_k$. Now, let $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots$ be such that

$$\frac{f(t_k\boldsymbol{y}_k)}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}_k-\boldsymbol{w}_k\|^2-\frac{1}{k}\leq\min\left\{\frac{f(t_k\boldsymbol{y})}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}-\boldsymbol{w}_k\|^2 : \boldsymbol{y}\in\mathbb{R}^n\right\}\quad\text{for all } k=1,2,\ldots,$$

and thus

$$\limsup_{k\to\infty}\left(\frac{f(t_k\boldsymbol{y}_k)}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}_k-\boldsymbol{w}_k\|^2\right)\leq\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}).$$

We claim that $\limsup_{k\to\infty}\|\boldsymbol{y}_k\|<\infty$. To this end, we first observe that

$$\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) = \lim_{k\to\infty}\min\left\{\frac{f(t_k\boldsymbol{y})}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}-\boldsymbol{w}_k\|^2 : \boldsymbol{y}\in\mathbb{R}^n\right\}\leq\limsup_{k\to\infty}\frac{1}{\lambda}\|\boldsymbol{w}_k\|^2<\infty,$$

as $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots$ is a convergent sequence. Then, we estimate

$$\frac{f(t_k\boldsymbol{y}_k)}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}_k-\boldsymbol{w}_k\|^2\geq-\rho\|\boldsymbol{y}_k\|^2+\frac{1}{\lambda}\|\boldsymbol{y}_k-\boldsymbol{w}_k\|^2=\left(\frac{1}{\lambda}-\rho\right)\|\boldsymbol{y}_k\|^2-\frac{2}{\lambda}\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{w}_k+\frac{1}{\lambda}\|\boldsymbol{w}_k\|^2,$$

where the inequality follows from (Davis & Drusvyatskiy, 2019, Lemma 2.1(3)). However, as $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots$ is convergent, it must be uniformly bounded; i.e., $\|\boldsymbol{w}_k\|\leq M$ for some $M>0$ and all $k=1,2,\ldots$. This implies that

$$\left(\frac{1}{\lambda}-\rho\right)\|\boldsymbol{y}_k\|^2-\frac{2}{\lambda}\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{w}_k+\frac{1}{\lambda}\|\boldsymbol{w}_k\|^2\geq\left(\frac{1}{\lambda}-\rho\right)\|\boldsymbol{y}_k\|^2-\frac{2M}{\lambda}\|\boldsymbol{y}_k\|\to\infty\quad\text{as }\|\boldsymbol{y}_k\|\to\infty,$$

as $1/\lambda-\rho>0$ as assumed. As a result, if $\limsup_{k\to\infty}\|\boldsymbol{y}_k\|=\infty$, then

$$\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w})\geq\limsup_{k\to\infty}\left(\frac{f(t_k\boldsymbol{y}_k)}{t_k^2/2}+\frac{1}{\lambda}\|\boldsymbol{y}_k-\boldsymbol{w}_k\|^2\right)\geq\limsup_{k\to\infty}\left(\left(\frac{1}{\lambda}-\rho\right)\|\boldsymbol{y}_k\|^2-\frac{2M}{\lambda}\|\boldsymbol{y}_k\|\right)=\infty,$$

a contradiction. It then follows from the claim that $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots$ is uniformly bounded, and thus admits a subsequence $\boldsymbol{y}_{k_1},\boldsymbol{y}_{k_2},\ldots$ with $\lim_{\ell\to\infty}\boldsymbol{y}_{k_\ell}=\boldsymbol{y}_0$ for some $\boldsymbol{y}_0\in\mathbb{R}^n$ by the

Bolzano–Weierstrass theorem, which further implies that

$$\begin{aligned}
\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) &\geq \limsup_{k\to\infty} \left(\frac{f(t_k \boldsymbol{y}_k)}{t_k^2/2} + \frac{1}{\lambda}\|\boldsymbol{y}_k - \boldsymbol{w}_k\|^2 \right) \\
&\geq \limsup_{\ell\to\infty} \left(\frac{f(t_{k_\ell} \boldsymbol{y}_{k_\ell})}{t_{k_\ell}^2/2} + \frac{1}{\lambda}\|\boldsymbol{y}_{k_\ell} - \boldsymbol{w}_{k_\ell}\|^2 \right) \\
&\geq \liminf_{\ell\to\infty} \frac{f(t_{k_\ell} \boldsymbol{y}_{k_\ell})}{t_{k_\ell}^2/2} + \frac{1}{\lambda}\|\boldsymbol{y}_0 - \boldsymbol{w}\|^2 \\
&\geq \liminf_{t\searrow 0,\, \boldsymbol{y}'\to\boldsymbol{y}_0} \frac{f(t\boldsymbol{y}')}{t^2/2} + \frac{1}{\lambda}\|\boldsymbol{y}_0 - \boldsymbol{w}\|^2 \\
&= \mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{y}_0) + \frac{1}{\lambda}\|\boldsymbol{y}_0 - \boldsymbol{w}\|^2 \\
&\geq \min\left\{ \mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{u}) + \frac{1}{\lambda}\|\boldsymbol{u} - \boldsymbol{w}\|^2 : \boldsymbol{u} \in \mathbb{R}^n \right\} \\
&= 2\left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0}) \right)_\lambda (\boldsymbol{w});
\end{aligned}$$

i.e.,

$$\frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \geq \left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0}) \right)_\lambda (\boldsymbol{w}) \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^n,$$

as desired. □

We are now ready to prove Theorem 5.3.

Proof of Theorem 5.3. If $\boldsymbol{0} \in \partial f(\boldsymbol{x})$ and $\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \geq 0$ for all $\boldsymbol{w} \in \mathbb{R}^n$, then $\nabla f_\lambda(\boldsymbol{x}) = \boldsymbol{0}$ by (Davis & Drusvyatskiy, 2022, Lemma 2.5(4)) and

$$\frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) = \left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0}) \right)_\lambda (\boldsymbol{w}) = \min\left\{ \frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{u}) + \frac{1}{2\lambda}\|\boldsymbol{u} - \boldsymbol{w}\|^2 : \boldsymbol{u} \in \mathbb{R}^n \right\} \geq 0.$$

If $\nabla f_\lambda(\boldsymbol{x}) = \boldsymbol{0}$ and $\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \geq 0$ for all $\boldsymbol{w} \in \mathbb{R}^n$, then $\boldsymbol{0} \in \partial f(\boldsymbol{x})$ by (Davis & Drusvyatskiy, 2022, Lemma 2.5(4)) and

$$\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \geq \left(\frac{1}{2}\mathrm{d}^2 f(\boldsymbol{x};\boldsymbol{0}) \right)_\lambda (\boldsymbol{w}) = \frac{1}{2}\mathrm{d}^2 f_\lambda(\boldsymbol{x};\boldsymbol{0})(\boldsymbol{w}) \geq 0,$$

as desired. □

While Theorem 5.3 already reduces the problem of computing a Rockafellar SOSP of a potentially nonsmooth function to a simpler setting (i.e., L-smooth functions) without any loss, it is still natural to wonder whether the reduced problem can in fact be even easier; i.e., whether the class of Moreau envelopes generated by all weakly convex functions can be a proper subset of all L-smooth functions. The following realizability result of Moreau envelopes, which follows as an immediate consequence of the proximal-hull techniques in (Rockafellar & Wets, 2009), offers a refutation of such a possibility through a stronger statement: Every L-smooth function can be generated as the Moreau envelope of another L'-smooth function, rather than a weakly convex function.

Proposition F.2 *Let $L > 0$ and $0 < \lambda < 1/L$ be arbitrary, and $g : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. We have*

$$g = \left(-(-g)_\lambda\right)_\lambda, \quad \textit{where } -(-g)_\lambda \textit{ is } \max\left\{ \frac{1}{\lambda}, \frac{L}{1-\lambda L} \right\}\textit{-smooth.}$$

Proof. As g is L-smooth, $-g$ is L-smooth, and thus L-weakly convex; see, e.g., (Pougkakiotis & Kalogerias, 2023, Proposition 2.3) and (Vial, 1983, Proposition 4.12). Hence, the function

$$-g + \frac{1}{2\lambda}\|\bullet\|^2 = \left(-g + \frac{L}{2}\|\bullet\|^2 \right) + \left(\frac{1}{2\lambda} - \frac{L}{2} \right)\|\bullet\|^2$$

is convex. It then follows from (Rockafellar & Wets, 2009, Example 11.26(d)) that $-g$ is λ-proximal; i.e.,

$$-g = -\big(-(-g)_\lambda\big)_\lambda \quad \Longleftrightarrow \quad g = \big(-(-g)_\lambda\big)_\lambda$$

by (Rockafellar & Wets, 2009, Example 1.44). It remains to note that as $-g$ is L-weakly convex, $(-g)_\lambda$ is $\max\{1/\lambda, L/(1-\lambda L)\}$-smooth by (Davis & Drusvyatskiy, 2022, Lemma 2.5(3)), and $-(-g)_\lambda$ is thus $\max\{1/\lambda, L/(1-\lambda L)\}$-smooth as well. □

G Proof of Proposition 5.4

Proof of Proposition 5.4. We shall apply (Khanh et al., 2025a, Theorem 3.5) to show that $\boldsymbol{x}$ is also a local minimizer of f_λ. To this end, we first claim that a prox-bounded threshold of f (in the sense of (Khanh et al., 2025a, Lemma 3.1(iii))) can be $1/\rho$; i.e., $f_\mu(\boldsymbol{x}) > -\infty$ for all $\mu \in (0, 1/\rho)$ and $\boldsymbol{x} \in \mathbb{R}^n$. But this is trivial, as $\boldsymbol{y} \mapsto f(\boldsymbol{y}) + \frac{1}{2\mu}\|\boldsymbol{y} - \boldsymbol{x}\|^2$ is strongly convex (see, e.g., (Khanh et al., 2025b, Remark 2.5(iii)) for more details), and thus lower bounded. Hence, by (Khanh et al., 2025a, Theorem 3.5), we have $\boldsymbol{x}$ is a local minimizer of f_λ. As f_λ is $\max\{\lambda^{-1}, \rho/(1-\lambda\rho)\}$-smooth by (Davis & Drusvyatskiy, 2022, Lemma 2.5(3)), the desired claims then immediately follow by applying Fermat's rule (see, e.g., (Rockafellar & Wets, 2009, Theorem 10.1)) and the second-order optimality condition (Hiriart-Urruty et al., 1984, Theorem 3.1). □

H An example in bilevel optimization

Example H.1 *Consider the functions $u, \ell : \mathbb{R}^2 \to \mathbb{R}$ and $r : \mathbb{R} \to \mathbb{R}$, given by $u(x,y) := x^2 + y$, $\ell(x,y) := (y-x)^2/2 + r(y)$, and*

$$r(y) := \begin{cases} \frac{1}{6}|y|^3, & |y| \le 1, \\ \frac{1}{2}y^2 - \frac{1}{2}|y| + \frac{1}{6}, & |y| > 1, \end{cases}$$

respectively. It can be calculated that

$$r'(y) = \begin{cases} \frac{1}{2}y|y|, & |y| \le 1, \\ y - \frac{1}{2}\operatorname{sign}(y), & |y| > 1, \end{cases} \quad \text{and} \quad r''(y) = \begin{cases} |y|, & |y| \le 1, \\ 1, & |y| > 1, \end{cases}$$

which imply all the assumptions in Lemma 5.7 except the lower boundedness of h, and that $\ell(x, \bullet)$ is not three times differentiable. To see that h is indeed lower bounded, we calculate

$$y^\star(x) = \begin{cases} (2x-1)/4, & x \le -3/2, \\ 1 - \sqrt{1-2x}, & -3/2 < x \le 0, \\ \sqrt{2x+1} - 1, & 0 < x \le 3/2, \\ (2x+1)/4, & x > 3/2, \end{cases}$$

and thus

$$\min\{h(x) : x \in \mathbb{R}\} = \min\{x^2 + y^\star(x) : x \in \mathbb{R}\} \approx -0.1823 > -\infty,$$

as desired. However, as $y^\star$ is not twice differentiable at 0, neither is h.